applied sciences

MDPI

Article

A Novel Dataset for Multi-View Multi-Player Tracking in Soccer Scenarios

Xubo Fu [1,†], Wenbin Huang [2,3,4,†], Yaoran Sun [4], Xinhua Zhu [5], Julian Evans [2,3], Xian Song [6], Tongyu Geng [2,3] and Sailing He [2,3,4,5,*]

1 Department of Public Physical and Art Education, Zhejiang University, Hangzhou 310058, China; fycc@zju.edu.cn (X.F.)
2 National Engineering Research Center for Optical Instruments, Zhejiang University, Hangzhou 310058, China; Bingo@zju.edu.cn (W.H.)
3 Centre for Optical and Electromagnetic Research, Zhejiang University, Hangzhou 310058, China
4 Hangzhou Zhuxing Information Technology Co., Ltd., Hangzhou 311100, China
5 Shanghai Institute for Advanced Study, Zhejiang University, Shanghai 200135, China
6 Department of Applied Biology and Chemical Technology, The Hong Kong Polytechnic University, Hong Kong SAR, China
* Correspondence: sailing@jorcep.org
† These authors contributed equally to this work.

Abstract: Localization and tracking in multi-player sports present significant challenges, particularly in wide and crowded scenes where severe occlusions can occur. Traditional solutions relying on a single camera are limited in their ability to accurately identify players and may result in ambiguous detection. To overcome these challenges, we proposed fusing information from multiple cameras positioned around the field to improve positioning accuracy and eliminate occlusion effects. Specifically, we focused on soccer, a popular and representative multi-player sport, and developed a multi-view recording system based on a $1 + N$ strategy. This system enabled us to construct a new benchmark dataset and continuously collect data from several sports fields. The dataset includes 17 sets of densely annotated multi-view videos, each lasting 2 min, as well as 1100+ min multi-view videos. It encompasses a wide range of game types and nearly all scenarios that could arise during real game tracking. Finally, we conducted a thorough assessment of four multi-view multi-object tracking (MVMOT) methods and gained valuable insights into the tracking process in actual games.

Keywords: multi-view;tracking; soccer; system; benchmark;

Citation: Fu, X.; Huang, W.; Sun, Y.; Zhu, X.; Evans, J.; Song, X.; Geng, T.; He, S. A Novel Dataset for Multi-View Multi-Player Tracking in Soccer Scenarios. *Appl. Sci.* **2023**, *13*, 5361. https://doi.org/10.3390/app13095361

Academic Editors: Antonio Fernández-Caballero and Seokwon Yeom

Received: 12 February 2023
Revised: 17 April 2023
Accepted: 20 April 2023
Published: 25 April 2023

1. Introduction

In recent years, there has been a surge of interest in sports analysis. This analysis enables a range of applications, such as gathering data for coaching purposes [1] and improving the quality of sports broadcasts [2–4]. While most studies are focused on on-ball behavior, there is relatively little research on off-ball movement, which accounts for more than 95% of a player's playing time and plays a key role in game understanding [5–7]. Player tracking is the fundamental task of all off-ball research.

Recent advancements in video technology and the availability of large-scale annotated single-view pedestrian datasets [8–12] have significantly improved the accuracy and efficiency of deep-learning-based pedestrian trackers. Many modern algorithms [13–16] can learn transferable features across datasets, resulting in inspiring performance. As a result, single-view camera-based player tracking systems have become increasingly popular. However, difficulties in multi-object tracking occur when they are far away from the observing sensor or if their appearances are quite similar. In wide but highly cluttered sports scenarios especially [17], such situations make monocular detection and tracking insufficient.

Player tracking using multi-camera systems [18–27] is a promising solution for data acquisition, which aims to calculate the position of every target at any time, and to assemble

the complete trajectories from multi-view video streams. Using joint visual information from multiple synchronized cameras [28–30] will provide reliable estimates in crowded scenes.

Although multi-view multi-object tracking has flourished in the past few years, there is currently no large-scale and high-quality public dataset that focuses on tracking players on the field. The most commonly used MVMOT standard datasets [31–35] all target pedestrians in limited scenes. The most similar dataset LH0 [17], recorded junior players running by scripts on a 9 m $\times$ 26 m square, with annotations that are limited, making it difficult to train and test algorithms. Additionally, the cameras are positioned only slightly above the average human height, meaning that throughout the sequence, occluded situations are very difficult to deal with. The lack of relevant datasets makes it impossible to verify whether the existing methods can solve the problems in sports scenarios.

To accelerate the sports analysis research, we focused our attention on soccer, the most challenging and the most popular multi-player sport. More specifically, we installed the multi-view video recording system at several soccer fields for continuous data collection. The system was designed under the $1 + N$ strategy, which consists of 5$\sim$7 static super-high-resolution cameras with overlapping Fields of View (FOV). Then we created a large-scale player tracking dataset capturing unscripted games in various soccer scenarios on these fields. The dataset comprises three parts: 17 groups of time-synchronized 2 min long videos from multiple cameras with 20 fps tracking annotation for each player, precise joint camera calibrations that include intrinsic and extrinsic parameters, and a total of 1100+ min of multi-view videos.

Our experiments have demonstrated that trackers designed using publicly available datasets perform poorly in challenging sports environments due to differences in viewpoints and domain gaps.

We anticipate that the release of this multi-view multi-player tracking dataset will inspire further research in this field. Additionally, the dataset holds substantial value for related tasks such as multi-view human detection and tracking multiple individuals with a single camera.

To summarize, our workflow proceeded as shown in this Figure 1, and our contributions are as follows:

1. We designed and installed $1 + N$ based multi-view video recording systems in four soccer fields for continuous data collection.
2. We designed a novel $1 + N$ MVMOT annotation strategy based on the systems.
3. We constructed the largest densely annotated multi-view multi-player tracking dataset and provided a total of 1100+ min of multi-view videos to encourage research in automatic tracking of players in soccer scenarios.
4. We evaluated four state-of-the-art tracking models to understand the challenges and characteristics of our dataset.

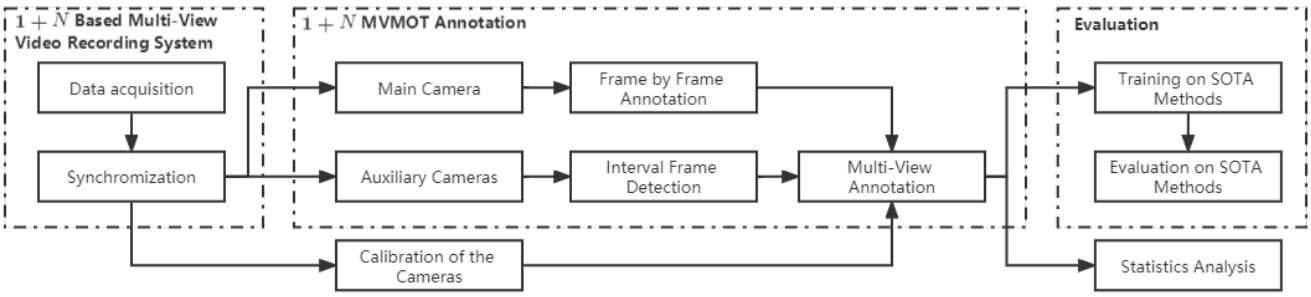


Figure 1. Our workflow proceeded as shown in this figure.

2. Related Works

2.1. Datasets

The most commonly used multi-view tracking datasets are listed in Table 1.

Table 1. Public Multi-view Datasets.

Dataset	Resolution	Cameras	FPS	IDs	Annotations	Size/Duration	Scenarios	Field Size
KITTI [12]	1392 × 480	4	10	10	389 ‡	7 min	campus roads	
Laboratory [21]	320 × 240	4	25	6	264 (1 fps)	4.4 min	laboratory †	5.5 × 5.5 m^2
Terrace [21]	320 × 240	4	25	9	1023 (1 fps)	3.5 min	terrace †	10 × 10 m^2
Passageway [21]	320 × 240	4	25	13	226 (1 fps)	20 min	square †	10 × 6 m^2
Campus [32]	1920 × 1080	4	30	25	240 (1 fps)	16 min	gargen †	
CMC * [31]	1920 × 1080	4	4	15	11,719 (4 fps)	1494 frames	laboratory †	7.67 × 3.41 m^2
SALSA [33]	1024 × 768	4	15	18	1200 (0.3 fps)	60 min	lobby †	
EPFL-RLC [35]	1920 × 1080	3	60	-	6132 ‡	8000 frames	lobby	
PETS-2009 [36]	768 × 576 720 × 576	7	7	19	4650 (7 fps)	795 frames	campus roads †	
WildTrack [34]	1920 × 1080	7	60	313	66,626 (2 fps)	60 min	square	36 × 12 m^2
APIDIS [37]	1600 × 1200	7	22	12	86,870 (25 fps)	1 min	Basketball Court †	28 × 15 m^2
LH0 * [17]	1920 × 1080	8	25	26	26,000 (2 fps)	1.3 min	soccer field †	38 × 7 m^2
Ours	3840 × 2160 4736 × 1400 5950 × 1152	6 5 5 7	20	316	727,179 (20 fps) +137,846 (1 fps)	1100 min	soccer field	40 × 20 m^2 40 × 20 m^2 68.3 × 48.5 m^2 101.8 × 68.5 m^2

† The collection of the dataset is performed by the script. ‡ The dataset only provides annotations for detection.
* The dataset is not open to all researchers.

DukeMTMC [38] was not included in Table 1, because the dataset's camera field of views do not strictly overlap, with only two cameras having a slight overlap.

The Federal Institute of Technology in Lausanne [21] pioneered the creation of the multi-view datasets for multi-person detection and tracking, including Laboratory Terrace and Passageway. Campus [32] incorporated four sequences into its dataset, including Garden, Auditorium, and Parking Lot. CMC [31] focuses on walking people, with five different sequences featuring varying levels of person density and occlusion. SALSA [33] was gathered during a social event involving 18 individuals in a limited indoor environment. The participants wore sociometric badges equipped with microphones, accelerometers, Bluetooth, and infrared sensors, and the event was split into two sections of roughly equal length. However, the annotation provided by these six datasets has a low frame rate, making it hard to train and evaluate algorithms. Furthermore, the range of activities and the number of characters are limited, and the activity trajectories are pre-arranged, which is not representative of the actual motion of players on the sports field.

EPFL-RLC [35] is yet another multi-view dataset, captured using only three cameras. It offers a balanced mix of positive and negative multi-camera annotations, albeit within a limited field of view. The primary objective of the dataset is to determine if a particular location is occupied by a pedestrian or not. Originally, only a tiny portion of the 8000 total frames, specifically the last 300, were labeled for testing purposes.

The PETS-2009 [36] challenge aims to detect three types of crowd surveillance characteristics/events in a public outdoor scene. However, due to the broad scope of the challenge, only a portion of the dataset is devoted to people tracking. Additionally, the presence of a slope in the scene creates a calibration issue that can significantly degrade homography mapping, resulting in substantial inaccuracies in the projection of 3D points between views.

WildTrack [34] is the most widely used dataset with an overlapping camera setup. Similar to sports scenes, the multi-view camera focuses on a 36 × 12 m^2 rectangular area. The dataset annotates pure passengers without any scripted movements, with a larger number of annotated people and annotations compared to previous datasets. However, the dataset only includes one scene from a single time period, which limits its richness of data.

APIDIS [37] and LH0 [17] are among the few datasets that focus on sports scenes, and are of particular interest to us. These datasets were created by recording short clips of basketball and soccer training, respectively, using seven and eight static cameras. While the duration of these clips is very short, and they are not formal game videos, they still provide valuable information for the application of multi-view algorithms in sports scenes.

Due to the requirement for total automation, long-term stability, and precise player tracking, the models developed using the aforementioned datasets are insufficient. It is crucial to create a specialized dataset tailored to these specific needs.

2.2. Methods

Since 2000, there has been a significant body of literature on multi-view multi-object tracking (MVMOT), as documented in the citation [39]. In this paper, we provide a comprehensive summary of works related to various aspects of MVMOT.

2.2.1. Object Detection

Object detection is a critical component of MVMOT, with recent methods relying on deep learning models based on Convolutional Neural Networks (CNNs) [40–44]. Building on previous approaches such as YOLO [42], Kong et al. [45] introduced the Foveabox model, which employs deep learning to generate a robust description for candidate parts. Such anchor-free methods have gained popularity for object detection due to their ability to detect objects in images without the need for the two-stage R-CNN framework used by most other methods [46,47].

2.2.2. Feature Extraction

The goal of feature extraction is to generate a robust and discriminative description. In recent years, deep learning has been the basis for much of the research on object description, with numerous studies conducted [48–54].

Many approaches use local information to provide a robust, part-based description model for objects. The local part can offer additional information that helps describe the object more accurately. There are two categories of part generation, based on whether additional information is used: pose-based part generation [51,54] and unsupervised part generation [49,53]. In pose-based part generation, the person's part is localized based on the key points generated by existing pose estimation methods. In contrast, unsupervised part generation is an attention learning method that generates the local part using unsupervised part learning. Yao et al. [53] proposed an unsupervised part generation method for description based on the consistency of activation of neurons in the Convolutional Neural Network.

Global feature learning is a widely used method for representation learning, in addition to part-based methods. Some techniques [48,52] fine-tune networks on multi-target tracking data to produce robust visual descriptions. For instance, [52] applies the VGG architecture, converting its fully connected and softmax layers into convolution layers. Other research [55,56] seeks to fuse visual features with pose information to create distinctive features. Simultaneously, a challenge called Multi-Person PoseTrack [56] was proposed to combine pose estimation and tracking. For instance, Iqbal et al. [56] designed a spatio-temporal graph to solve pose estimation and tracking simultaneously. Additionally, some research (e.g., [55,57]) incorporates optical flow information into feature descriptions, with [55] utilizing the Siamese network to fuse pixel values and optical flow to obtain spatio-temporal features.

2.2.3. Data Association

Data association can be categorized as either intra-camera or cross-camera data associations.

Most existing methods address intra-camera data association through either tracklet-to-target or target-to-target matching. In [20,58], the LSTM network was used to obtain the historical tracklet features, while Bae et al. [18] proposed using the tracklet confidence

to measure the performance of each tracklet. Some other methods [19,24,26] apply the Kuhn–Munkres algorithm [59] to optimize the assignment problem. Zhou et al. [27] proposed a tensor-based high-order graph matching algorithm to flexibly integrate high-order information.

Various cross-camera data association methods have been proposed to merge single-camera trajectories across different cameras for cross-camera tracking. These methods include assignment methods based on trajectory features such as hierarchical clustering [22,32,60–63], greedy matching [28], multiple hypothesis tracking [25,64], K-shortest path [65], and camera topology [66]. Bredereck et al. [28] proposed a Greedy Matching Association (GMA) method that matches local tracklets obtained from different cameras one by one. Jiang et al. [66] took camera topology into consideration to reduce the influence of inconsistent appearance and spatio-temporal constraints in different cameras. In contrast to these tracklet-to-tracklet methods, other methods [67,68] solve the cross-camera tracklet matching problem using tracklet-to-target assignment. For instance, He et al. [68] focused on tracklet-to-target assignment and proposed the Restricted Non-negative Matrix Factorization algorithm to compute the optimal assignment solution. Graph-based approaches [68–72] based on the two-step framework have gradually been adopted by many researchers to solve the data associations across different frames and cameras. Gmcp-tracker [71] utilized Generalized Minimum Clique Graphs to solve the optimization problem of our data association method. Hypergraphs [69] introduced a combined maximum a posteriori (MAP) formulation and a flow graph, which jointly model multi-camera reconstruction as well as global temporal data association. Chen [72] developed a global approach by integrating these two steps via an equalized global graph model.

Recently, some outstanding works used a deep learning network to replace the graph-model to complete the data association. Reference [73] proposed a self-supervised learning framework to solve the MvMHAT problems. In References [74,75], the authors built novel networks and achieved end-to-end real-time online tracking. Finally, in [76–78], the researchers meticulously trained their networks to identify crowds and utilized self-supervision to effectively harness temporal consistency across frames. This allowed them to merge information from multiple perspectives while preserving the accurate location of individuals.

3. The System and The Datasets

3.1. Hardware and Data Acquisition

In order to continuously collect data, compared with the temporary equipment of the latest datasets [12,17,34], we built the multi-view video recording systems on four completely different fields according to different game types. This system can not only ensure that we collect high-quality and stable data, but also help reduce the difficulty of data post-processing.

1 + *N* based multi-view video recording system There is a significant difference between our dataset and other multi-view datasets. Our approach adopts the $1 + N$ strategy, which consists of 1 main camera capable of covering the entire field, complemented by N auxiliary cameras placed around the stadium for additional coverage. Providing this main camera has the following advantages. (i). Although most existing algorithms are currently evaluated based on a single-view metric, many fail to clearly describe how this metric is calculated or from which view the evaluation should be performed. To address this issue, we proposed a standardized evaluation scale based on the main camera view, enabling direct comparison of different algorithms and reducing ambiguity between them. (ii). Algorithm designers can easily assess potential failure points and identify problems that require solutions through visualization using the main camera view. (iii). A main camera view that captures the entire field not only makes it easier to visualize data but also meets the needs of data analysis. This allows coaches and players to intuitively understand the data without frequently switching views.

To effectively reduce the impact of occlusion on the vision algorithm, it is advisable to set up the camera at a high place. For larger fields, it becomes necessary to mount the camera higher. As a result, instead of positioning the cameras just above the average player height [17,34], we strived to make use of the existing facilities in the field and set up cameras from different angles as high as possible. Moreover, according to the $1 + N$ strategy, we ensured that players standing in any area of the field are covered by at least two cameras, providing clearer shots from multiple camera views. The camera layout is highly overlapping as shown in Figures A1–A4. The main cameras are identified as camera-1 in the figure.

Data acquisition. As shown in Table 1, the dataset we constructed is not only the largest known multi-view dataset with respect to video and annotation data, but it is also the first publicly available multi-view dataset recorded for soccer player tracking during real games. We selected 17 groups of video clips with dense annotations from real competitions, which included varying lighting conditions and weather, and covered almost all types of situations that occur in real games. A total of 316 players and referees from 34 different teams were involved in this dataset. The details of the dataset are described in Table 2 Furthermore, we have made available to all researchers a comprehensive multi-view video dataset of several games, totaling over 1100+ min.

Synchronization. The time synchronization difference between videos from different cameras in the same group was obtained with 50 microseconds $(1/fps)$ accuracy, the accuracy of which can be observed in Figure 2.

Figure 2. Synchronization of multi-cameras.

Based on this system, we have generated a public benchmark with high frame rate annotations, focusing on different challenging characteristics that occur during soccer scenarios. Through our data acquisition system and carefully labeled dataset, we hoped topromote the development of algorithms in this field and ultimately achieve the purpose of automatically localizing and tracking players.

3.2. Calibration of the Cameras

Since multi-view tracking relies on integrating information using unified world coordinates, it is crucial to have a reliable calibration method that provides stable and accurate data. We used surveillance-level zoomable cameras, but we were unable to locate the calibration parameters for our specific models in the equipment manual or on the official website. As a result, we utilized the pinhole camera model, as described in [79], to jointly calibrate the cameras for each field.

For a specific field, given a set of existing landmark points $\mathcal{W} = \{\mathbf{w}_i\}$, $\mathbf{w}_i = [x_i, y_i, z_i = 0] \in \mathbb{R}^3$, we manually annotated their corresponding pixel coordinates $\mathcal{M} = \{M^c\}$, $M^c = \{\mathbf{m}_i^c\}$, $\mathbf{m}_i^c = [x_i^c, y_i^c] \in \mathbb{R}^2$ on the image plane of each camera, more precisely, $\mathcal{C} = \{c_i\}$ denotes the number of cameras in this field.

Let $\mathbf{P} = \{\mathbf{P}^c\} = \{\{\mathbf{I}^c, \mathbf{E}^c\}\}$ denote the calibration parameters of each camera, where $\mathbf{I}$ and $\mathbf{E}$ denote the intrinsic and the extrinsic parameters, respectively. To this end, we

solved the projection relationship between $\mathcal{W}$ and $\mathcal{M} = \{M^c\}$ jointly through bundle adjustment [80] to achieve $\mathbf{P}$, as we explain in Equation (2).

$$\mathbf{m}_i^c = \mathbf{P}^c \cdot \mathbf{w}_i \tag{1}$$

$$\mathbf{P}^* = \underset{\mathbf{P},\mathcal{W},\mathcal{M}}{\arg\min} \sum_{c=1}^{|\mathcal{C}|} \sum_{i=1}^{|W|} \alpha_i^c \|\mathbf{m}_i^c - \mathbf{P}^c \cdot \mathbf{w}_i\| \tag{2}$$

$\|\cdot\|$ denotes the Euclidean distance in the image plane, and α_i^c is the indicator variable equal to 0 when the landmark point w_i is not available in view c, and equal to 1 when the landmark point w_i is visible. In other words, we have formulated this problem as a non-linear least squares approach. In this approach, the error is represented as the squared $L2$ norm of the difference between the projection points of the real world $3D$ points $\mathbf{w}$ on the image plane and their corresponding annotated pixel coordinates $\mathbf{m}^c$.

To improve the versatility and robustness of the calibration method, we exclusively utilized the original landmarks on the field to calibrate the camera. These landmarks comprised the penalty point, the kick-off point, the intersection of the marking line, the intersection of the boundary line between the deep and shallow turf and the marking line, etc.

For the 11-a-side set, camera-1 used four pinhole cameras to stitch the image plane, and therefore we calibrated each part of the image plane separately. Similarly, we used a similar method to calibrate the 8-a-side camera-1, which was stitched from two pinhole cameras.

Take the 11-a-side field as an example, Figure 3 illustrates the calibration precision of this method. (1) The field areas actually observed by each camera are sampled at an interval of 0.2 m to obtain $D^c = \{\mathbf{d}_i^c\}$, $\mathbf{d}_i^c = [x_i^d, y_i^d, z_i^d = 0] \in \mathbb{R}^3$. Then the projection results of D^c on each image plane, $J^c = \{\mathbf{j}_i^c\}$, $\mathbf{j}_i^c = [x_i^j, y_i^j] \in \mathbb{R}^2$ are calculated using Equation (1). We draw J^c on the image plane in the form of small red dots, as shown in Figure 3. (2) Based on landmark points W^c and manually annotated pixel coordinates M^c, we re-projected M^c to the world plane to obtain $K^c = \{\mathbf{k}_i^c\}$, $\mathbf{k}_i^c = [x_i^k, y_i^k, y_i^k = 0] \in \mathbb{R}^3$ using Equation (1). W^c and K^c were drawn on the world plane using big red dots and small blue dots, respectively.

Table 2. Video Statistics.

Game Type	Game ID	Cameras	Players	Length/Fps	Resolution (Main)	Resolution (Auxiliary)	Field Size	Split Type
5-A-Side Figure A1	1	6	16	120 s/20	3840 × 2160	3840 × 2160	40 × 20 m^2	train
	2	6	12	120 s/20	3840 × 2160	3840 × 2160	40 × 20 m^2	train
	3	6	13	120 s/20	3840 × 2160	3840 × 2160	40 × 20 m^2	train
	4	6	15	120 s/20	3840 × 2160	3840 × 2160	40 × 20 m^2	test
7-A-Side Figure A2	5	5	15	120s/20	3840 × 2160	3840 × 2160	44 × 35 m^2	train
	6	5	16	120 s/20	3840 × 2160	3840 × 2160	44 × 35 m^2	train
	7	5	16	120 s/20	3840 × 2160	3840 × 2160	44 × 35 m^2	train
	8	5	17	120 s/20	3840 × 2160	3840 × 2160	44 × 35 m^2	test
8-A-Side Figure A3	9	5	19	120 s/20	5120 × 1400	3840 × 2160	68.3 × 48.5 m^2	train
	10	5	20	120 s/20	5120 × 1400	3840 × 2160	68.3 × 48.5 m^2	train
	11	5	12	120 s/20	5120 × 1400	3840 × 2160	68.3 × 48.5 m^2	train
	12	5	12	120 s/20	5120 × 1400	3840 × 2160	68.3 × 48.5 m^2	train
	13	5	12	120 s/20	5120 × 1400	3840 × 2160	68.3 × 48.5 m^2	test
11-A-Side Figure A4	14	7	25	120 s/20	5950 × 1450	4000 × 3000	101.8 × 68.5 m^2	train
	15	7	25	120 s/20	5950 × 1450	4000 × 3000	101.8 × 68.5 m^2	train
	16	7	25	120 s/20	5950 × 1450	4000 × 3000	101.8 × 68.5 m^2	train
	17	7	25	120 s/20	5950 × 1450	4000 × 3000	101.8 × 68.5 m^2	test

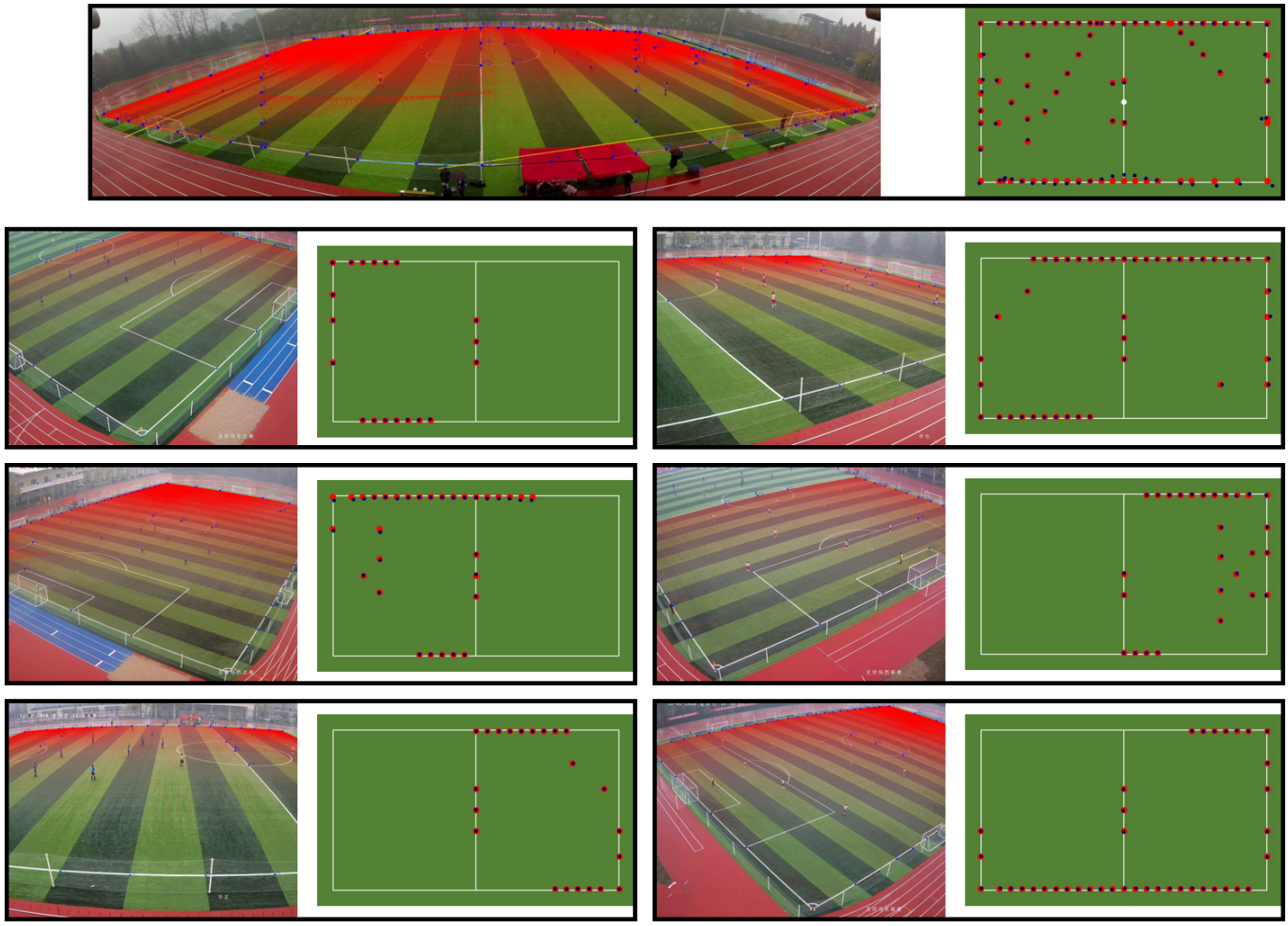


Figure 3. Visualization of the camera calibration precision. Details are illustrated in Section 3.2.

3.3. Annotations Process

$1 + N$ **MVMOT annotation strategy.** Manually labeling a multi-view multi-player tracking dataset is a time-consuming task.

Traditional labeling methods, as seen in other datasets listed in Table 1, employ an interval sampling approach for all camera views. This may lead to incomplete trajectory data during algorithm evaluation. However, labeling all views frame by frame would require a significant investment of time and resources. Our $1 + N$ MVMOT annotation strategy allows us to meet evaluation requirements while also saving on annotation costs. Annotating on the main camera view frame by frame (20 fps) is sufficient to obtain the complete trajectory of all players, as it provides full coverage of the field. At this point, training MVMOT algorithm models to perform tasks such as cross-view ReID matching and cross-view data association, only requires annotating the down-sampled frames of auxiliary camera views at 1 fps. Finally, an evaluation is conducted on the main view to ensure evaluation stability and reduce ambiguity.

More specifically, we firstly used the public annotation tool DarkLabel [81] to annotate each target frame by frame in the main camera views. Next, the pre-trained detector performed preliminary player detection on the auxiliary views. After that, we used self-designed annotation software to load the annotation data of the main camera view and the detection predictions of the auxiliary views. Finally, we adjusted the detection predictions of the auxiliary views and matched them to the corresponding identities in the main view. It is worth noting that the software also loads the calibration parameters of all cameras, which enables quick positioning and bounding box connections across all camera views. Further details of the self-designed software are described in Figure 4. Since annotating the multi-

view multi-player dataset requires annotators to mentally perform spatial transformations, these calibration parameters can significantly reduce annotation errors.

Figure 4. GUI of our self-designed multi-camera annotation software. The software allows annotators to quickly assign corresponding identities and adjust the size of detection predictions in auxiliary cameras by connecting them to bounding boxes in camera-1. (1) The bounding boxes in camera-1 are based on manual annotations, whereas the bounding boxes in other cameras are generated by the detector. (2) The purple lines serve to demarcate the target area [9]. (3) The blue bounding box has been adjusted to the correct size and aligned with the identities across multiple views. (4) The yellow bounding boxes are being adjusted and aligned. (5) The red bounding boxes are waiting to be adjusted and aligned. (6) The number above the bounding box indicates its identity. (7) The software utilizes calibration parameters to approximately locate the detection predictions in auxiliary camera views.

Each player in the dataset was assigned a unique identity, regardless of the number of times they appear. The objective of multi-view multi-player tracking is to identify the complete 3D trajectory of a particular player across a camera network. To provide a rigorous evaluation of the trackers, it was decided that these unique IDs would be used as the benchmark, which requires greater precision compared to standard tracking tasks.

3.4. Statistics

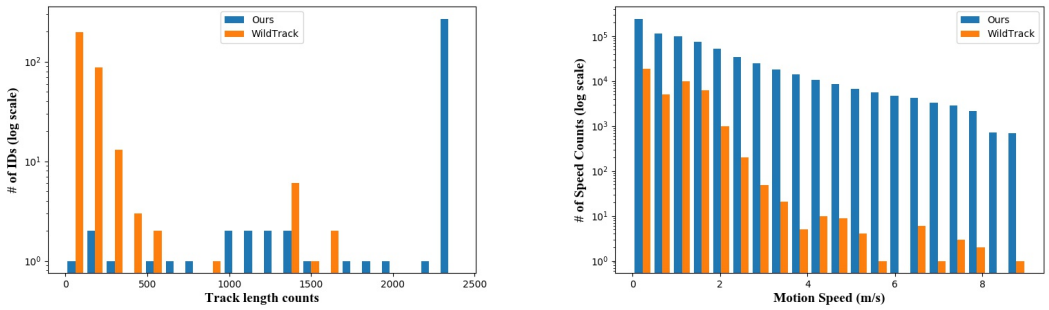
Annotated Frames and Videos. Our dataset, as shown in Table 1, is not only the closest to soccer scenarios, but also provides the largest amount of annotations. We have completed annotations for a total of 40,800 frames of data, with 2040 frames providing complete annotations from all camera views. For all main camera views, we have completed 727,179 identity annotations, and for all auxiliary camera views, we have completed a total of 137,846 identity annotations. These 865,025 annotations are an order of magnitude larger than the next largest dataset [34,37]. Additionally, a significant proportion of the videos are

unlabeled, and we will publicly release complete multi-view game videos that total over 1100 min. This will be of significant value for unsupervised methods and can serve as a benchmark for the annotated portion.

Field Size and Resolution. Multi-player sports, especially soccer, have high field size requirements. Table 2 shows that the size of the fields involved in our dataset ranges from 40×20 m^2 to 101.8×68.5 m^2, and the area where the players are active is much larger than other datasets. To capture sufficient detail, we used zoomable ultra-high-resolution cameras in all fields.

Track Length and Motion Speed. In Figure 5, we present a detailed analysis of our dataset and compare it to the WildTrack dataset [34], which is the standard dataset used for evaluating most state-of-the-art models.

In Figure 5a, we present the distribution of object track lengths, while Figure 5b shows the distribution of motion speed. Due to WildTrack's annotation frame rate of only 1 fps, we increased the data amount by interpolation. In WildTrack, the target length is generally shorter, and the motion speed is slower, mainly due to most pedestrians walking directly through the square, with only a few people standing still and chatting, making them relatively easier to track. In contrast, our dataset focuses on players on the field who move constantly throughout the clip, with more running and sprinting. Consequently, our dataset provides longer track lengths and faster speeds for the players, requiring tracking methods that can stably track all players for an extended duration.



(**a**) (**b**)

Figure 5. Comparison between our dataset and WildTrack dataset. (**a**) Track length comparison. (**b**) Motion speed comparison.

Data split. Our dataset was designed as an evaluation benchmark, so we divided it into two subsets—train and test. The training subset consists of 13 videos, while the testing subset contains 4 videos. This division was performed to ensure that the videos in each subset are distinct from each other. The trackers were trained using the videos in the training subset and then tested using the corresponding videos in the testing subset. For more information on how the subsets were constructed, refer to Column 9 of Table 2.

4. Benchmark

4.1. Evaluation Protocol

WildTrack [34] and other methods [28,61,74] adopted HOTA [82] and clear metrics [83,84] directly on the world plane. However, these metrics were designed for evaluation on the image plane. Directly implementing on the world plane will impair the fairness of these metrics. Since we had accurate camera parameters, we reported all performance on the main perspective plane for multi-camera tracking models. As described in NPSPT [9], we used HOTA [82], DetA, MOTA [84], IDF1, and IDSW as the evaluation metrics.

4.2. Evaluated Methods

For multiple object tracking, we provided results of the four state-of-the-art methods, so that we can better observed the challenges in our dataset in these evaluations

TRACTA [68] (tracking by tracklet-to-target assignment) treats multi-view multi-object tracking as a cross-camera tracklet matching problem. There are four major modules in the method: (1) local tracklet construction, (2) tracklet similarity calculation, (3) cross-camera tracklet connection, and (4) global trajectory construction. The TRACTA system employs the Restricted Non-negative Matrix Factorization (RNMF) algorithm to determine the optimal assignment solution that adheres to practical constraints. This enabled TRACTA to mitigate tracking errors caused by occlusions and missed detections in local tracklets and generate a comprehensive global trajectory for each target across all cameras.

MvMHAT [73] proposed a self-supervised learning approach to address the MVMOT challenge. The researchers started by selecting several frames from different views and using convolutional neural networks (CNNs) to extract the embedding features for each subject. Next, they introduced a spatial-temporal association network to consider both the temporal relationships over time and the spatial relationships across views. This generates a matching matrix which is self-supervised by symmetric similarity and transitive similarity losses. During the implementation stage, they utilized a novel joint tracking and association strategy to tackle the problem.

DAN4Ass [74] approached the MVMOT problem for $C > 2$ views as a constraint optimization challenge and created a novel end-to-end solution. This solution consists of two components: building the affinity matrix and assigning multi-view multi-cliques. The affinity network calculates the similarity between subjects detected in different views, while the Deep Assignment Network represents the multi-view constraints as differentiable loss functions during unsupervised training. DAN integrated four specific constraints into the model and combined image feature extraction, affinity matrix calculation, and assignment optimization into one unified framework for joint training.

MVFlow [76] proposed a weakly supervised approach to detect people flow given only detection supervision. The approach starts by incorporating a detection network that predicts people flow in a weakly supervised manner. The existing association algorithms are then modified to generate clear tracks using the predicted flows. This is followed by transforming two consecutive sets of multi-view frames into human flow using the proposed multi-view prediction model. The human flow is then utilized to rebuild detection heatmaps. The flow-based framework benefits from the temporal consistency across video frames and enforces consistency in scale and motion over various viewpoints.

To ensure the fairness of the experiment, we used the open source code provided by the authors to train, and used the unified evaluation method mentioned above to evaluate. After getting the results close enough to the results in the authors' papers, we used the same training methods to train and test on each subset of our dataset separately.

5. Results Analysis

Table 3 presents the tracking results of the four candidate trackers on different subsets of our dataset. The parameters of each method were optimized using the HOTA metric on the same training data as the corresponding detector. Our results reveal that the dataset poses a substantial tracking challenge, with lower HOTA metric scores compared to those observed in previous benchmark tests such as [34].

The reason why TRACTA has the worst tracking performance is because tracklets are generated within each view, and multi-view fusion is based on a manually set algorithm that has certain limitations in different scenes. On the other hand, the other three methods integrate information from multiple perspectives when generating detection. In the data association stage, MvMHAT relies on the similarity matrix to establish connections. However, players on the same team wear jerseys of the same color, making it difficult to distinguish them by their appearance. On the other hand, the data association methods adopted by DAN4Ass and MVFlow are more robust. Therefore, the comprehensive performance of these two methods is better.

Table 3. Comparison of the candidate methods on our dataset.

GameType	Algorithms	HOTA ↑	DetA ↑	MOTA ↑	IDF1 ↑	IDSW ↓
5-A-Side	TRACTA	57.42	72.77	90.12	70.30	155
	MvMHAT	58.87	71.65	91.27	71.97	143
	DAN4Ass	60.75	73.38	92.68	74.28	137
	MVFlow	62.52	75.95	93.08	74.11	124
7-A-Side	TRACTA	70.33	80.89	94.27	80.41	79
	MvMHAT	75.91	82.55	95.02	82.50	70
	DAN4Ass	77.52	82.29	95.26	83.97	51
	MVFlow	76.82	81.89	94.83	82.89	46
8-A-Side	TRACTA	67.33	80.26	90.99	77.71	101
	MvMHAT	68.80	80.98	91.57	80.75	89
	DAN4Ass	70.22	85.75	93.50	81.87	80
	MVFlow	72.57	87.31	94.89	83.75	76
11-A-Side	TRACTA	60.75	79.59	81.89	62.33	111
	MvMHAT	60.62	80.11	80.25	64.35	107
	DAN4Ass	63.27	82.86	79.75	63.71	102
	MVFlow	64.89	83.77	85.73	69.92	93

↑ indicates that a larger value is desirable; ↓ indicates that a smaller value is desirable.

To our surprise, 5-a-side games are not necessarily the simplest to track, and 11-a-side games are not necessarily the most challenging. The difficulty of tracking and detection is not solely determined by the number of players or the size of the image. Other factors, such as the camera setup, the proportion of the target within the image, and others, also play a role in influencing tracking performance.

The experiments demonstrate that the existing algorithms are far from perfect. There is still much work to be explored in this challenging scenario

6. Conclusions

The advancement of multi-view multi-object tracking methods for sports scenes is hindered by the limited availability of suitable datasets. To address this issue, we have developed multi-view recording systems and introduced a new dataset with high-quality and dense annotations. Our approach provides researchers with a rich and diverse set of data, allowing them to conduct more accurate and detailed analyses of the soccer scenarios. By leveraging our new dataset and recording systems, researchers can gain deeper insights into the behaviors and interactions of the players, ultimately leading to more advanced and impactful research in a range of fields.

Despite significant advances in computer vision and machine learning, our experiments revealed that the accuracy of four leading methods is not optimal, indicating that tracking in soccer and sports poses significant challenges to the research community. Addressing these challenges could have a significant impact on various areas, by enabling more precise and informative insights into player tracking.

Furthermore, the dataset is distinguished by a considerable volume of unlabeled video clips, in addition to a substantial amount of annotated data. This unique feature of the dataset makes it highly valuable for unsupervised learning techniques, as it presents the opportunity to evaluate the performance of such models on both the labeled and unlabeled portions of the dataset, enabling the potential discovery of new patterns and insights in the data.

To further automate game analysis, we recognize the limitations and challenges of our current system, which relies on commercial surveillance cameras. When strong backlighting occurs, it becomes difficult to obtain clear features of athletes. Additionally, the cameras are mounted at a high height and cover a large area, which makes it impossible to accurately recognize the players' postures and the movements of the football. To address these limitations, we plan to incorporate more optical sensors such as Lidar and UWB sensors into our recording system. This will allow us to obtain more comprehensive and in-depth information about the game. Additionally, we will release more labeled data that captures

the combination of human and ball movement. This richer and deeper dataset will enable machine learning algorithms to better understand the complexities of game situations.

As we continue to develop our system, we remain open to feedback from the research community regarding its effectiveness and limitations. We believe that this feedback will be crucial in helping us improve the quality of our data and the accuracy of our automated analysis. Ultimately, our goal is to create a more accurate and reliable system for automated game analysis that can benefit players, coaches, and researchers.

Author Contributions: W.H. and X.F. proposed the idea of the paper. T.G. helped manage the annotation group and helped clean the raw annotations. W.H. and T.G. designed the annotation software. W.H. conducted all experiments. W.H. and X.F. wrote the manuscript. X.S., X.Z. and J.E. revised and improved the text. X.F., S.H. and Y.S. are the people in charge of this project. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [The National Natural Science Foundation of China] grant number [62007029], [Key Projects of Major Humanities and Social Sciences of Zhejiang's Universities 2019-2020] grant number [2021QN043], [Zhejiang Province's "14th Five-Year Plan" graduate teaching reform project].

Institutional Review Board Statement: Our study did not require ethical approval.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Our dataset will be available at https://www.true-think.com (accessed on 31 March 2023).

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

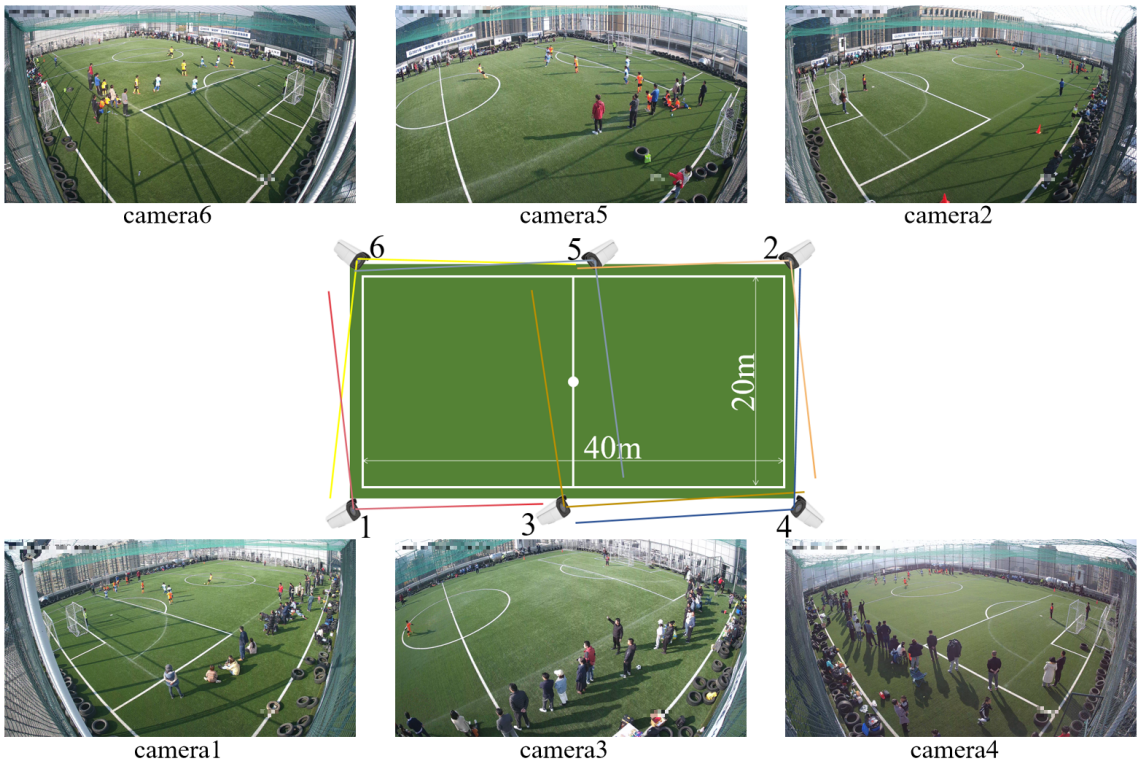


Figure A1. Camera layout of **5-a-side.** (1) Camera 1 is the main camera. (2) All these six cameras are installed 3 m away from the restricted area and are 6 m high.

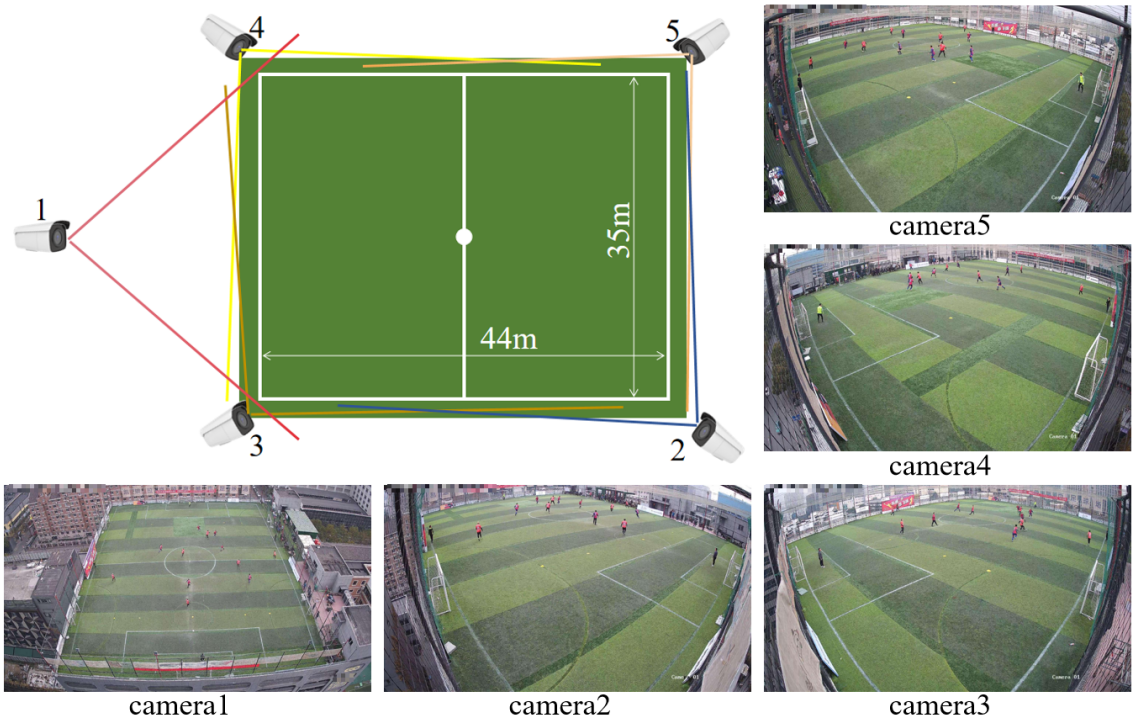


Figure A2. Camera layout of **7-a-side.** (1) Camera 1 is the main camera, it is installed 5 m away from the post line and 18 m high. (2) Cameras 2, 3, 4, and 5 are installed 0.5 m away from the corner point and are 6 m high.

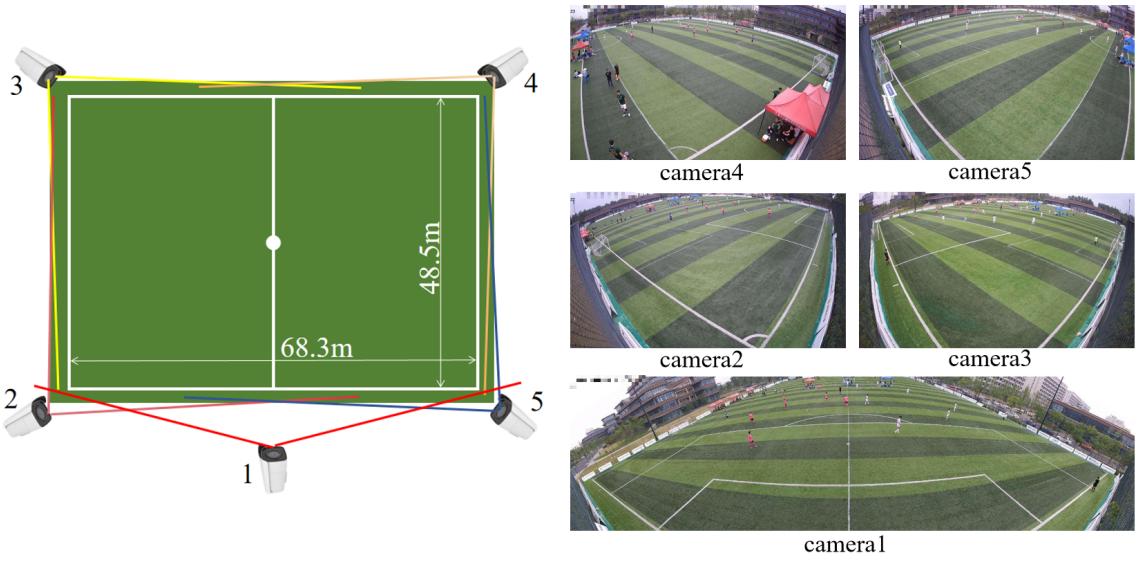


Figure A3. Camera layout of **8-a-side** (1) Camera 1 is the main camera, it is installed 1 m away from the middle-side-line and is 6 m high. (2) Cameras 2, 3, 4, and 5 are installed 1 m away from the corner point and are 6 m high.

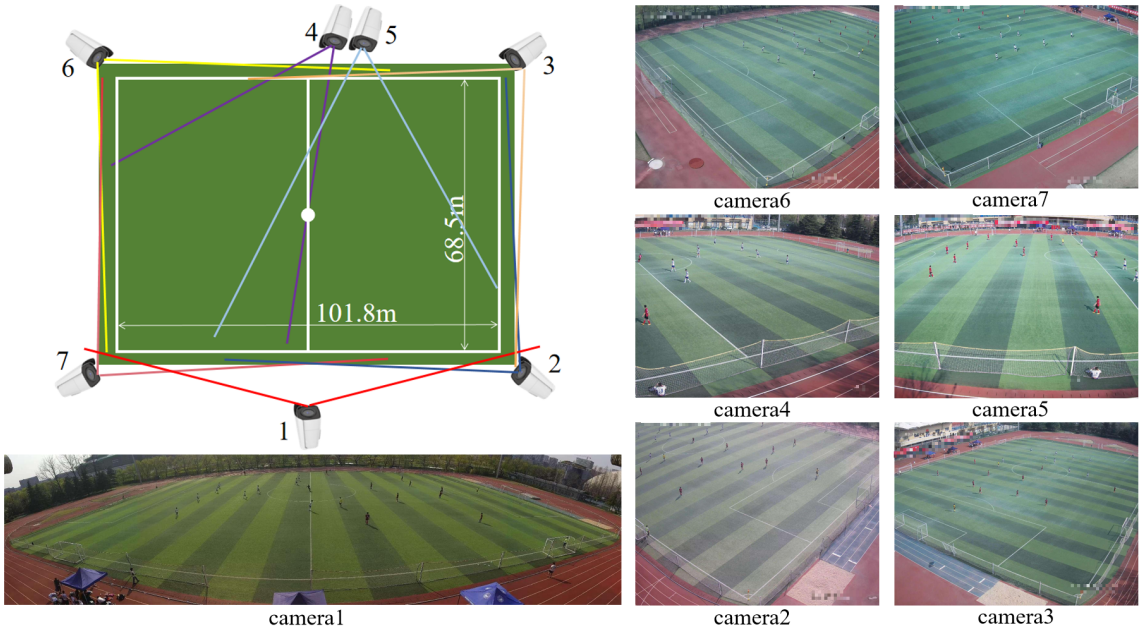


Figure A4. Camera layout of **11-a-side** (1) Camera 1 is the main camera, it is installed 15 m away from the middle-side-line and is 15 m high. (2) Cameras 2, 3, 6, and 7 are installed 15 m away from the corner point and are 30 m high. (3) Cameras 4 and 5 are installed 10 m away from the middle-side-line and are 12 m high.

References

1. Niu, Z.; Gao, X.; Tian, Q. Tactic analysis based on real-world ball trajectory in soccer video. *Pattern Recognit.* **2012**, *45*, 1937–1947. [CrossRef]
2. D'Orazio, T.; Leo, M. A review of vision-based systems for soccer video analysis. *Pattern Recognit.* **2010**, *43*, 2911–2926. [CrossRef]
3. Lu, X.; Wang, W.; Ma, C.; Shen, J.; Shao, L.; Porikli, F. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 3623–3632.
4. Wang, W.; Shen, J.; Porikli, F.; Yang, R. Semi-supervised video object segmentation with super-trajectories. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 985–998. [CrossRef] [PubMed]
5. Bornn, L.; Dan, C.; Fernandez, J. Soccer analytics unravelling the complexity of "the beautiful game". *Significance* **2018**, *15*, 26–29. [CrossRef]
6. Fernandez, J.; Bornn, L. Wide open spaces: A statistical technique for measuring space creation in professional soccer. In Proceedings of the Sloan Sports Analytics Conference, Boston, MA, USA, 23–24 February 2018; Volume 2018.
7. Narizuka, T.; Yamazaki, Y.; Takizawa, K. Space evaluation in football games via field weighting based on tracking data. *Sci. Rep.* **2021**, *11*, 1–8. [CrossRef]
8. Dave, A.; Khurana, T.; Tokmakov, P.; Schmid, C.; Ramanan, D. Tao: A large-scale benchmark for tracking any object. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 436–454.
9. Huang, W.; He, S.; Sun, Y.; Evans, J.; Song, X.; Geng, T.; Sun, G.; Fu, X. Open dataset recorded by single cameras for multi-player tracking in soccer scenarios. *Appl. Sci.* **2022**, *12*, 7473. [CrossRef]
10. Pappalardo, L.; Cintia, P.; Rossi, A.; Massucco, E.; Ferragina, P.; Pedreschi, D.; Giannotti, F. A public data set of spatio-temporal match events in soccer competitions. *Sci. Data* **2019**, *6*, 1–15. [CrossRef]
11. Voigtlaender, P.; Krause, M.; Osep, A.; Luiten, J.; Sekar, B.B.G.; Geiger, A.; Leibe, B. Mots: Multi-object tracking and segmentation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 7942–7951.
12. Geiger, A.; Lenz, P.; Urtasun, R. Are we ready for autonomous driving? The kitti vision benchmark suite. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 3354–3361.
13. Sun, P.; Cao, J.; Jiang, Y.; Zhang, R.; Xie, E.; Yuan, Z.; Wang, C.; Luo, P. Transtrack: Multiple object tracking with transformer. *arXiv* **2020**, arXiv:2012.15460.
14. Xu, Y.; Ban, Y.; Delorme, G.; Gan, C.; Rus, D.; Alameda-Pineda, X. Transcenter: Transformers with dense queries for multiple-object tracking. *arXiv* **2021**, arXiv:2103.15145.
15. 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.* **2021**, *129*, 3069–3087. [CrossRef]

16. Zhou, X.; Koltun, V.; Krähenbühl, P. Tracking objects as points. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 474–490.
17. Yang, Y.; Zhang, R.; Wu, W.; Peng, Y.; Xu, M. Multi-camera sports players 3d localization with identification reasoning. In Proceedings of the 25th International Conference on Pattern Recognition (ICPR), Milano, Italy, 10–15 January 2021; pp. 4497–4504.
18. Bae, S.-H.; Yoon, K.-J. Confidence-based data association and discriminative deep appearance learning for robust online multi-object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 595–610. [CrossRef] [PubMed]
19. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; pp. 3464–3468.
20. Fang, K.; Xiang, Y.; Li, X.; Savarese, S. Recurrent autoregressive networks for online multi-object tracking. In Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Tahoe, CA, USA, 12–15 March 2018; pp. 466–475.
21. Fleuret, F.; Berclaz, J.; Lengagne, R.; Fua, P. Multicamera people tracking with a probabilistic occupancy map. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *30*, 267–282. [CrossRef] [PubMed]
22. Li, P.; Li, G.; Yan, Z.; Li, Y.; Lu, M.; Xu, P.; Gu, Y.; Bai, B.; Zhang, Y.; Chuxing, D. Spatio-temporal consistency and hierarchical matching for multi-target multi-camera vehicle tracking. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 222–230.
23. Tang, S.; Andriluka, M.; Andres, B.; Schiele, B. Multiple people tracking by lifted multicut and person re-identification. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3539–3548.
24. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 3645–3649.
25. Yoon, K.; Song, Y.; Jeon, M. Multiple hypothesis tracking algorithm for multi-target multi-camera tracking with disjoint views. *Iet Image Process.* **2018**, *12*, 1175–1184. [CrossRef]
26. Yu, F.; Li, W.; Li, Q.; Liu, Y.; Shi, X.; Yan, J. Poi: Multiple object tracking with high performance detection and appearance feature. In Proceedings of the Computer Vision–ECCV 2016 Workshops, Amsterdam, The Netherlands, 8–10, 15–16 October 2016; Proceedings, Part II 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 36–42.
27. Zhou, Z.; Xing, J.; Zhang, M.; Hu, W. Online multi-target tracking with tensor-based high-order graph matching. In Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR), Beijing, China, 20–24 August 2018; pp. 1809–1814.
28. Bredereck, M.; Jiang, X.; Körner, M.; Denzler, J. Data association for multi-object-tracking-by-detection in multi-camera networks. In Proceedings of the 2012 Sixth International Conference on Distributed Smart Cameras (ICDSC), Hong Kong, China, 30 October–2 November 2012; pp. 1–6.
29. Hou, Y.; Zheng, L.; Gould, S. *Multiview Detection with Feature Perspective Transformation*; Computer Vision—ECCV 2020; Vedaldi, A., Bischof, H., Brox, T., Frahm, J., Eds.; Springer International Publishing: Cham, Switzerland, 2020; pp. 1–18.
30. Hsu, H.-M.; Huang, T.-W.; Wang, G.; Cai, J.; Lei, Z.; Hwang, J.-N. Multi-camera tracking of vehicles based on deep features re-id and trajectory-based camera link models. In Proceedings of the CVPR Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 416–424.
31. Ong, J.; Vo, B.; Vo, B.-N.; Kim, D.Y.; Nordholm, S. A bayesian filter for multi-view 3d multi-object tracking with occlusion handling. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *44*, 2246–2263. [CrossRef] [PubMed]
32. Xu, Y.; Liu, X.; Liu, Y.; Zhu, S.-C. Multi-view people tracking via hierarchical trajectory composition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4256–4265.
33. Alameda-Pineda, X.; Staiano, J.; Subramanian, R.; Batrinca, L.; Ricci, E.; Lepri, B.; Lanz, O.; Sebe, N. Salsa: A novel dataset for multimodal group behavior analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *38*, 1707–1720. [CrossRef]
34. Chavdarova, T.; Baqué, P.; Bouquet, S.; Maksai, A.; Jose, C.; Bagautdinov, T.; Lettry, L.; Fua, P.; Gool, L.V.; Fleuret, F. Wildtrack: A multi-camera hd dataset for dense unscripted pedestrian detection. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 5030–5039.
35. Chavdarova, T.; Fleuret, F. Deep multi-camera people detection. In Proceedings of the 16th IEEE International Conference on Machine Learning and Applications (ICMLA), Cancun, Mexico, 18–21 December 2017; pp. 848–853.
36. Ferryman, J.; Shahrokni, A. Pets2009: Dataset and challenge. In Proceedings of the 2009 Twelfth IEEE International Workshop on Performance Evaluation of Tracking and Surveillance, Snowbird, UT, USA, 7–9 December 2009; pp. 1–6.
37. De Vleeschouwer, C.; Chen, F.; Delannay, D.; Parisot, C.; Chaudy, C.; Martrou, E.; Cavallaro, A. Distributed video acquisition and annotation for sport-event summarization. *Nem Summit* **2008**, *8* , 1010–1016.
38. Ristani, E.; Solera, F.; Zou, R.; Cucchiara, R.; Tomasi, C. Performance measures and a data set for multi-target, multi-camera tracking. In Proceedings of the Computer Vision–ECCV 2016 Workshops,Amsterdam, The Netherlands, 8–16 October 2016; pp. 17–35.
39. Krumm, J.; Harris, S.; Meyers, B.; Brumitt, B.; Hale, M.; Shafer, S. Multi-camera multi-person tracking for easyliving. In Proceedings of the Third IEEE International Workshop on Visual Surveillance, Dublin, Ireland, 1 July 2000; pp. 3–10.
40. LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]

41. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision—ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
42. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
43. Zaidi, S.S.A.; Ansari, M.S.; Aslam, A.; Kanwal, N.; Asghar, M.; Lee, B. A survey of modern deep learning based object detection models. *Digit. Signal Process.* **2022**, *126*, 103514. [CrossRef]
44. Ouyang, W.; Wang, X. Joint deep learning for pedestrian detection. In Proceedings of the 2013 IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 2056–2063.
45. Kong, T.; Sun, F.; Liu, H.; Jiang, Y.; Li, L.; Shi, J. Foveabox: Beyound anchor-based object detection. *IEEE Trans. Image Process.* **2020**, *29*, 7389–7398. [CrossRef]
46. Song, T.; Sun, L.; Xie, D.; Sun, H.; Pu, S. Small-scale pedestrian detection based on topological line localization and temporal feature aggregation. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 536–551.
47. Zhang, S.; Wen, L.; Bian, X.; Lei, Z.; Li, S.Z. Occlusion-aware r-cnn: Detecting pedestrians in a crowd. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 637–653.
48. Chen, L.; Ai, H.; Shang, C.; Zhuang, Z.; Bai, B. Online multi-object tracking with convolutional neural networks. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 645–649.
49. Li, D.; Chen, X.; Zhang, Z.; Huang, K. Learning deep context-aware features over body and latent parts for person re-identification. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 384–393.
50. Li, Y.; Yao, H.; Duan, L.; Yao, H.; Xu, C. Adaptive feature fusion via graph neural network for person re-identification. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 2115–2123.
51. Su, C.; Li, J.; Zhang, S.; Xing, J.; Gao, W.; Tian, Q. Pose-driven deep convolutional model for person re-identification. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 3960–3969.
52. Sun, S.; Akhtar, N.; Song, H.; Mian, A.; Shah, M. Deep affinity network for multiple object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *43*, 104–119. [CrossRef]
53. Yao, H.; Zhang, S.; Hong, R.; Zhang, Y.; Xu, C.; Tian, Q. Deep representation learning with part loss for person re-identification. *IEEE Trans. Image Process.* **2019**, *28*, 2860–2871. [CrossRef]
54. Zhao, H.; Tian, M.; Sun, S.; Shao, J.; Yan, J.; Yi, S.; Wang, X.; Tang, X. Spindle net: Person re-identification with human body region guided feature decomposition and fusion. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp.1077–1085.
55. Leal-Taixé, L.; Canton-Ferrer, C.; Schindler, K. Learning by tracking: Siamese cnn for robust target association. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 33–40.
56. Iqbal, U.; Milan, A.; Gall, J. Posetrack: Joint multi-person pose estimation and tracking. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2011–2020.
57. Choi, W. Near-online multi-target tracking with aggregated local flow descriptor. In Proceedings of the 2015 IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 3029–3037.
58. Wan, X.; Wang, J.; Kong, Z.; Zhao, Q.; Deng, S. Multi-object tracking using online metric learning with long short-term memory. In Proceedings of the 25th IEEE International Conference on Image Processing (ICIP), Athens, Greece, 7–10 October 2018; pp. 788–792.
59. Kuhn, H.W. The hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [CrossRef]
60. Hou, Y.; Zheng, L.; Wang, Z.; Wang, S. Locality aware appearance metric for multi-target multi-camera tracking. *arXiv* **2019**, arXiv:1911.12037.
61. Ristani, E.; Tomasi, C. Features for multi-target multi-camera tracking and re-identification. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6036–6046.
62. Tesfaye, Y.T.; Zemene, E.; Prati, A.; Pelillo, M.; Shah, M. Multi-target tracking in multiple non-overlapping cameras using constrained dominant sets. *arXiv* **2017**, arXiv:1706.06196.
63. Zhang, Z.; Wu, J.; Zhang, X.; Zhang, C.a Multi-target, multi-camera tracking by hierarchical clustering: Recent progress on dukemtmc project. *arXiv* **2017**, arXiv:1712.09531.
64. Yoo, H.; Kim, K.; Byeon, M.; Jeon, Y.; Choi, J.Y. Online scheme for multiple camera multiple target tracking based on multiple hypothesis tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 454–469. [CrossRef]
65. Berclaz, J.; Fleuret, F.; Turetken, E.; Fua, P. Multiple object tracking using k-shortest paths optimization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1806–1819. [CrossRef]
66. Jiang, N.; Bai, S.; Xu, Y.; Xing, C.; Zhou, Z.; Wu, W. Online inter-camera trajectory association exploiting person re-identification and camera topology. In Proceedings of the 26th ACM International Conference on Multimedia, Seoul, Republic of Korea, 22–26 October 2018; pp. 1457–1465.

67. He, Y.; Han, J.; Yu, W.; Hong, X.; Wei, X.; Gong, Y. City-scale multi-camera vehicle tracking by semantic attribute parsing and cross-camera tracklet matching. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Virtual, 14–19 June 2020; pp. 576–577.
68. He, Y.; Wei, X.; Hong, X.; Shi, W.; Gong, Y. Multi-target multi-camera tracking by tracklet-to-target assignment. *IEEE Trans. Image Process.* **2020**, *29*, 5191–5205. [CrossRef]
69. Hofmann, M.; Wolf, D.; Rigoll, G. Hypergraphs for joint multi-view reconstruction and multi-object tracking. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3650–3657.
70. Jiang, X.; Rodner, E.; Denzler, J. Multi-person tracking-by-detection based on calibrated multi-camera systems. In Proceedings of the International Conference on Computer Vision and Graphics, Moscow, Russia, 1–5 October 2012; Springer: Berlin/Heidelberg, Germany, 2012; pp. 743–751.
71. Zamir, A.R.; Dehghan, A.; Shah, M. Gmcp-tracker: Global multi-object tracking using generalized minimum clique graphs. In Proceedings of the European Conference on Computer Vision, Florence, Italy, 7–13 October 2012; pp. 343–356.
72. Chen, W.; Cao, L.; Chen, X.; Huang, K. An equalized global graph model-based approach for multicamera object tracking. *IEEE Trans. Circuits Syst. Video Technol.* **2016**, *27*, 2367–2381. [CrossRef]
73. Gan, Y.; Han, R.; Yin, L.; Feng, W.; Wang, S. Self-supervised multi-view multi-human association and tracking. In Proceedings of the 29th ACM International Conference on Multimedia, Virtual, 20–24 October 2021; pp. 282–290.
74. Han, R.; Wang, Y.; Yan, H.; Feng, W.; Wang, S. Multi-view multi-human association with deep assignment network. *IEEE Trans. Image Process.* **2022**, *31*, 1830–1840. [CrossRef]
75. You, Q.; Jiang, H. Real-time 3d deep multi-camera tracking. *arXiv* **2020**, arXiv:2003.11753.
76. Engilberge, M.; Liu, W.; Fua, P. Multi-view tracking using weakly supervised human motion prediction. In Proceedings of the 2023 IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–7 January 2023; pp. 1582–1592.
77. Liu, W.; Salzmann, M.; Fua, P. Estimating people flows to better count them in crowded scenes. In Proceedings of the European Conference on Computer Vision, Virtual, 23–28 August 2020; pp. 723–740.
78. Tanaka, Y.; Iwata, T.; Kurashima, T.; Toda, H.; Ueda, N. Estimating latent people flow without tracking individuals. In Proceedings of the 32nd International Joint Conference on Artificial Intelligence, Macao, China, 19–25 August 2023; pp. 3556–3563.
79. Zhang, Z. A flexible new technique for camera calibration. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 1330–1334. [CrossRef]
80. Zach, C. Robust bundle adjustment revisited. In Proceedings of the European Conference on Computer Vision, Zurich, Switzerland, 5–12 September 2014; pp. 772–787.
81. Dark Programmer. Dark Programmer. Darklabel. Available online: https://github.com/darkpgmr/DarkLabel (accessed on 11 February 2023).
82. 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. *Int. J. Comput. Vis.* **2021**, *129*, 548–578. [CrossRef]
83. Milan, A.; Leal-Taixé, L.; Reid, I.; Roth, S.; Schindler, K. Mot16: A benchmark for multi-object tracking. *arXiv* **2016**, arXiv:1603.00831.
84. Luo, W.; Xing, J.; Milan, A.; Zhang, X.; Liu, W.; Kim, T.-K. Multiple object tracking: A literature review. *Artif. Intell.* **2021**, *293*, 103448. [CrossRef]