

SportsMOT: A Large Multi-Object Tracking Dataset in Multiple Sports Scenes

Yutao Cui* Chenkai Zeng* Xiaoyu Zhao* Yichun Yang* Gangshan Wu Limin Wang[✉]

State Key Laboratory for Novel Software Technology, Nanjing University, China

<https://github.com/MCG-NJU/SportsMOT>

Abstract

Multi-object tracking (MOT) in sports scenes plays a critical role in gathering players statistics, supporting further applications, such as automatic tactical analysis. Yet existing MOT benchmarks cast little attention on this domain. In this work, we present a new large-scale multi-object tracking dataset in multiple sports scenes, coined as SportsMOT, where all players on the court are supposed to be tracked. It consists of 240 video sequences, over 150K frames (almost 15× MOT17) and over 1.6M bounding boxes (3× MOT17) collected from 3 sports categories, including basketball, volleyball and football. Our dataset is characterized with two key properties: 1) fast and variable-speed motion and 2) similar yet distinguishable appearance. We expect SportsMOT to encourage the MOT trackers to promote in both motion-based association and appearance-based association. We benchmark several state-of-the-art trackers and reveal the key challenge of SportsMOT lies in object association. To alleviate the issue, we further propose a new multi-object tracking framework, termed as MixSort, introducing a MixFormer-like structure as an auxiliary association model to prevailing tracking-by-detection trackers. By integrating the customized appearance-based association with the original motion-based association, MixSort achieves state-of-the-art performance on SportsMOT and MOT17. Based on MixSort, we give an in-depth analysis and provide some profound insights into SportsMOT.

1. Introduction

Multi-object tracking (MOT) has been a fundamental computer vision task for recent decades, aiming to locate the objects and associate them in video sequences. Researchers have cast much focus on various practical use cases like crowded street scenes [9, 27], static dancing scenes [33] and driving scenarios [14], achieving consid-

erable progress [2, 4, 30, 37, 38, 40, 41] in MOT. MOT for sports scenes however is overlooked, where typically only the players on the court should be tracked for further analysis, such as counting the players' running distance or average speed and automatic tactical analysis.

Generally, prevailing state-of-the-art trackers [1, 6, 11, 39, 44] consist of several components to accomplish the tracking task: objects localization module, motion based objects association module and appearance based association module. Biased to the data distribution of specific human tracking benchmarks, e.g. MOT17 [27], MOT20 [9] and DanceTrack [33], the components of these trackers have difficulty adapting to sports scenes. Firstly, motivated by surveillance or self-driving applications, current human tracking benchmarks provide tracks for almost all persons in the scenes. While for sports scenes like basketball or football games, generally only the players on the court are what we focus on, hence a specialized training platform is required to make the detectors suitable for sports scenes. More importantly, in MOT17 and MOT20, these trackers highlight Kalman Filter [16] based IoU matching for object association, due to the slow and regular motion of pedestrians. DanceTrack highlights diverse motion rather than fast movement [33], that is, dancers frequently switch the motion direction and relative position. However in sports scenes, we observe *fast and variable-speed movement* of objects on adjacent frames, i.e. players usually possess high speed and frequently change their running speed in professional sports events, thus constituting barriers in existing motion based association. For instance, as visualized in Fig. 1, the adjacent IoU and Kalman-Filter-based IoU in sports scenes remain lower than that on MOT17 and DanceTrack (More detailed comparison can be found in Fig. 2 and Fig. 3). As a consequence, more suitable motion based association for sports scenes is required. Additionally, compared to the MOT17 and MOT20 datasets in street scenes, the objects appearances in sports scenes are less distinguishable, since not only the players inherently are in similar clothes but also the players are frequently blurred caused by fast camera motion or targets motion. Different from DanceTrack, where generally the dancers are in almost the same clothes and

* Equal contribution. ✉ : Corresponding author.

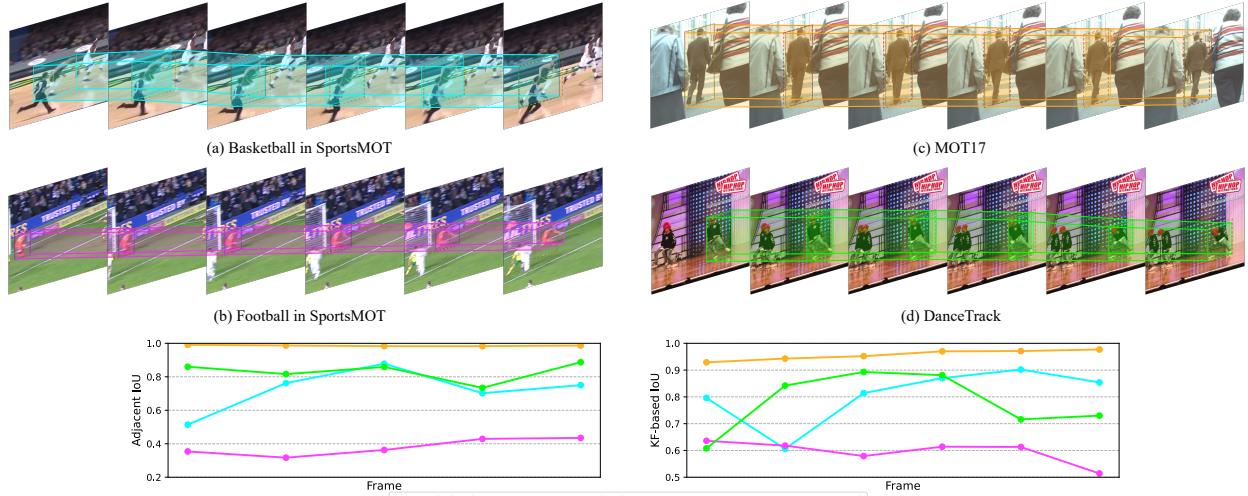


Figure 1. Sampled sequences from the categories of basketball and football of SportsMOT, MOT17 and DanceTrack. There exist two key properties of SportsMOT: 1) *fast and variable-speed motion*, i.e. players usually possess high speed and frequently change their running speed (the visualized adjacent IoU and Kalman Filter based adjacent IoU can indicate the property); 2) *similar yet distinguishable appearance*, that is, players in sports scenes inherently wear jerseys with different numbers and usually display distinct postures. We expect SportsMOT to encourage the MOT trackers to promote in both motion-based association and appearance based association.

thus having indistinguishable appearance, players in sports scenes inherently wear jerseys with different numbers and usually display distinct postures. Thereby, we argue that objects in sports scenes are with *similar yet distinguishable appearance*, which necessitates the appearance model developing more discriminative and extensive representations.

Considering the analysis above, to advance the development of tracking and sports analysis, we propose a multi-object tracking dataset in sports scenes, termed as **SportsMOT**. The dataset is large-scale, high-quality and contains dense annotations for every player on the court in various sports scenes. It consists of 240 videos, over 150K frames (almost $15 \times$ MOT17 [27]) and over 1.6M bounding boxes ($3 \times$ MOT17 [27]) collected from 3 categories of sports, including basketball, volleyball and football. To provide a platform for making the trackers suitable for sports scenes, we split the dataset into training, validation and test subsets, consisting of 45, 45 and 150 video sequences respectively. There exist two core properties of SportsMOT: (1) ***fast and variable-speed motion***, requiring more suitable motion modeling association; (2) ***similar yet distinguishable appearance***, which necessitates the appearance model developing more discriminative and extensive representations. Altogether, we expect SportsMOT to encourage the trackers to promote in both the certain aspects, i.e. motion based association and appearance based association.

Given the large-scale multi-object tracking dataset SportsMOT, we benchmark some recent tracking approaches and retrain all of them on the training split. We observe IDF1 and AssA metrics are lower than that on MOT17 while the DetA is quite high, indicating that the main chal-

lenge of SportsMOT lies in objects association rather than objects localization. To alleviate the issue, we propose a new multi-object tracking framework, dubbed as **MixSort**, with introducing a MixFormer-like [8] structure as appearance based association to prevailing tracking-by-detection trackers (e.g. ByteTrack [44], OC-SORT [6]). By integrating the original motion based objects association and the designed appearance based association, the performance gets boosted on both SportsMOT and MOT17 benchmarks. Based on MixSort, we perform extensive exploration studies and provide some profound insights into SportsMOT.

The main contributions are summarized as follows:

- We build a new large-scale multi-object tracking dataset in diverse sports scenes, SportsMOT, equipped with two key properties of 1) fast and variable-speed motion and 2) similar yet distinguishable appearance, aiming to advance the development of both tracking and sports analysis.
- We benchmark some prevailing trackers on SportsMOT, which reveals that the key challenge lies in objects association and hopefully can facilitate further research.
- We propose a new multi-object tracking framework MixSort, with introducing a MixFormer-like structure as appearance based association model to prevailing tracking-by-detection trackers, so as to boost the objects association. Based on MixSort, we perform extensive studies and provide some profound insights into SportsMOT.

Dataset	Videos	Frames	Length (s)	Bbox	Tracks
MOT17	14	11,235	463	292,733	1,342
MOT20	8	13,410	535	1,652,040	3,456
DanceTrack	100	105,855	5,292	-	990
SportsMOT	240	150,379	6,015	1,629,490	3,401

Table 1. Comparison of statistics between existing human MOT datasets and our SportsMOT.

2. Related Work

Multi-object tracking datasets. Existing Multi-object tracking datasets usually focus on different scenes, such as autonomous driving, pedestrians on roads, and dancing. For autonomous driving, there are KITTI [14], KITTI360 [22] and BDD100K [42], which focus on pedestrians and vehicles. Besides, other datasets, focusing on only the pedestrians, collect the videos from static and moving cameras. One of the earliest is PETS [12], but it’s too simple in some scenes. MOT15 [20] proposes the first large-scale benchmark for Multi-object tracking, followed by MOT17 [27] and MOT20 [9]. It’s worth noting that MOT20 focuses on extremely crowded scenes where many pedestrians are occluded, increasing the tracking difficulty greatly in both detection and association. Recently, DanceTrack [33], focusing on dancing scenes, is proposed to encourage trackers to rely less on visual discrimination and depend more on motion analysis. The emphasized properties are uniform appearance and diverse motion. While in SportsMOT, the appearance is similar yet distinguishable, and the players’ motion is fast and with variable speed. We expect SportsMOT to encourage algorithms to promote in both appearance and motion association. Besides, SoccerNet-Tracking [7] is presented to track elements in football scenarios. The main difference lies in that, it only contains soccer scenes and tracks almost all elements (players, goalkeepers, referees, balls) on the court without distinction. While SportsMOT contains three types of sports where the objects in basketball scenes are more crowded and thus more challenging (refer to Section 5.3), and only focuses on the players to support further statistics and tactical analysis.

Object association in tracking. Association is a very important task in tracking, where trackers need to associate detections in new frames with existing tracks. For most of the trackers, a similarity matrix (or cost matrix) between new detections and tracks is computed based commonly on motion and appearance cues, which is later fed into Hungarian algorithm [18] to perform association. For example, SORT [4] uses Kalman Filter to predict the location of objects and computes the IoU of detected and predicted bounding boxes as similarity matrix. IOU-Tracker [5] directly computes the IoU without prediction. ByteTrack [44] adds an association phase for detection with low confidence score, which can boost the performance. OC-SORT [6] tries to address the limitation of Kalman Filter.

Appearance cues also play an important role in association. DeepSORT [39] crops the detection from frame images, which are then used by networks to generate re-ID features. Then the motion cues and distance of re-ID features are fused to perform association. FairMOT [45] uses a re-ID branch on a backbone shared by the detection branch to generate re-ID features. In CenterTrack [46], the previous frame is used to help the prediction of tracks. Recently, Transformer [36] is used by some work such as TrackFormer [26] and MOTR [43] to boost the association quality. Our proposed MixSort integrates the motion and appearance cues with a motion modeling component and the designed MixFormer-like structure respectively.

3. SportsMOT Dataset

3.1. Dataset Construction

Video Collection. We select three worldwide famous sports, football, basketball, and volleyball, and collect videos of high-quality professional games including NCAA, Premier League, and Olympics from Multi-Sports [21], which is a large dataset in sports area focusing on spatio-temporal action localization. Each category has typical players’ formations and motion patterns, and they can effectively represent the diversity of sports scenarios. Only the overhead shots of sports game scenes are used, guaranteeing certain extreme situations do not occur. The proposed dataset consists of 240 video sequences in total, each of which is 720P and 25 FPS. Following the principles of multi-object tracking, each video clip is manually checked to ensure that there are no abrupt viewpoint switches within the video.

Annotation Pipeline. We annotate the collected videos according to the following guidelines.

- The entire athlete’s limbs and torso, excluding any other objects like balls touching the athlete’s body, are required to be annotated.
- The annotators are asked to predict the bounding box of the athlete in the case of occlusion, as long as the athletes have a visible part of body. However, if half of the athletes’ torso is outside the view, annotators should just skip them.
- We ask the annotators to confirm that each player has a unique ID throughout the whole clip.

We provide a customized labeling tool for SportsMOT and a corresponding manual book to annotators. Once they start annotating a new object, the labeling tool automatically assigns a new ID to the object and propagates the bounding box of previous state to the current state, with the help of the single object tracker KCF [15]. Then the generated bounding boxes should be refined by the annotators, so as to improve annotation quality. After carefully reviewing each

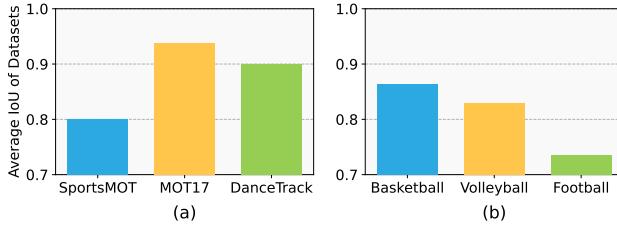


Figure 2. IoU on adjacent frames. (a) Compared to MOT17 and DanceTrack, SportsMOT has a lower score, indicating that objects have faster motion. (b) In SportsMOT, the category of football has the lowest IoU score, which means that football players often have fast motion.

Category	Frames	Tracks	Track gap len.	Track len.	Bboxes per frame
Basketball	845.4	10	68.7	767.9	9.1
Volleyball	360.4	12	38.2	335.9	11.2
Football	673.9	20.5	116.1	422.1	12.8
Total	626.6	14.2	96.6	479.1	10.8

Table 2. Detailed statistics of the three categories in SportsMOT. annotation result, we refine the bounding boxes and IDs that do not satisfy the standards, hence building a high-quality dataset. Finally, the bounding boxes with too small size, *i.e.* $w < 5$ or $h < 5$, are deleted.

3.2. Dataset Statistic

Overview. SportsMOT is a large-scale and high-quality MOT dataset, aiming to advance the development of both sports analysis and multi-object tracking. Table 1 compares the statistics of SportsMOT with the prevailing human tracking datasets, including MOT17, MOT20 and DanceTrack. According to the statistics, SportsMOT has a large number of bounding boxes of over 1.6M, which is comparable to MOT20 and significantly larger than MOT17. Besides, SportsMOT has a large number of video clips, tracks ($2.5 \times$ MOT17, $3.4 \times$ DanceTrack), frames ($13.4 \times$ MOT17). As shown in Table 2, we also compare the basic statistics of each category of SportsMOT. SportsMOT solely provides fine annotations of the players on the court, which are supposed to be tracked for further analysis. To provide a platform for making the trackers suitable for sports scenes, we split the dataset into training, validation and test subsets, consisting of 45, 45 and 150 video sequences respectively.

Fast and Variable-Speed Motion. Motion cues play an important role in object association for multi-object tracking. The existing human-tracking datasets (*e.g.* MOT17, MOT20 and DanceTrack) generally have certain motion patterns that are distinct with sports scenes, constituting barriers in players tracking. For instance, in MOT17 and MOT20, pedestrians are featured by linear motion with constant speed, which is easily hacked by association strategies with constant velocity assumption. Besides, Dance-

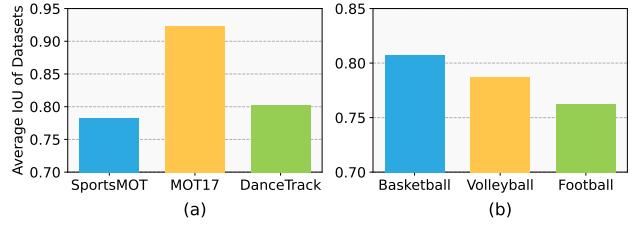


Figure 3. Kalman-Filter-based IoU on adjacent frames. (a) Compared to MOT17 and DanceTrack, SportsMOT has a lower score, indicating that objects have more variable-speed motion. (b) In SportsMOT, the category of football has the lowest Kalman-Filter-based IoU score.

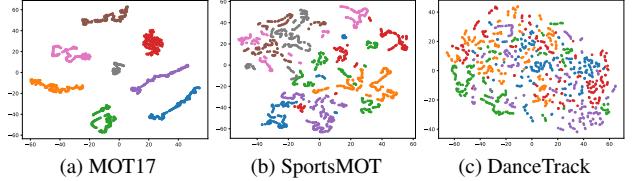


Figure 4. Visualization of re-ID features from sampled videos in MOT17, SportsMOT and DanceTrack dataset using t-SNE [35]. The same object is coded by the same color. It indicates that object appearance of SportsMOT is less distinguishable than that of MOT17, while more distinguishable than that of DanceTrack. We expect the appearance model to capture more discriminative and extensive representation for object association.

Track highlights diverse motion rather than fast motion, that is, dancers usually move in more diverse directions with relatively low speed. In contrast, SportsMOT has distinct motion patterns, *i.e. fast and variable-speed motion*, where the players typically move fast with their running speed or camera speed frequently changing. As illustrated in Fig. 2, among the three datasets, SportsMOT has the lowest IoU score of the objects bounding boxes on adjacent frames, indicating the fast movement. We use the ground truth of previous frames for Kalman Filter prediction. The result and current ground truth are used to calculate Kalman-Filter-based IoU. Seen from Fig. 3, SportsMOT also has the lowest Kalman-Filter-based IoU score on adjacent frames, which suggests that the motion can not be easily modeled by prevailing methods due to the variable-speed movement. Specifically, football has the smallest adjacent IoU and Kalman-Filter-based IoU, which is closely related to the fast running speed, abrupt acceleration or stops. It poses a major challenge for trackers based on simple motion assumptions and also encourages them to model object motion in more dynamic and adaptive ways.

Similar yet Distinguishable Appearance. Object appearance is another kind of cue on which MOT trackers often rely to distinguish different objects. In MOT17 and MOT20, pedestrians are usually distinct in body size and wear different clothes, yielding discriminative visual features. In contrast, the objects in DanceTrack typically wear nearly identical outfits, leading to indistinguishable

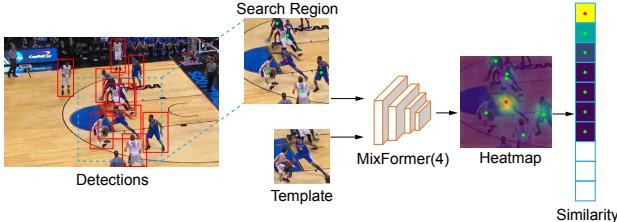


Figure 5. Paradigm for computing visual similarity matrix of tracks and detections in MixSort. The center of ground-truth target detection is marked with red dots, the others are green. The blue dashed box indicates the cropped search region. The blank part of similarity vector means that for detections not in the search region, the corresponding value is set to 0.

appearances. Thereby, DanceTrack highlights solely relying on motion-based association rather than appearance-based association. In SportsMOT, the players also have very similar appearances. However, the players wear jerseys with different numbers and usually display distinct postures, thus resulting in *similar yet distinguishable* appearance. In Fig. 4, we provide visualization of re-ID feature from sampled videos in MOT17, SportsMOT and DanceTrack dataset using t-SNE [35]. It implies that the re-ID features in SportsMOT are similar yet distinguishable compared to MOT17 and DanceTrack. We aim to encourage the trackers to learn more discriminative visual representations for more robust object association.

Evaluation Metrics. MOTA [3] is the main metric for existing MOT evaluations. However, MOTA focuses more on measuring the accuracy of detection. To highlight the performance of object association, we recommend HOTA [25], AssA and IDF1 [31] as the major evaluation metrics in SportsMOT dataset. HOTA aims to measure the accuracy of detection and association equally.

4. Multi-Object Tracking on SportsMOT

In this section, we present our proposed multi-object tracking framework, called *MixSort*. This framework is designed to enhance the appearance-based association performance and can be applied to any trackers that follow the tracking-by-detection paradigm, such as ByteTrack [44] and OC-SORT [6].

4.1. MixFormer for Appearance-based Association

MixFormer. In this paragraph, we discuss the use of MixFormer [8] in our proposed framework *MixSort*. MixFormer is designed to extract target-specific discriminative features and perform extensive communication between the target and search area, therefore, it is the key component that enables MixSort to compute visual similarities between the templates of tracked objects and detected objects in the search region of the current frames. The original MixFormer uses a corner-based localization head to predict the

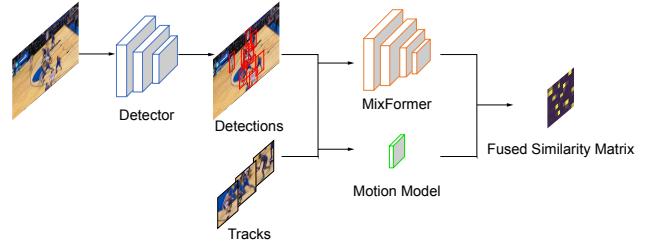


Figure 6. The pipeline of MixSort. We use motion model and MixFormer to generate fused similarity matrix for association.

top-left and bottom-right corners of the input template in the search region. However, we modify the corner head by using a heatmap prediction head that predicts the center of the template and generates a confidence heatmap. This allows us to compute the similarity between the detection and the template. To make MixSort suitable for multi-object tracking and accelerate inference speed, we reduce the number of mixed attention modules in MixFormer from 12 to 4. The steps involved in computing the visual similarity matrix are illustrated in Figure 5.

Association Strategy. In order to perform association between detections and existing tracks, we use a mixed similarity matrix generated by computing the visual similarity between the target template and the detected objects in the search region of the current frame. Specifically, we obtain the heatmap response at the center of each detection as its visual similarity to the template. The resulting similarity matrix is then combined with the IoU matrix using the Hungarian Algorithm.

To start, for each existing track t , we use the Kalman Filter to predict its new location. Then, we crop the current frame centered at the predicted location with a certain scale to obtain the search region s . By feeding s and the template t into MixFormer, we generate a heatmap \mathcal{H} that represents the similarity between the template and search region.

Next, for each detection d whose center is in the search region s , we set its similarity to track t as the response in the heatmap \mathcal{H} ; the similarity values of other detections are set to 0. Finally, we fuse the visual similarity and the IoU score to obtain the mixed similarity matrix

$$M = \alpha \cdot \text{IoU} + (1 - \alpha) \cdot V \quad (1)$$

where α is the weight coefficient and V represents the visual similarity matrix calculated using MixFormer.

4.2. MixSort Tracking

Based on the tracking-by-detection paradigm, the pipeline of *MixSort* can be generalized as follows. As shown in Figure 6, we first obtain detections using an object detector. Then, we employ a motion model (e.g., Kalman Filter) to predict new locations of existing tracks. Based on the new locations and templates of tracks, we compute a

	Training Setup	HOTA \uparrow	IDF1 \uparrow	AssA \uparrow	MOTA \uparrow	DetA \uparrow	LocA \uparrow	IDs \downarrow	Frag \downarrow
CenterTrack [46]	Train	62.7	60.0	48.0	90.8	82.1	90.8	10481	5750
FairMOT [45]	Train	49.3	53.5	34.7	86.4	70.2	83.9	9928	21673
QDTrack [29]	Train	60.4	62.3	47.2	90.1	77.5	88.0	6377	11850
TransTrack [34]	Train	68.9	71.5	57.5	92.6	82.7	91.0	4992	9994
GTR [47]	Train	54.5	55.8	45.9	67.9	64.8	89.0	9567	14525
ByteTrack [44]	Train	62.8	69.8	51.2	94.1	77.1	85.6	3267	4499
OC-SORT [6]	Train	71.9	72.2	59.8	94.5	86.4	92.4	3093	3474
ByteTrack	Train+Val	64.1	71.4	52.3	95.9	78.5	85.7	3089	4216
OC-SORT	Train+Val	73.7	74.0	61.5	96.5	88.5	92.7	2728	3144
MixSort-Byte	Train+Val	65.7 (+1.6)	74.1 (+2.7)	54.8 (+2.5)	96.2	78.8	85.7	2472	4009
MixSort-OC	Train+Val	74.1 (+0.4)	74.4 (+0.4)	62.0 (+0.5)	96.5	88.5	92.7	2781	3199

Table 3. Tracking performance of investigated algorithms on our proposed SportsMOT. The best results are shown in **bold**.

fused similarity matrix as described above and use it to associate tracks and detections by the Hungarian Algorithm. Finally, for matched tracks and detections, we update the online templates. For unmatched tracks, we keep them until the threshold is reached. For unmatched detections with confidence scores higher than the threshold, new tracks are initialized.

4.3. Training and Inference

Training. We only consider the training of MixFormer here since the detector remains the same as initial method (*e.g.* ByteTrack). The original MixFormer is trained on SOT datasets, so we first modify the format of ground truth of MOT datasets, that is, converting the ground truth trajectory of every single player into TrackingNet format [28].

For each ground-truth bounding box, we compute its corresponding center location (c_x, c_y) in the low-resolution heatmap. Following CornerNet [19], the ground-truth heatmap response is generated using 2D Gaussian kernel:

$$h_{xy} = \exp\left(-\frac{(x - c_x)^2 + (y - c_y)^2}{2\sigma^2}\right) \quad (2)$$

where σ is adaptive to the size of the bounding box. The training loss is a pixel-wise logistic regression with focal loss [23]:

$$L = - \sum_{xy} \begin{cases} (1 - \hat{h}_{xy})^\gamma \log(\hat{h}_{xy}) & , h_{xy} = 1; \\ (1 - h_{xy})^\beta (\hat{h}_{xy})^\gamma \log(1 - \hat{h}_{xy}) & , \text{otherwise.} \end{cases} \quad (3)$$

where γ and β are hyper-parameters in focal loss, and we set $\gamma = 2$, $\beta = 4$ following CornerNet.

Inference. For each track, we maintain only one template for keeping a balance between accuracy and speed. When a detection is matched to an existing track, we directly replace the original template to the new detection, if and only if the ratio of its uncovered (*i.e.* overlapping with any detected objects) area is larger than a certain threshold, so as to reduce the impact of misleading representations.

5. Experiments and Analysis

5.1. Experiment Setup

Dataset Split. In benchmark experiments, we follow the default split described in Sec. 3.2. In exploration study, we split the original MOT17 training set into two sets, used for training and validation respectively following CenterTrack.

Implementation Details. Following ByteTrack and OC-SORT, we use YOLOX [13] as our detector. Using COCO-pretrained model as the initialized weights, we first train the model on CrowdHuman [32] for 80 epochs and then train on SportsMOT for another 80 epochs. The remaining settings are the same as that in ByteTrack.

For MixFormer, we initialize the backbone with the model trained on VOT datasets and then fine-tune it on SportsMOT for 300 epochs with learning rate initialized as $1e-4$ and decreased to $1e-5$ at epoch 200. The optimizer is ADAM [17] with weight decay 10^{-4} . The sizes of search images and templates are 224×224 and 96×96 respectively. The max sample interval is set to 10. For every tracking result, we apply linear interpolation as post-processing, with maximum gap set to 20.

5.2. Benchmark Results

We evaluate several representative methods of three kinds on our dataset. ByteTrack [44], OC-SORT [6] and QDTrack [29] are trackers in tracking-by-detection paradigm. CenterTrack [46] and FairMOT [45] perform joint detection and tracking in one stage. TransTrack [34] and GTR [47] are trackers based on Transformer. Most of the current best multi-object tracking algorithms belong to tracking-by-detection paradigm, however, due to the separation of detection and tracking, the information cannot be shared completely. Joint-detection-and-tracking paradigm couples the two modules, with the goal of boosting the performance of each. Transformer-based-tracking methods are relatively new but have achieved great performance. Despite its huge potential, the model complexity and calcula-

α	1	0.9	0.8	0.7	0.6	0.5	0.4	0.3	0.2	0.1	0
basketball	65.9	66.1	66.2	65.9	65.3	64.8	63.6	60.7	56.7	47.1	26.9
volleyball	76.0	76.8	76.5	76.4	76.5	76.9	76.4	75.5	73.7	69.1	40.9
football	71.9	72.4	72.3	72.4	72.5	72.8	73.2	72.9	72.6	71.4	65.7

Table 4. Comparison of the HOTA metric of basketball, volleyball and football under different fusion parameters α on SportsMOT test set. The models are trained on SportsMOT training set.

tion cost are much higher, resulting in large memory and long training time.

All training settings including the number of epochs and change of learning rate are consistent with original papers. According to different default settings, we follow the commonly used pretraining datasets, such as CrowdHuman [32], COCO [24] and ImageNet [10], and apply SportsMOT-train with or without other datasets for finetuning for different methods. We compare the results in Tab. 3.

In sports scenes, the clear appearance and sparse density of objects allow current mature detection frameworks to generate bounding boxes with high accuracy. However, specialized detectors need to be trained for not detecting audience and referees. The key challenges are fast speed and motion blur, which forces us to pay more attention to improve the association performance. Besides, The wide range of HOTA and MOTA denotes SportsMOT is more distinguishable among different kinds of algorithms.

5.3. Exploration Study

In this section, we perform extensive studies on the proposed MixSort and SportsMOT.

Effectiveness of the proposed association module. We evaluate the effectiveness of MixSort by applying it to two state-of-the-art trackers, OC-SORT [6] and ByteTrack [44], which follow the tracking-by-detection paradigm and use YOLOX as their detector. The evaluation is conducted on the SportsMOT test set, and the results are presented in Tab. 3. Our experiments show that MixSort significantly improves the performance of both trackers, with OC-SORT achieving a 0.4 HOTA increase and ByteTrack achieving a 1.6 HOTA increase on SportsMOT. This demonstrates the effectiveness of MixSort in enhancing the association.

Appearance-based vs. Motion-based association. We have demonstrated that MixSort can improve association performance. In this paragraph, we investigate the impact of the fusion weight α on the ability of appearance cues to aid in conventional motion-based association. We evaluate OC-SORT with MixSort on the three categories in the SportsMOT test set using α values ranging from 1 to 0 in Eq. (1). The results, presented in Tab. 4, reveal that pure motion-based association ($\alpha = 1$) outperforms pure appearance-based association ($\alpha = 0$) in all categories, underscoring the significance of motion cues in sports scenes. Moreover, fused association surpasses both

IoU	Motion	Mix.	HOTA \uparrow	IDF1 \uparrow	AssA \uparrow	MOTA \uparrow	IDs \downarrow
✓		✓	71.5	71.2	58.1	95.9	4329
		✓	64.2	63.9	48.7	91.1	25947
✓	✓		64.1	71.4	52.3	95.9	3089
✓		✓	73.8	74.4	61.6	96.6	3203
✓	✓	✓	65.7	74.1	54.8	96.1	2469

Table 5. Results of the ablation experiment on SportsMOT test set. IoU means computing IoU between detections and the last location of existing tracks for association, while Motion means using Kalman filter to predict the location of tracks. The models are trained on SportsMOT training and validation set.

pure motion-based and appearance-based association, suggesting that both motion and appearance cues should be considered jointly for optimal results.

Our analysis of the three categories reveals that appearance cues provide the most significant improvement for football videos (+1.3), followed by volleyball (+0.9) and basketball (+0.3). Combining this with Fig. 2, which indicates that the adjacent IoU of football games is the smallest (i.e., motion is fastest) among the three categories, we can conclude that scenes with faster motion are more dependent on appearance cues.

Ablation study on MixSort. We ablate important components of our tracker (MixSort based on ByteTrack) including *IoU*, *Motion* (Kalman Filter) and *MixSort* on SportsMOT test set. The results are presented in Tab. 5. Surprisingly, we found that simple IoU without using motion prediction outperformed IoU with motion prediction by a large margin (from 64.1 HOTA to 71.5 HOTA), indicating that the Kalman filter, which assumes linear motion models, performed poorly on SportsMOT, where the motion patterns are far more complex than in previous datasets.

Furthermore, we observed that MixSort played a crucial role in boosting the performance of the tracker significantly. By fusing the IoU and MixSort cues, our method achieved the best performance of 73.8 HOTA compared to 71.5 HOTA of simple IoU in our experiments.

Analysis on different categories of SportsMOT. While the three SportsMOT categories share some common characteristics such as fast motion and similar appearance, they also have distinct features due to the different types of games. In this section, we analyze the results of our experiments on the three categories.

We first use the best HOTA metric from Tab. 4 to represent the *overall difficulty* of a category. Based on this metric, we find that basketball videos (66.17) are the most difficult, followed by football (73.19), and finally volleyball (76.91).

Next, we consider the *appearance-based* association (i.e. $\alpha = 0$ in Tab. 4). We observe a notable gap between the HOTA metrics of the three categories, where basketball (26.94) has a much lower HOTA than volleyball (40.94) and football (65.65). Similarly, basketball remains the most dif-

	MOT17-test					IDs↓	MOT17-val					
	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑		HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑	IDs↓
QDTrack [29]	53.9	66.3	52.7	68.7	55.6	3378	-	-	-	-	-	-
MOTR [43]	57.2	68.4	55.8	71.9	58.9	2115	-	-	-	-	-	-
GTR [47]	59.1	71.5	57.0	75.3	61.6	2859	63.0	75.9	66.2	71.3	60.4	-
ByteTrack [44]	63.1	77.3	62.0	80.3	64.5	2196	-	79.7	-	76.7	-	159
OC-SORT [6]	63.2	77.5	63.4	78.0	63.2	1950	68.0	79.3	69.9	77.9	-	-
MixSort-Byte	64.0	78.7	64.2	79.3	64.1	2235	69.4	81.1	71.3	79.9	68.2	155
MixSort-OC	63.4	77.8	63.2	78.9	63.8	1509	69.2	80.6	71.5	78.9	67.4	135

Table 6. Comparison of the state-of-the-art methods under the “private detector” protocol on MOT17-test set and MOT17-val set.

	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑
ByteTrack	47.1	51.9	31.5	88.2	70.5
MixSort-Byte	46.7	53.0	31.9	85.8	68.6

Table 7. Comparison of ByteTrack and MixSort-Byte on DanceTrack validation set. For MixSort-Byte, the fuse parameter α is 0.9, which results in the highest HOTA among {0.6, 0.7, 0.8, 0.9, 0.95}. The models are trained on DanceTrack training set.

	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑
SportsMOT	overall	65.7	74.1	54.8	96.1
	basketball	60.8	67.8	46.8	97.3
	volleyball	72.5	87.0	66.8	96.5
	football	66.4	73.6	56.3	94.9
SoccerNet	62.9	73.9	55.5	87.8	71.5

Table 8. Results of MixSort-Byte on SportsMOT and SoccerNet test set. The models are trained on SportsMOT training and validation set and SoccerNet training set respectively and the hyper-parameters are the same.

ficult in the *motion-based* association (*i.e.* $\alpha = 1$ in Tab. 4), while the volleyball becomes the easiest.

We believe that the differences in difficulty arise from several factors, including the size of the game court and the degree of physical confrontation among players. For instance, basketball scenes are played on smaller courts and involve more physical contact between players than football scenes. This can lead to more occlusion and blur in basketball videos, making the association task more challenging than in football scenes.

Comparison with DanceTrack. We evaluate MixSort-Byte on the DanceTrack validation set and compare the results with that of ByteTrack as shown in Tab. 7. Unlike the results on SportsMOT where MixSort brings significant improvement, on DanceTrack the original ByteTrack performs better instead, with HOTA, MOTA and DetA metrics all higher than MixSort-Byte. This indicates that appearances in our proposed dataset SportsMOT are *similar yet distinguishable*, while those in DanceTrack are much harder to distinguish. Therefore SportsMOT highlights both the motion-based and appearance-based associations.

Comparison with SoccerNet. We conducted experiments to compare SoccerNet that focuses only on soccer

	HOTA↑	IDF1↑	AssA↑	MOTA↑	DetA↑
ByteTrack	64.1	71.4	52.3	95.9	78.5
ByteTrack+ReID	64.8	72.2	53.4	96.1	78.8
MixSort-Byte	65.7	74.1	54.8	96.1	78.8

Table 9. Comparison of ByteTrack, ByteTrack with ReID model and MixSort-Byte on SportsMOT test set. The ReID model is the same as in DeepSORT [4] and all the models are trained on SportsMOT training set and the best results are shown in **bold**.

scenes with SportsMOT. Results shown in Tab. 8 suggest that SportsMOT is a challenging dataset with varying levels of difficulty across different sports. Specifically, basketball is proved to be the most difficult with the lowest HOTA of 60.8, while volleyball and football are relatively easier. Notably, MixSort obtains higher HOTA on overall SportsMOT than on SoccerNet, which is mainly because all the elements on the court are to be tracked in SoccerNet, leading to more false detections and much lower DetA (71.5 vs. 78.8). Yet, with a higher AssA on SoccerNet despite these false detections, it underscores SportsMOT’s challenging association and significance for sports tracking.

MixSort vs. ReID models. To validate the effectiveness of the MixFormer-like model in MixSort, we present a comparison with ReID models in Tab. 9. The results show that ByteTrack with the ReID model outperforms the original ByteTrack, highlighting the value of appearance-based association on SportsMOT. Notably, MixSort-Byte further improves upon ByteTrack with ReID by 0.9, 1.9, and 1.4 in HOTA, IDF1, and AssA. This underscores MixSort’s superior appearance modeling, capable of extracting broader and more discriminative representations and facilitating efficient offline learning.

More results on MOT17 The propose MixSort can be applied to any trackers following this paradigm and achieve state-of-the-art performance on SportsMOT. To further validate the effectiveness of MixSort, we compare MixSort with the prevailing trackers on MOT17 validation set and test set under the private detection protocol in Tab. 6. MixSort-Byte and MixSort-OC outperform these trackers in HOTA, IDF1.

Detailed Challenging attributes of SportsMOT. For a comprehensive statistical analysis and performance evalua-

Attr.	Definition of sequence	Attr.	Definition of target box
CM	Abrupt motion of the camera	LR	The target box is smaller than a threshold
OV	The target completely leaves the video frame	FOC	The area of target box intersecting with others is larger than 0.9
MB	The target region is blurred due to target or camera motion	POC	The area of target box intersecting with others is in range [0.4, 0.9]
FM	The total average IoU of target boxes in adjacent frames is smaller than a threshold	DEF	The sum of the relative changes in the length and width of the target boxes in adjacent frames is greater than a threshold

Table 10. Definition of attributes. For the attributes in the right half, we first calculate the proportion of boxes in each sequence that meet the criteria defined in the table. If this proportion exceeds a certain threshold, we consider the sequence to exhibit that attribute.

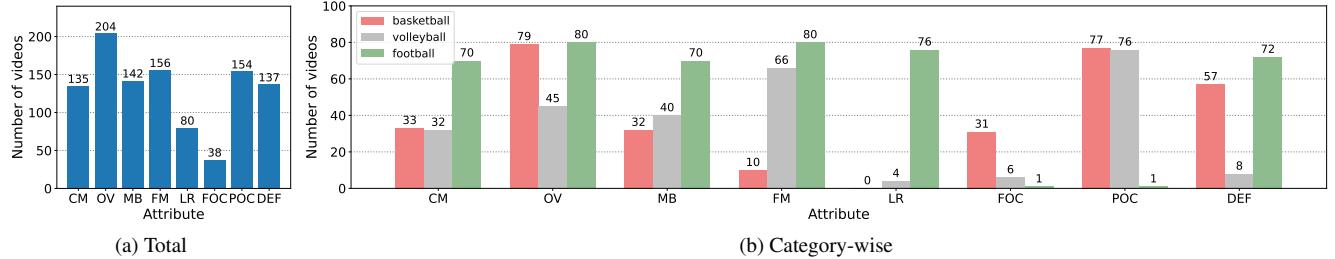


Figure 7. Distribution of sequences in each attribute on SportsMOT dataset.

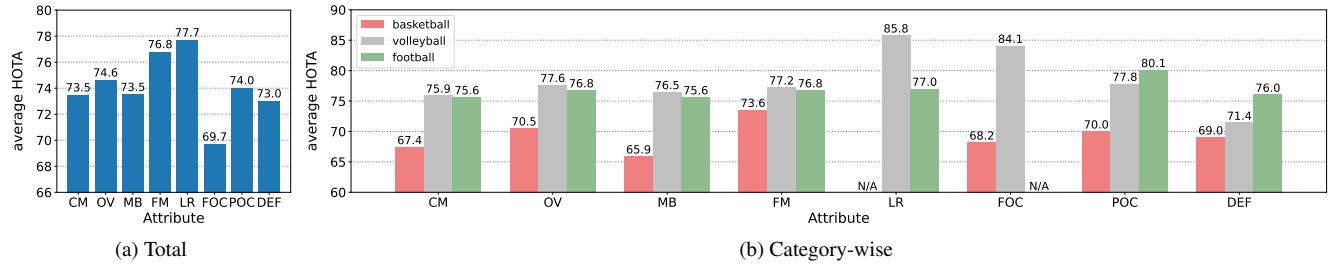


Figure 8. Average HOTA of sequences in each attribute on SportsMOT dataset. MixSort-OC is employed for analysis.

tion of the SportsMOT dataset, we categorize and present eight distinct challenging attributes: camera motion (CM), out of view (OV), motion blur (MB), fast motion (FM), low resolution (LR), full occlusion (FOC), partial occlusion (POC), and large deformation (DEF). The definitions for each attribute are specified in Tab. 10, with their corresponding distribution illustrated in Fig. 7. Human selection is used to identify the CM, OV, and MB attributes, while metrics calculation helps determine the others. The diversity and complexity of the scenarios featured in Fig. 7a suggest that our three chosen sports games provide a representative snapshot of broader sports analysis based on tracking states. Furthermore, in comparison to SoccerNet-Tracking [7], our SportsMOT offers a *more diverse* set of challenges observed in sports. These challenges range from the *frequent occlusion* seen in basketball and volleyball, to the *low resolution* typical of football, as depicted in Fig. 7b. We also assess the difficulty level posed by different attributes in the following.

Difficulty analysis based on HOTA scores. We employ the average HOTA scores of sequences corresponding to each attribute as an indicator of that attribute’s difficulty. The results are depicted in Fig. 8. The low HOTA score for *full occlusion*, evident in Fig. 8a, highlights its challeng-

ing nature in MOT within the sports field. Furthermore, an analysis of average HOTA scores shows that basketball consistently has lower scores than volleyball and football for all attributes (Fig. 8b), indicating its increased complexity.

6. Conclusion

In this paper, we have introduced *SportsMOT*, a large-scale multi-object tracking dataset in sports scenes. SportsMOT is characterized with two key properties: 1) fast and variable-speed motion and 2) similar yet distinguishable appearance. We have empirically investigated several prevailing MOT trackers on the SportsMOT dataset. We have also proposed a new MOT framework *MixSort*, introducing a MixFormer-like association module. Hopefully, SportsMOT can provide a platform for facilitating both sports analysis and multi-object tracking.

Acknowledgements. This work is supported by National Key R&D Program of China (No. 2022ZD0160900), National Natural Science Foundation of China (No. 62076119, No. 61921006), Fundamental Research Funds for the Central Universities (No. 020214380091, No. 020214380099), and Collaborative Innovation Center of Novel Software Technology and Industrialization.

References

- [1] Nir Aharon, Roy Orfaig, and Ben-Zion Bobrovsky. Bot-sort: Robust associations multi-pedestrian tracking. *arXiv preprint arXiv:2206.14651*, 2022. 1
- [2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019. 1
- [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. 5
- [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. IEEE, 2016. 1, 3, 8
- [5] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *2017 14th IEEE international conference on advanced video and signal based surveillance (AVSS)*, pages 1–6. IEEE, 2017. 3
- [6] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirrodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9686–9696, 2023. 1, 2, 3, 5, 6, 7, 8
- [7] 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. 3, 9
- [8] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13608–13618, 2022. 2, 5
- [9] 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. 1, 3
- [10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 7
- [11] Yunhao Du, Yang Song, Bo Yang, and Yanyun Zhao. Strongsort: Make deepsort great again. *arXiv preprint arXiv:2202.13514*, 2022. 1
- [12] James Ferryman and Ali Shahrokni. Pets2009: Dataset and challenge. In *2009 Twelfth IEEE international workshop on performance evaluation of tracking and surveillance*, pages 1–6. IEEE, 2009. 3
- [13] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. Yolox: Exceeding yolo series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 6
- [14] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012. 1, 3
- [15] Sion Hannuna, Massimo Camplani, Jake Hall, Majid Mirmehdi, Dima Damen, Tilo Burghardt, Adeline Paiement, and Lili Tao. Ds-kcf: a real-time tracker for rgb-d data. *Journal of Real-Time Image Processing*, 16(5):1439–1458, 2019. 3
- [16] Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45, 1960. 1
- [17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [18] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955. 3
- [19] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *Proceedings of the European conference on computer vision (ECCV)*, pages 734–750, 2018. 6
- [20] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. Motchallenge 2015: Towards a benchmark for multi-target tracking. *arXiv preprint arXiv:1504.01942*, 2015. 3
- [21] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13536–13545, October 2021. 3
- [22] Yiyi Liao, Jun Xie, and Andreas Geiger. Kitti-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 3
- [23] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 6
- [24] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 7
- [25] 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(2):548–578, 2021. 5
- [26] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixé, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8844–8854, 2022. 3
- [27] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016. 1, 2, 3
- [28] Matthias Müller, Adel Bibi, Silvio Giancola, Salman Al-Subaihi, and Bernard Ghanem. Trackingnet: A large-scale

- dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision, ECCV*, 2018. 6
- [29] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-dense similarity learning for multiple object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 164–173, 2021. 6, 8
- [30] Jinlong Peng, Changan Wang, Fangbin Wan, Yang Wu, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In *European conference on computer vision*, pages 145–161. Springer, 2020. 1
- [31] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016. 5
- [32] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 6, 7
- [33] 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. 1, 3
- [34] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 6
- [35] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008. 4, 5
- [36] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 3
- [37] Qiang Wang, Yun Zheng, Pan Pan, and Yinghui Xu. Multiple object tracking with correlation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3876–3886, 2021. 1
- [38] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards real-time multi-object tracking. In *European Conference on Computer Vision*, pages 107–122. Springer, 2020. 1
- [39] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017. 1, 3
- [40] Jialian Wu, Jiale Cao, Liangchen Song, Yu Wang, Ming Yang, and Junsong Yuan. Track to detect and segment: An online multi-object tracker. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12352–12361, 2021. 1
- [41] Yihong Xu, Yutong Ban, Guillaume Delorme, Chuang Gan, Daniela Rus, and Xavier Alameda-Pineda. Transcenter: Transformers with dense queries for multiple-object tracking. *arXiv preprint arXiv:2103.15145*, 2021. 1
- [42] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashishth Madhavan, and Trevor Darrell. Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2636–2645, 2020. 3
- [43] Fangao Zeng, Bin Dong, Yuang Zhang, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. In *European Conference on Computer Vision*, pages 659–675. Springer, 2022. 3, 8
- [44] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. Bytetrack: Multi-object tracking by associating every detection box. In *European Conference on Computer Vision*, pages 1–21. Springer, 2022. 1, 2, 3, 5, 6, 7, 8
- [45] 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(11):3069–3087, 2021. 3, 6
- [46] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *European Conference on Computer Vision*, pages 474–490. Springer, 2020. 3, 6
- [47] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8771–8780, 2022. 6, 8