SoccerNet Game State Reconstruction: End-to-End Athlete Tracking and Identification on a Minimap

Vladimir Somers[1,3,8]* Victor Joos[1]* Anthony Cioppa[2,4]* Silvio Giancola[4]*
Seyed Abolfazl Ghasemzadeh[1] Floriane Magera[2,7] Baptiste Standaert[1] Amir M. Mansourian[5]
Xin Zhou[6] Shohreh Kasaei[5] Bernard Ghanem[4]
Alexandre Alahi[3] Marc Van Droogenbroeck[2] Christophe De Vleeschouwer[1]
[1] UCLouvain [2] University of Liège [3] EPFL [4] KAUST [5] SUT [6] Baidu Research [7] EVS [8] Sportradar

Abstract

Tracking and identifying athletes on the pitch holds a central role in collecting essential insights from the game, such as estimating the total distance covered by players or understanding team tactics. This tracking and identification process is crucial for reconstructing the game state*, defined by the athletes' positions and identities on a 2D top-view of the pitch, (*i.e. *a* minimap*). However, reconstructing the game state from videos captured by a single camera is challenging. It requires understanding the position of the athletes and the viewpoint of the camera to localize and identify players within the field. In this work, we formalize the task of Game State Reconstruction and introduce* SoccerNet-GSR*, a novel Game State Reconstruction dataset focusing on football videos.* SoccerNet-GSR *is composed of* 200 *video sequences of* 30 *seconds, annotated with* 9.37 *million line points for pitch localization and camera calibration, as well as over* 2.36 *million athlete positions on the pitch with their respective role, team, and jersey number. Furthermore, we introduce* GS-HOTA*, a novel metric to evaluate game state reconstruction methods. Finally, we propose and release an end-to-end baseline for game state reconstruction, bootstrapping the research on this task. Our experiments show that GSR is a challenging novel task, which opens the field for future research. Our dataset and codebase are publicly available at* https://github.com/SoccerNet/sn-gamestate.

1. Introduction

Recently, sports companies and teams have shown a growing interest in collecting athlete-centric data. One key focus area lies in tracking and identifying athletes on the sports field throughout the entire game, using available video footage. These analytics hold immense value for a broad spectrum of sports applications, ranging from **(i)** supporting team coaching and athlete training, **(ii)** assisting scouters in discovering new talents, **(iii)** offering valuable insights for medical staff, and **(iv)** boosting fan engagement through personalized content creation [11, 12, 27].

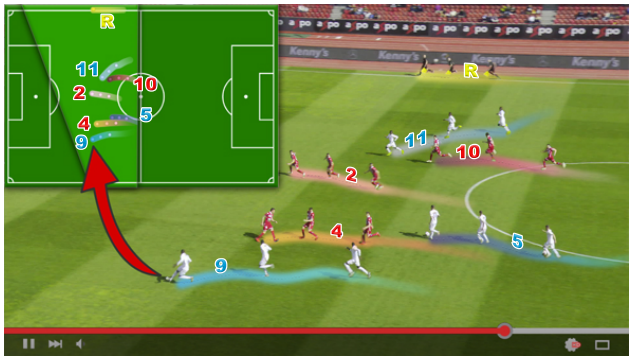


Figure 1. **SoccerNet-GSR.** We introduce a novel Game State Reconstruction task, dataset, evaluation metric and baseline. Our SoccerNet-GSR dataset contains unique identifications for players along with their localization on the pitch, for 200 video sequences.

However, the manual generation of such data by human annotators is time-consuming and costly. Sensor-based solutions offer a time-efficient alternative, but require athletes to wear special, sometimes expensive, equipment. Recently, automatic solutions based on optical tracking systems have gained prominence. These systems necessitate the installation of sophisticated, well-calibrated static multi-camera setups in stadiums. Hence, they come with significant drawbacks in terms of cost and scalability, which restricts their use to elite competitions, exemplified by their deployment at events like the 2022 Qatar World Cup.

Meanwhile, recent progress in computer vision opened up a growing potential to automatically and reliably extract athlete localization and identification data solely from broadcast camera feeds. In line with this objective, Multi-Object Tracking (MOT) methods have long been popular

(*) Equal contributions. Data/code available at www.soccer-net.org.

for sports video analysis. However, they offer only a partial solution to the aforementioned requirements. Indeed, the bounding-box-based tracking data produced by MOT (1) lacks critical identification information necessary to analyze specific athletes and (2) lacks interpretability due to the absence of grounding in a real-world coordinate system. These significant limitations hinder the usability of such tracking data for many downstream sports applications.

To address the above limitations, we introduce the concept of **G**ame **S**tate **R**econstruction (**GSR**), a novel computer vision task tailored for sports analytics. GSR aims to recognize the state of a sports game by identifying and tracking all athletes on the pitch based on input videos captured by a single camera. Moreover, game state data can be visualized in a minimap of the game, as depicted in Fig. 1, offering a concise representation of the ongoing gameplay dynamics. To support research on this task, we publicly release *SoccerNet-GSR*, the first dataset for Game State Reconstruction, consisting of 200 30-second fully annotated clips. Our proposed GSR annotations include over 9.37 million line points for football pitch registration, and over 2.36 million athlete positions on the pitch with unique identification information, including their role, team, and jersey number. Since existing metrics for Multi-Object Tracking [5, 54] are not suited for our proposed task, we introduce the *GS-HOTA*, a new evaluation metric to benchmark GSR methods. Finally, we propose *GSR-Baseline*, the first end-to-end and open-source pipeline for game state reconstruction, built upon state-of-the-art tracking, re-identification, team affiliation, jersey number recognition, pitch localization, and camera calibration methods. Our analysis underscores the complexity of Game State Reconstruction and highlights the importance of introducing this new benchmark. This initiative establishes an ideal platform for future research in the field, aiming to democratize access to this valuable game state data for all leagues.

Contributions. We summarize our contributions as follows. **(i)** We introduce and concretely define the concept of *Game State Reconstruction*, a task aiming to track and identify all athletes on a minimap of the pitch. **(ii)** We publicly release **SoccerNet-GSR**, the first open-source sports video dataset for Game State Reconstruction. **(iii)** We introduce **GS-HOTA**, a new metric to evaluate game state reconstruction methods. **(iv)** We propose **GSR-Baseline**, the first end-to-end GSR pipeline for football videos.

2. Related Work

Game State Reconstruction relates to the general topic of sports video understanding and, more particularly, to tracking, identification, and sports field registration.

Sports Video Understanding. Sports video understanding has emerged as a prominent research topic over the past decade [60, 62, 82]. Some works focused on low-level semantics, aiming to build a bottom-up understanding of the game [14], such as segmenting [15] or detecting [65, 69, 73, 84] players and keypoints [30, 53]. Recent advances in computer vision allowed for a higher semantic understanding of the game, focusing for instance on the action spotting task, aiming to spot a series of events during the game [9, 16, 32, 34, 36, 76, 87, 96, 97]. Fortunately, these works can rely on the availability of large-scale datasets [19, 20, 23, 35, 41, 59, 64, 72, 91] and challenges [21, 33, 40, 45, 58, 83]. Our novel Game State Reconstruction task stands in between low- and high-level semantics, providing both local information about the players but also global information about the whole state of the game through time. This information can later be used to better understand player actions [18], enhance the generation of engaging captions [10, 59], improve visualizations [8, 28, 70, 98], or derive high-level analytics [1, 3, 22, 48, 63, 66]. In this work, we complement the literature in sports video understanding by proposing a novel task of Game State Reconstruction that aggregates several tasks ranging from field to player understanding.

Player Tracking and (Re-)Identification. Multiple Object Tracking (MOT) has often been approached through the tracking-by-detection paradigm [4, 6, 7, 79, 81, 86, 94, 95]. However, applying the tracking-by-detection paradigm to sports introduces unique challenges compared to generic scenarios. Previous works [17, 20, 37, 68, 75, 84, 89] tackled the challenges of similar appearances and fast motion of people and object in sports. Furthermore, unlike generic MOT scenarios, athletes come in and out of the camera view, requiring long-term Re-Identification (ReID) [31, 46, 57, 92, 93]. Finally, uniquely identifying actors in a sports scene has been widely investigated in the literature. Some approaches focused on athletes' role (*e.g.*, player, referee, coach, *etc.*) [19, 57, 84], their team [39, 57], or jersey numbers [2, 29, 51, 52, 61, 85, 90]. Different from previous works, our new game state reconstruction task combines athlete tracking and identification, including the role, team, and jersey number under a single task.

Sports Field Registration. Mapping the video tracking data into a real-world coordinate system requires camera calibration. Sports games come naturally with a coordinate system based on the sports pitch. Hence, combining the location of the field [71, 89] with video camera calibration [26, 67, 74, 80], one can reconstruct a game state as illustrated in Fig. 1. Unifying tracking and camera calibration as proposed in this paper has been investigated in previous works.[18, 44, 72, 78] Scott *et al.* [72] collected data from fish-eye camera, drone, and GNSS, while Karungaru *et al.* [44] focused on the mapping of players onto the field in one video frame. Cioppa *et al.* [18] leveraged tracking and camera calibration to reproject players' posi-

tions on the pitch for the task of action spotting. Maglo *et al*. [56] introduced a robust player tracking method, incorporating test-time fine-tuning and a novel football field registration technique, which were combined to explore player localization on a minimap. However, due to the lack of annotations, they did not perform either player identification or quantitative evaluations of their localization results. Finally, Theiner *et al*. [81] introduced a pipeline to localize players on a pitch minimap from broadcast videos but omitted player identification and tracking. Different from previous work, our proposed GSR benchmark addresses the combined athlete pitch localization and identification task.

3. Game State Reconstruction Task

Game State Reconstruction (**GSR**) is a form of video compression task aiming to extract high-level information about the dynamics of a sports game from an input video. It includes **(1)** the 2D position of all athletes on the sports pitch, **(2)** their roles in the game (e.g., "player", "goalkeeper", or "referee"), and **(3)**, for players, their jersey number and team affiliation. This information can be visualized on a 2D top-view of the pitch, or minimap, as illustrated in Fig. 1. In the following, we refer to all individuals to be identified and localized, irrespective of their specific roles, as "**athletes**". GSR is a multifaceted task that requires addressing various intricate sub-tasks, including: **(a)** pitch localization and camera calibration, **(b)** athlete detection, re-identification, and tracking, and **(c)** role classification, team affiliation, and jersey number recognition.

We formalize the Game State Reconstruction task as follows. Given a team sports video composed of T frames, the objective is to predict a set of detections d_i^t for each frame t, where i indexes the detections within frame t. A detection encapsulates each athlete's location on the pitch ($pitch\_x$, $pitch\_y$) in a real-world coordinate system, and their *role*, *team*, and *jersey number*. A detection is therefore represented as follows:

$$d_i^t = \{\underbrace{pitch\_x, pitch\_y}_{\text{localization}}, \underbrace{role, team, jersey\_number}_{\text{identification}}\}. \quad (1)$$

While our main focus is football, the definition of the GSR task can extend to other team sports.

4. SoccerNet-GSR Dataset

Our dataset expands upon SoccerNet-Tracking [20], which consists of 200 30-second clips split into train, validation, test, and a segregated challenge set. In the original dataset, each frame includes bounding box annotations for the localization of players, referees, and balls tracked over time with extra role, team, and jersey number attributes. Despite the comprehensive annotations, SoccerNet-Tracking lacks information like pitch localization, camera calibration[1], and athlete positions on the pitch, critical for the Game State Reconstruction task. In subsequent sections, we detail how we augmented the SoccerNet-Tracking annotations to create our proposed SoccerNet-GSR dataset. The new annotations now include over 9.37 million line points for pitch localization and camera calibration, as well as over 2.36 million athlete positions on the pitch with their respective role, team, and jersey number. Since the SoccerNet-GSR videos are uncut broadcast sequences captured by a single moving camera, only a portion of the football pitch is visible at any given time. As a result, the GSR task is limited to players within the camera's field of view.

4.1. Athlete Localization on the Pitch

Expressing the 2D image location of athletes in the real-world pitch coordinates requires pitch localization and camera calibration. Together, these information enable precise mapping of player positions from the image onto the pitch. Our new annotations described in this section therefore include: (1) pitch 2D positions, (2) camera parameters, and (3) positions on the pitch.

Pitch localization. Following the same procedure as SoccerNet-v3 [19], we manually annotate every line on the football pitch by placing a series of points along its length to accurately define its shape, including curves such as the circles or the ones due to camera distortions. We categorize each line (*e.g*., side line left, side line top, *etc*.) and part of the goals, (left and right posts and the crossbar), totaling 26 classes. Next, we continuously track all these annotations over time using key frame annotations and interpolations in-between when it is appropriate, mirroring the player tracking data as described in [20], resulting in a densely marked pitch annotation throughout the entire video. This annotation process is core for calibrating the camera through time.

Camera calibration. Camera calibration is the process of determining the camera parameters for each frame, allowing to link the image-plane to the 3D world. It is required to compensate for the lack of a pre-calibrated camera. Usually, this process requires correspondences between a known 3D object and its image. In the context of football, the pitch is a convenient object [38] to obtain correspondences from. In this work, we assume that the pitch has a conventional size of 105 by 68 meters. However, as the pitch is only partially visible in the images, the calibration of broadcast cameras is a challenging task. Hence, due to the lack of visible lines, some frames may not be calibrated correctly and are discarded in the evaluation. For the frames presenting a sufficient amount of pitch line annotations, the camera pa-

[1] A camera calibration and pitch localization dataset was already introduced for the SoccerNet Camera Calibration challenge, but the corresponding annotations were provided on a separate set of data.

rameters are obtained from the best of several open-source techniques [13, 55] or an industrial tool [25]. The complete process is described in the supplementary materials.

Position on the pitch. The point of calibrating the cameras is to derive positions in the real-world. Our 3D world reference axis system is centered on the pitch center mark, the X-axis points to the right goal, the Y-axis follows the middle line towards the camera and the Z-axis is perpendicular to the XY – or the pitch – plane, pointing towards earth's center. Once the camera parameters are known, the inverse of the camera projection function applied to a point gives a 3D ray that can be intersected with the pitch plane to derive the 3D position. We assume that the athlete's feet, and more specifically the center of the bottom part of their detection bounding boxes lies on the pitch. Unfortunately, this approximation limits the precision of the estimated locations in the case of jumps. Hence, we remove the ball as it spends significant time in the air. A precise 3D localization of all elements would require tracking hardware, which is unavailable for open-science research at the moment.

4.2. Athlete Identification

To identify athletes during a game, we leverage three distinct manual annotations provided in the SoccerNet-tracking dataset that have been previously overlooked in standard multi-object tracking: *role*, *team*, and *jersey number*. The following paragraphs detail each annotation and the utilization of an additional *track id* for cases where targets cannot be uniquely identified by their attributes.

Role. In the SoccerNet-GSR dataset, athletes are categorized into four distinct roles during the game: 'player', 'goalkeeper', 'referee', or 'other'. The 'other' role encompasses individuals entering the pitch, such as coaches, medical staff, and any additional person not falling into the previous three categories. For the first version of the SoccerNet-GSR benchmark, both referee responsibilities (i.e. main, bottom/top assistants) and ball detections are ignored.

Team. Detections with the 'player' and 'goalkeeper' roles are annotated with a 'team' attribute, which can be assigned one of two values: 'left' or 'right'. Since the dataset consists of 30-second sequences captured from a single camera, we determine the 'left' and 'right' teams based on their goal's position relative to the camera viewpoint.

Jersey Number. Players and goalkeepers in the SoccerNet-GSR dataset are annotated with an additional 'jersey number' attribute. However, unlike the role and team attributes, which are always available, a jersey number may not be visible at any point during the entire 30-second video sequence. In such cases, players with invisible shirt numbers are assigned a 'null' value for this attribute. If a player's jersey number is visible in at least one frame of the sequence, then the entire tracklet is annotated with that jersey number. Therefore, a jersey number assigned to a detection does not necessarily mean that it is visible in that particular frame.

Track Id. We utilize the combination of role, team, and jersey number attributes to identify each athlete during a game. However, athletes cannot always be uniquely identified by their attributes. This occurs, for example, when two players from the same team do not have visible jersey numbers or when multiple individuals with the role of 'referee' or 'other' appear simultaneously. Although these cases represent a small proportion of all annotated athletes, they prevent unique identification using attributes alone. To address this, we also include the standard 'track id' annotation from standard MOT. This requires methods for the SoccerNet-GSR task to output four values per detection: role, team, jersey number, and track id. The impact of non-uniquely identifiable targets is further discussed in Sec. 5.

5. GS-HOTA Evaluation Metric

Game State Reconstruction (GSR) is a novel computer vision task closely related to multi-object tracking (MOT). Yet, standard evaluation metrics for MOT, such as MOTA [5] and HOTA [54], cannot be used to evaluate GSR for two main reasons. First, these metrics do not account for additional attributes predicted on the tracked targets, such as team, role, and jersey numbers. Second, these metrics rely on an IoU score to match predicted and ground truth bounding boxes in the image space, while GSR operates on 2D points within the pitch coordinate system.

To address these issues, we introduce *GS-HOTA*, a novel evaluation metric to measure the ability of a GSR method to correctly track and identify all athletes on the sports pitch. GS-HOTA is derived from the HOTA [54] metric, which is formulated as follows:

$$\text{HOTA} = \int_{0<\alpha\leq 1} \sqrt{\text{DetA}_\alpha \cdot \text{AssA}_\alpha} \tag{2}$$

DetA/*AssA* are the detection/association accuracy respectively, and α is a similarity threshold. To compute these two underlying accuracy metrics, ground truth and predicted detections must first be matched according to a similarity score. Pairs with a similarity score below the α threshold are not matched. For predictions (P) and ground truth (G) represented as bounding boxes in the image space, the Intersection-over-Union (IoU) is employed as the similarity score for the HOTA metric. The key distinction setting GS-HOTA apart from HOTA is the use of a new similarity score, that accounts for the specificities of the GSR task, i.e. the additional target attributes (jersey number, role, team) and the detections provided as 2D points instead of bounding boxes. This new similarity score, denoted $Sim_{GS-HOTA}(P, G)$, is formulated as follows:

$$Sim_{\text{GS-HOTA}}(P, G) = \text{LocSim}(P, G) \times \text{IdSim}(P, G), \tag{3}$$

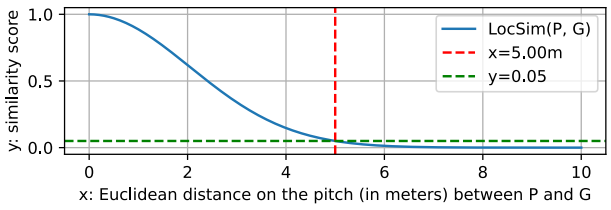


Figure 2. The localization similarity function for $\tau = 5$ meters.

$$\text{with LocSim}(P, G) = e^{\ln(0.05)\frac{\|P-G\|_2^2}{\tau^2}}, \tag{4}$$

$$\text{and IdSim}(P, G) = \begin{cases} 1 & \text{if all attributes match,} \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

$Sim_{\text{GS-HOTA}}$, is therefore a combination of two similarity metrics. The first metric, the localization similarity $\text{LocSim}(P, G)$, computes the Euclidean distance $\|P - G\|_2$ between prediction P and ground truth G in the pitch coordinate system. This distance is subsequently processed using a Gaussian kernel with a special *distance tolerance parameter* τ, resulting in a final score falling within the $[0, 1]$ range. The second metric, the identification similarity $\text{IdSim}(P, G)$, is set to one only if all attributes match, i.e. role, team, and jersey numbers. Attributes not provided in G are ignored, e.g. jersey numbers for referees[2]. Finally, once P and G are matched, DetA and AssA are computed and integrated into a final GS-HOTA score, following the original formulation of the HOTA metric in Eq. (2).

5.1. GS-HOTA Distance Tolerance Parameter

Our GS-HOTA metric relies on a single τ parameter introduced in Eq. (4). In practice, the continuous integral in Eq. (1) is computed over a discrete interval $\alpha \in [0.05, 0.95]$ with 0.05 steps. This means that (P, G) pairs with a similarity below or equal to 0.05 are never matched. Hence, our distance tolerance parameter τ defines the maximum distance in meters for a prediction P and a ground truth G to be matched, as illustrated in Fig. 2. Furthermore, since all similarity thresholds in the range $[0.05, 0.95]$ are considered, a distance smaller than τ meters between P and G still results in a higher GS-HOTA. This way, methods are still incentivized to produce athlete localization closer to the ground truth. In this work, we define τ as 5 meters, considering it a reasonable distance tolerance given the average dimensions of a soccer pitch (68×105 meters) and the substantial distance between the camera and the athletes.

[2]GSR methods must ignore the team and jersey number for non-player roles, as well as the jersey number when it is not visible in the video.

5.2. Motivation and Discussion

As introduced in Sec. 4.2, we consider the combination of attributes (role, team, jersey) as a way to identify athletes. If each person was uniquely identifiable by the combination of its attributes, association would become trivial, and as a consequence, a simple average of the Detection Accuracy across all identities would suffice as a robust performance metric. However, as explained in Sec. 4.2, not all identities in our SoccerNet GSR dataset can be uniquely identified by their attributes. Therefore, the Association Accuracy must also be taken into account to account for identity switches among athletes sharing the same attributes (e.g. players from the same team with no visible jersey number).

Finally, a key difference that sets GSR apart from MOT — and by extension, GS-HOTA from HOTA — is the necessity to identify athletes by their attributes. This requirement is specified by Eq. (5), according to which failing to correctly predict at least one attribute turns the corresponding detection into a False Positive. Requiring the correct prediction of all attributes simultaneously is a strict constraint, which we justify based on the severe impact that incorrectly assigning localization data to a nonexistent or incorrect identity can have on downstream applications.

6. GSR Baseline

In this section, we introduce the GSR-Baseline, a pipeline designed to reconstruct the game state of any broadcast football video. Our baseline splits the Game State Reconstruction task into several sub-tasks, selecting popular and open-source state-of-the-art methods for each sub-task. To facilitate the development of such a complex video processing pipeline, we leverage TrackLab [43], a research-oriented PyTorch-based framework for multi-object tracking. The overall architecture of the GSR-Baseline is depicted in Fig. 3, and a detailed description of each of the pipeline modules is provided hereafter.

6.1. Athlete Detection and Tracking

We employ a pre-trained **YOLOv8** [42] model as our athlete detector, without fine-tuning it on the SoccerNet dataset, since it already provides decent performance on football videos. We filter the model's output to retain only the "person" class detections. To leverage existing strong multi-object trackers, our GSR-Baseline performs tracking in the image space based on bounding boxes. As illustrated in Fig. 3, these bounding boxes are converted into 2D pitch positions later within the pipeline. Next, we employ **StrongSORT** [24] as our multi-object tracker, for its SOTA performance and its ability to leverage both spatio-temporal and appearance cues, the latter being provided by the re-identification model PRTreID [57] described in Sec. 6.3.

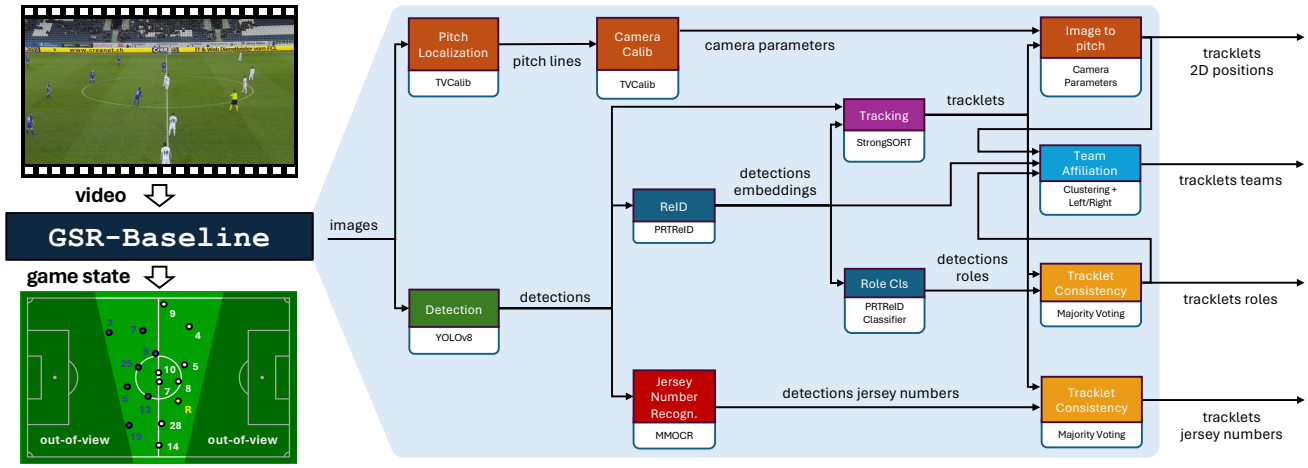


Figure 3. **Architecture overview of our proposed baseline.** GSR-Baseline takes a video as input and outputs the complete game state. Two modules are first applied on the input images: an object detector and a pitch localization model. Then, PRTreID [57] produces a ReID embedding for each detection, that is identity, team, and role aware. These embeddings are then forwarded to subsequent modules to perform role classification, team affiliation, and multi-object tracking. Finally, the pitch localization output is used for camera calibration, which enables the tracked bounding boxes to be transformed into 2D positions on the pitch coordinate system.

6.2. Pitch Localization and Camera Calibration

Camera calibration is performed using **TVCalib** [80], which is composed of two modules. The first module performs pitch localization through semantic segmentation. The second estimates the camera calibration parameters by iteratively minimizing the pitch segments reprojection errors. Once the camera has been calibrated, its corresponding homography is used to transform image bounding boxes into 2D positions on the pitch. For this purpose, we assume that the bottom of the bounding box lies on the ground field.

6.3. Athlete Identification

Athlete identification is performed by two key models: **PRTreID** [57] to produce team and role-aware ReID embeddings, and **MMOCR** for jersey number recognition. The output of these two models is further processed for tracklet consistency, team affiliation, and role classification, to produce the final game state identification data.

Re-Identification. The sportsperson representation model **PRTreID** [57] is designed to jointly solve person re-identification, role classification, and team affiliation with a single backbone. Therefore, it produces an embedding that is team, role, and identity discriminative, thanks to a multi-task learning setup with three learning objectives. PRTreID builds upon the SOTA part-based ReID method BPBreID [77]. During the PRTreID training procedure, re-identification and team affiliation are formulated as deep metric learning tasks, where persons with the same identity/team are pulled close to each other in the embedding space with a triplet loss. Role prediction is framed as a classification task with four target classes, employing a focal loss to address class imbalance. At inference in the GSR-Baseline pipeline, PRTreID produces an embedding for each input detection, that is forwarded to subsequent modules to perform tracking, team clustering with left/right labeling, and role classification.

Role Classification. The embeddings described above are processed by the **PRTreID** classification layer to output the target's role: *player*, *goalkeeper*, *referee*, or *other*.

Jersey Number Recognition. Jersey numbers recognition is performed in two separate steps with the open-source optical character recognition library **MMOCR** [47]. First, the YOLOv8 detections are fed to the **DBNet** [50] text detection model. Subsequently, the detected texts are forwarded to the **SAR** [49] text recognition model. Finally, the highest-scored detected text containing a number is considered as the player's jersey number.

Tracklet Consistency. As described, jersey numbers and roles are predicted independently for each detection, potentially leading to inconsistencies within tracklets. We adopt a **majority voting** approach within each tracklet to select the most common role and jersey number, ensuring uniformity.

Team Affiliation. Team affiliation is performed in three steps for tracklets having the "player" role assigned. First, the PRTReID embeddings of all detections within each tracklet are averaged to create a single tracklet-level representation of the player. Next, these tracklet-level embeddings are separated by a **K-means clustering** algorithm into two clusters representing two teams. Finally, the average 2D positions of each team on the pitch are compared to determine which team is positioned more to the left or right.

7. Experiments

7.1. Implementation details

To provide a baseline that is generic, we employ mostly pretrained networks that were not finetuned on SoccerNet. The only exceptions are PRTReid [57] and TVCalib [80]. We use the standard weights provided by TVCalib's authors in our baseline. Finally, PRTReid [57] is trained on the SoccerNet-GSR train set using parameters from the original paper. For more implementation details, we invite readers to visit our project's GitHub repository and Tracklab[3].

7.2. Evaluation

To evaluate the performance of our proposed method on the Game State Reconstruction task, we employ the GS-HOTA metric introduced in Sec. 5. In the supplementary materials, we evaluate the performance of our GSR-Baseline in the image plane on the standard Multi-Object Tracking (MOT) task. Unless specified otherwise, all experiments are performed on the SoccerNet-GSR test set.

7.3. Results and Analysis

Main Results and GS-HOTA Analysis. We report the performances of our GSR-Baseline in Tab. 1, which achieves 22.26% in GS-HOTA on the test set. All experiments in this table correspond to slight variations of the $Sim_{\text{GS-HOTA}}(P, G)$ introduced in Eq. (3). First, when "Pitch" is disabled, the *LocSim* function in Eq. (4) is replaced with the bounding boxes IoU in the image space: pitch localization and camera calibration have therefore no impact. Second, we ablate each attribute of the identification component in Eq. (5) (*IdSim* is set to 1 when all attributes are disabled). The first experiment in Tab. 1 falls back to the standard HOTA, *i.e.* with the IOU in image space as a similarity function. The remaining experiments illustrate how enabling attributes in Eq. (5) induces successive drops in performance, since it introduces additional predictions in the evaluation and therefore potential errors. Tab. 1 also highlights the key challenges of this task, showing that our GSR-Baseline struggles mostly with jersey number recognition, followed by pitch localization, team affiliation, and finally role classification. Finally, the influence of the GS-HOTA distance tolerance parameter τ introduced in Sec. 5 is illustrated in Fig. 4. According to this plot, picking $\tau = 5$ meters is a reasonable choice since performance quickly drops with a stricter tolerance.

Ablation Study of GSR-Baseline Modules. Table 2 illustrates the impact of each module on the overall performance. This study employs the ground truth as an oracle for all modules except the module of interest and its downstream modules in the pipeline. For instance, when examining the ReID module, the tracking, role classification, and team clustering modules are also activated. Dependencies between modules are depicted as a flowchart in Fig. 3. The first experiment (Exp. 1) shows that the heuristic chosen for team 'left'/'right' affiliation is highly effective, especially considering the significant impact that swapping two teams can have on GS-HOTA. Similarly, Exp. 2 demonstrates the solid performance of all modules depending on the ReID embeddings (*i.e.* tracking, role cls, and team aff.). Furthermore, Exp. 3 and 4 show the severe performance impact of enabling calibration and pitch localization, suggesting ample opportunities for improvements with these two modules. Similarly, Exp. 5 with the jersey number recognition module exposes it as another key weakness of the pipeline. Finally, performance in Exp. 6 is close to the complete baseline, since the object detector is the starting point for most of the pipeline, and ground truth data is therefore employed here only for pitch localization and camera calibration.

Our ablation study shows that while localization and identification are challenging alone, their intricate combination in GSR proves even more challenging.

Inference Time. Since the GSR-Baseline is an offline pipeline, each module processes its input in batches, where a single batch can span multiple images. The batch size and average frame rate of each module are reported in Tab. 2.

Table 1. **Main Results and GS-HOTA Analysis.** Attributes (Role, Team, Jersey) are ignored in the GS-HOTA computation when disabled. IoU in image space is used when Pitch is disabled.

Split	GS-HOTA components				GS-HOTA ↑
	Pitch	Role	Team	Jersey	
Test	✗	✗	✗	✗	57.64
	✓	✗	✗	✗	42.65
	✓	✓	✗	✗	40.76
	✓	✗	✓	✗	37.03
	✓	✗	✗	✓	25.65
	✗	✓	✓	✓	29.50
	✓	✓	✓	✓	**22.26**
Valid	✓	✓	✓	✓	**18.05**
Challenge	✓	✓	✓	✓	**23.36**

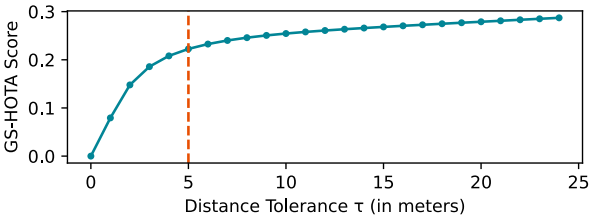
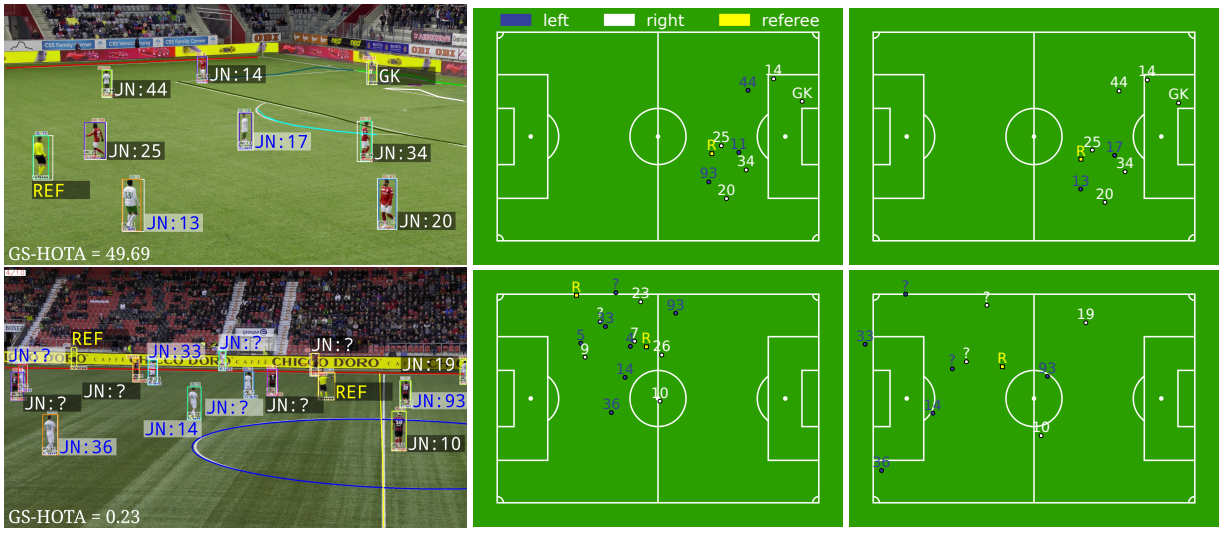


Figure 4. **Distance Tolerance Parameter** τ: its influence on the GS-HOTA score. We pick τ=5, illustrated by the orange line.

[3] https://github.com/TrackingLaboratory/tracklab



(a) Predictions on input image (b) Ground Truth minimap (c) Predicted minimap

Figure 5. **Qualitative results.** Output predictions of two frames from videos with different GS-HOTA values. (Top) High GS-HOTA (49.69%), with robust pitch localization and accurate athlete identification. (Bottom) Calibration failure (*e.g.* due to insufficient pitch elements) leads to completely erroneous athlete localization and poor GS-HOTA (0.23%).

Table 2. **GSR-Baseline Ablation Study.** We report the GS-HOTA for each GSR-Baseline module and its corresponding downstream modules by replacing other modules by a ground truth oracle. We also report their speed in FPS and their input batch sizes.

	Module	GS-HOTA ↑	Batch S.	FPS
(1)	Team Side	92.00	Video	1.5K
(2)	ReID (PRTReID)	87.42	16	14.5
(3)	Calibration (TVCalib)	51.39	512	7.6
(4)	Pitch (TVCalib)	49.99	16	2.9
(5)	Jersey N° (MMOCR)	56.75	32	3.8
(6)	BBox Det. (YOLOv8)	35.28	32	16.5
	Full Baseline	22.26	N/A	1.1

All inference speed tests are performed with an NVIDIA A100 32GB GPU. As illustrated, pitch localization, camera calibration, and jersey-number recognition emerge as the most time-consuming modules. It takes on average 11 minutes to process one 30s sequence from our dataset.

Qualitative Results. Fig. 5 illustrates two game state minimaps predicted by our GSR-Baseline and their respective ground truths. Our GSR-Baseline achieves a high GS-HOTA score of 49.69% on the video illustrated in the first row, accurately predicting most athletes' pitch positions and attributes. The bottom example, from a video with a GS-HOTA of 0.23%, exemplifies common failure cases, where even minor calibration inaccuracies can cause major pitch registration errors. In this frame, poor calibration is caused by the small number of visible salient points on the pitch.

8. Conclusion

Our work introduces the first Game State Reconstruction (GSR) benchmark for athlete identification and tracking on a minimap, comprising a new dataset, evaluation metric, and open-source baseline. Unlike previous efforts in sports video understanding that focused on specific subtasks, our approach stands out by benchmarking a complete pipeline, whose high-level game semantics outputs are directly relevant to a broad spectrum of downstream applications. Moreover, experiments with our proposed baseline reveal the inherent complexity of the GSR task and the significant interdependencies among its various subtasks. We hope that our introduced benchmark will pave the way for a new line of exciting research on specialized GSR methods. We anticipate future efforts to focus on (1) enhancing specific modules to increase performance, (2) implementing real-time pipelines, or even (3) developing end-to-end differentiable methods for tackling the task in one step.

Acknowledgments. This work was supported by SportRadar, the Service Public de Wallonie (SPW) Recherche, under the ReconnAIssance project and Grant N°8573, the F.R.S-FNRS, FRIA/FNRS, the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research through the Visual Computing Center (VCC) funding and the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI). Computational resources have been provided by the supercomputing facilities of the Université catholique de Louvain (CISM/UCLouvain).

References

[1] Adrià Arbués Sangüesa, Adrián Martín, Javier Fernández, Coloma Ballester, and Gloria Haro. Using player's body-orientation to model pass feasibility in soccer. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3875–3884, Seattle, WA, USA, 2020. Inst. Electr. Electron. Eng. (IEEE). 2

[2] Bavesh Balaji, Jerrin Bright, Harish Prakash, Yuhao Chen, David A. Clausi, and John Zelek. Jersey number recognition using keyframe identification from low-resolution broadcast videos. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, page 123–130, Ottawa, Ontario, Can., 2023. ACM. 2

[3] Ryan Beal, Georgios Chalkiadakis, Timothy J. Norman, and Sarvapali D. Ramchurn. Optimising game tactics for football. *arXiv*, abs/2003.10294, 2020. 2

[4] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 941–951, Seoul, North Korea, 2019. Inst. Electr. Electron. Eng. (IEEE). 2

[5] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP J. Image Video Process.*, 2008:1–10, 2008. 2, 4, 1

[6] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 3464–3468, Phoenix, AZ, USA, 2016. Inst. Electr. Electron. Eng. (IEEE). 2

[7] Erik Bochinski, Tobias Senst, and Thomas Sikora. Extending IOU based multi-object tracking by visual information. In *IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, pages 1–6, Auckland, New Zealand, 2018. Inst. Electr. Electron. Eng. (IEEE). 2

[8] Matthias Boeker and Cise Midoglu. Soccer athlete data visualization and analysis with an interactive dashboard. In *Int. Conf. Multimedia Retr.*, pages 565–576. Springer Int. Publ., 2023. 2

[9] Mengqi Cao, Min Yang, Guozhen Zhang, Xiaotian Li, Yilu Wu, Gangshan Wu, and Limin Wang. SpotFormer: A transformer-based framework for precise soccer action spotting. In *Int. Work. Multimedia Signal Process. (MMSP)*, pages 1–6, Shanghai, China, 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[10] Cheuk-Yiu Chan, Chun-Chuen Hui, Wan-Chi Siu, Sin-wai Chan, and H. Anthony Chan. To start automatic commentary of soccer game with mixed spatial and temporal attention. In *IEEE Region 10 Conference (TENCON)*, pages 1–6, Hong Kong, China, 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[11] Fan Chen and Christophe De Vleeschouwer. Personalized production of basketball videos from multi-sensored data under limited display resolution. *Comput. Vis. Image Underst.*, 114(6):667–680, 2010. 1

[12] Fan Chen and Christophe De Vleeschouwer. Automatic summarization of broadcasted soccer videos with adaptive fast-forwarding. In *IEEE Int. Conf. Multimedia Expo (ICME)*, pages 1–6, Barcelona, Spain, 2011. Inst. Electr. Electron. Eng. (IEEE). 1

[13] Jianhui Chen, Fangrui Zhu, and James J. Little. A two-point method for PTZ camera calibration in sports. In *IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, pages 287–295, Lake Tahoe, NV, USA, 2018. Inst. Electr. Electron. Eng. (IEEE). 4

[14] Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. A bottom-up approach based on semantics for the interpretation of the main camera stream in soccer games. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 1846–1855, Salt Lake City, UT, USA, 2018. 2

[15] Anthony Cioppa, Adrien Deliege, Maxime Istasse, Christophe De Vleeschouwer, and Marc Van Droogenbroeck. ARTHuS: Adaptive real-time human segmentation in sports through online distillation. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 2505–2514, Long Beach, CA, USA, 2019. Inst. Electr. Electron. Eng. (IEEE). 2

[16] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas B. Moeslund. A context-aware loss function for action spotting in soccer videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 13123–13133, Seattle, WA, USA, 2020. Inst. Electr. Electron. Eng. (IEEE). 2

[17] Anthony Cioppa, Adrien Deliège, Noor Ul Huda, Rikke Gade, Marc Van Droogenbroeck, and Thomas B. Moeslund. Multimodal and multiview distillation for real-time player detection on a football field. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 3846–3855, Seattle, WA, USA, 2020. 2

[18] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Floriane Magera, Olivier Barnich, Bernard Ghanem, and Marc Van Droogenbroeck. Camera calibration and player localization in SoccerNet-v2 and investigation of their representations for action spotting. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 4532–4541, Nashville, TN, USA, 2021. 2

[19] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Scaling up SoccerNet with multi-view spatial localization and re-identification. *Sci. Data*, 9(1):1–9, 2022. 2, 3

[20] 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 *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 3490–3501, New Orleans, LA, USA, 2022. Inst. Electr. Electron. Eng. (IEEE). 2, 3, 1

[21] Anthony Cioppa, Silvio Giancola, Vladimir Somers, Floriane Magera, Xin Zhou, Hassan Mkhallati, Adrien Deliège, Jan Held, Carlos Hinojosa, Amir M. Mansourian, Pierre Miralles, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdullah Kamal, Adrien Maglo, Albert Clapés, Amr Abdelaziz, Artur Xarles, Astrid Orcesi, Atom Scott, Bin Liu, Byoungkwon Lim, Chen Chen, Fabian Deuser, Feng Yan, Fufu Yu, Gal Shitrit, Guanshuo Wang, Gyusik Choi, Hankyul Kim, Hao Guo, Hasby Fahrudin, Hidenari

Koguchi, Håkan Ardö, Ibrahim Salah, Ido Yerushalmy, Iftikar Muhammad, Ikuma Uchida, Ishay Be'ery, Jaonary Rabarisoa, Jeongae Lee, Jiajun Fu, Jianqin Yin, Jinghang Xu, Jongho Nang, Julien Denize, Junjie Li, Junpei Zhang, Juntae Kim, Kamil Synowiec, Kenji Kobayashi, Kexin Zhang, Konrad Habel, Kota Nakajima, Licheng Jiao, Lin Ma, Lizhi Wang, Luping Wang, Menglong Li, Mengying Zhou, Mohamed Nasr, Mohamed Abdelwahed, Mykola Liashuha, Nikolay Falaleev, Norbert Oswald, Qiong Jia, Quoc-Cuong Pham, Ran Song, Romain Hérault, Rui Peng, Ruilong Chen, Ruixuan Liu, Ruslan Baikulov, Ryuto Fukushima, Sergio Escalera, Seungcheon Lee, Shimin Chen, Shouhong Ding, Taiga Someya, Thomas B. Moeslund, Tianjiao Li, Wei Shen, Wei Zhang, Wei Li, Wei Dai, Weixin Luo, Wending Zhao, Wenjie Zhang, Xinquan Yang, Yanbiao Ma, Yeeun Joo, Yingsen Zeng, Yiyang Gan, Yongqiang Zhu, Yujie Zhong, Zheng Ruan, Zhiheng Li, Zhijian Huangi, and Ziyu Meng. SoccerNet 2023 challenges results. *arXiv*, abs/2309.06006, 2023. 2

[22] Tom Decroos, Jan Van Haaren, and Jesse Davis. Automatic discovery of tactics in spatio-temporal soccer match data. In *ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, page 223–232. ACM, 2018. 2

[23] Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 4508–4519, Nashville, TN, USA, 2021. 2

[24] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. StrongSORT: Make DeepSORT great again. *IEEE Trans. Multimedia*, 25:8725–8737, 2023. 5

[25] EVS Broadcast Equipment. Multi-camera review system - Xeebra. https://evs.com/products/video-assistance/xeebra, 2022. 4

[26] D. Farin, S. Krabbe, and W. Effelsberg et.al. Robust camera calibration for sport videos using court models. In *Storage and Retrieval Methods and Applications for Multimedia*, pages 80–92, San Jose, California, USA, 2003. 2

[27] Ivan Alen Fernandez, Fan Chen, Fabien Lavigne, Xavier Desurmont, and Christophe De Vleeschouwer. Browsing sport content through an interactive H.264 streaming session. In *International Conferences on Advances in Multimedia*, pages 155–161, Athens, Greece, 2010. Inst. Electr. Electron. Eng. (IEEE). 1

[28] Maximilian T. Fischer, Daniel A. Keim, and Manuel Stein. Video-based analysis of soccer matches. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, page 1–9, Nice, France, 2019. ACM. 2

[29] Sebastian Gerke, Karsten Muller, and Ralf Schafer. Soccer jersey number recognition using convolutional neural networks. In *IEEE Int. Conf. Comput. Vis. Work. (ICCV Work.)*, pages 734–741, Santiago, Chile, 2015. Inst. Electr. Electron. Eng. (IEEE). 2

[30] Seyed Abolfazl Ghasemzadeh, Gabriel Van Zandycke, Maxime Istasse, Niels Sayez, Amirafshar Moshtaghpour, and Christophe De Vleeschouwer. DeepSportLab: a unified framework for ball detection, player instance segmentation and pose estimation in team sports scenes. *arXiv*, abs/2112.00627, 2021. 2

[31] Adhiraj Ghosh, Kuruparan Shanmugalingam, and Wen-Yan Lin. Relation preserving triplet mining for stabilising the triplet loss in re-identification systems. In *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pages 4829–4838, Waikoloa, HI, USA, 2023. Inst. Electr. Electron. Eng. (IEEE). 2

[32] Silvio Giancola and Bernard Ghanem. Temporally-aware feature pooling for action spotting in soccer broadcasts. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4490–4499, Nashville, TN, USA, 2021. 2

[33] Silvio Giancola, Anthony Cioppa, Adrien Deliège, Floriane Magera, Vladimir Somers, Le Kang, Xin Zhou, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdulrahman Darwish, Adrien Maglo, Albert Clapés, Andreas Luyts, Andrei Boiarov, Artur Xarles, Astrid Orcesi, Avijit Shah, Baoyu Fan, Bharath Comandur, Chen Chen, Chen Zhang, Chen Zhao, Chengzhi Lin, Cheuk-Yiu Chan, Chun Chuen Hui, Dengjie Li, Fan Yang, Fan Liang, Fang Da, Feng Yan, Fufu Yu, Guanshuo Wang, H. Anthony Chan, He Zhu, Hongwei Kan, Jiaming Chu, Jianming Hu, Jianyang Gu, Jin Chen, João V. B. Soares, Jonas Theiner, Jorge De Corte, José Henrique Brito, Jun Zhang, Junjie Li, Junwei Liang, Leqi Shen, Lin Ma, Lingchi Chen, Miguel Santos Marques, Mike Azatov, Nikita Kasatkin, Ning Wang, Qiong Jia, Quoc Cuong Pham, Ralph Ewerth, Ran Song, Rengang Li, Rikke Gade, Ruben Debien, Runze Zhang, Sangrok Lee, Sergio Escalera, Shan Jiang, Shigeyuki Odashima, Shimin Chen, Shoichi Masui, Shouhong Ding, Sin-wai Chan, Siyu Chen, Tallal El-Shabrawy, Tao He, Thomas B. Moeslund, Wan-Chi Siu, Wei Zhang, Wei Li, Xiangwei Wang, Xiao Tan, Xiaochuan Li, Xiaolin Wei, Xiaoqing Ye, Xing Liu, Xinying Wang, Yandong Guo, Yaqian Zhao, Yi Yu, Yingying Li, Yue He, Yujie Zhong, Zhenhua Guo, and Zhiheng Li. SoccerNet 2022 challenges results. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 75–86, Lisbon, Port., 2022. ACM. 2

[34] Silvio Giancola, Anthony Cioppa, Julia Georgieva, Johsan Billingham, Andreas Serner, Kerry Peek, Bernard Ghanem, and Marc Van Droogenbroeck. Towards active learning for action spotting in association football videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 5098–5108, Vancouver, Can., 2023. Inst. Electr. Electron. Eng. (IEEE). 2

[35] Jan Held, Anthony Cioppa, Silvio Giancola, Abdullah Hamdi, Bernard Ghanem, and Marc Van Droogenbroeck. VARS: Video assistant referee system for automated soccer decision making from multiple views. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 5086–5097, Vancouver, Can., 2023. Inst. Electr. Electron. Eng. (IEEE). 2

[36] James Hong, Haotian Zhang, Michaël Gharbi, Matthew

Fisher, and Kayvon Fatahalian. Spotting temporally precise, fine-grained events in video. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 33–51, Tel Aviv, Israël, 2022. Springer Nat. Switz. 2

[37] Hsiang-Wei Huang, Cheng-Yen Yang, Jiacheng Sun, Pyong-Kun Kim, Kwang-Ju Kim, Kyoungoh Lee, Chung-I Huang, and Jenq-Neng Hwang. Iterative scale-up ExpansionIoU and deep features association for multi-object tracking in sports. In *IEEE/CVF Winter Conf. Appl. Comput. Vis. Work. (WACVW)*, pages 163–172, Waikoloa, HI, USA, 2024. 2

[38] IFAB. Laws of the game. Technical report, The International Football Association Board, Zurich, Switzerland, 2022. 3

[39] Maxime Istasse, Julien Moreau, and Christophe De Vleeschouwer. Associative embedding for team discrimination. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 2477–2486, Long Beach, CA, USA, 2019. 2

[40] Maxime Istasse, Vladimir Somers, Pratheeban Elancheliyan, Jaydeep De, and Davide Zambrano. DeepSportradar-v2: A multi-sport computer vision dataset for sport understandings. In *Int. ACM Work. Multimedia Content Anal. Sports (MM-Sports)*, pages 23–29, Ottawa, Ontario, Can., 2023. ACM. 2

[41] Yudong Jiang, Kaixu Cui, Leilei Chen, Canjin Wang, and Changliang Xu. SoccerDB: A large-scale database for comprehensive video understanding. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, page 1–8, Seattle, WA, USA, 2020. ACM. 2

[42] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics, 2023. 5

[43] Victor Joos, Vladimir Somers, and Baptiste Standaert. TrackLab. https://github.com/TrackingLaboratory/tracklab, 2024. 5

[44] Stephen Karungaru, Hiroki Tanioka, and Kenji Matsuura. Soccer players real location determination using perspective transformation. In *Int. Conf. Soft Comput. Intell. Syst., Int. Symp. Adv. Intell. Syst. (SCIS&ISIS)*, pages 1–4, Ise, Japan, 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[45] Benjamin Kiefer, Matej Kristan, Janez Perš, Lojze Žust, Fabio Poiesi, Fabio Andrade, Alexandre Bernardino, Matthew Dawkins, Jenni Raitoharju, Yitong Quan, Adem Atmaca, Timon Höfer, Qiming Zhang, Yufei Xu, Jing Zhang, Dacheng Tao, Lars Sommer, Raphael Spraul, Hangyue Zhao, Hongpu Zhang, Yanyun Zhao, Jan Lukas Augustin, Eui-ik Jeon, Impyeong Lee, Luca Zedda, Andrea Loddo, Cecilia Di Ruberto, Sagar Verma, Siddharth Gupta, Shishir Muralidhara, Niharika Hegde, Daitao Xing, Nikolaos Evangeliou, Anthony Tzes, Vojtěch Bartl, Jakub Špaňhel, Adam Herout, Neelanjan Bhowmik, Toby P. Breckon, Shivanand Kundargi, Tejas Anvekar, Ramesh Ashok Tabib, Uma Mudenagudi, Arpita Vats, Yang Song, Delong Liu, Yonglin Li, Shuman Li, Chenhao Tan, Long Lan, Vladimir Somers, Christophe De Vleeschouwer, Alexandre Alahi, Hsiang-Wei Huang, Cheng-Yen Yang, Jenq-Neng Hwang, Pyong-Kun Kim, Kwangju Kim, Kyoungoh Lee, Shuai Jiang, Haiwen Li, Zheng Ziqiang, Tuan-Anh Vu, Hai Nguyen-Truong, Sai-Kit Yeung, Zhuang Jia, Sophia Yang, Chih-Chung Hsu, Xiu-Yu Hou, Yu-An Jhang, Simon Yang, and Mau-Tsuen Yang. 1st workshop on maritime computer vision (macvi) 2023: Challenge results. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 265–302, 2023. 2

[46] Minjung Kim, MyeongAh Cho, and Sangyoun Lee. Feature disentanglement learning with switching and aggregation for video-based person re-identification. In *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pages 1603–1612, Waikoloa, HI, USA, 2023. Inst. Electr. Electron. Eng. (IEEE). 2

[47] Zhanghui Kuang, Hongbin Sun, Zhizhong Li, Xiaoyu Yue, Tsui Hin Lin, Jianyong Chen, Huaqiang Wei, Yiqin Zhu, Tong Gao, Wenwei Zhang, Kai Chen, Wayne Zhang, and Dahua Lin. MMOCR: A comprehensive toolbox for text detection, recognition and understanding. In *Proceedings of the 29th ACM International Conference on Multimedia*, page 3791–3794. ACM, 2021. 6

[48] Karol Kurach, Anton Raichuk, Piotr Stanczyk, Michał Zajac, Olivier Bachem, Lasse Espeholt, Carlos Riquelme, Damien Vincent, Marcin Michalski, Olivier Bousquet, and Sylvain Gelly. Google research football: A novel reinforcement learning environment. In *AAAI Conf. Artif. Intell.*, pages 4501–4510. Association for the Advancement of Artificial Intelligence (AAAI), 2020. 2

[49] Hui Li, Peng Wang, Chunhua Shen, and Guyu Zhang. Show, attend and read: A simple and strong baseline for irregular text recognition. In *AAAI Conf. Artif. Intell.*, pages 8610–8617. Association for the Advancement of Artificial Intelligence (AAAI), 2019. 6

[50] Minghui Liao, Zhaoyi Wan, Cong Yao, Kai Chen, and Xiang Bai. Real-time scene text detection with differentiable binarization. In *AAAI Conf. Artif. Intell.*, pages 11474–11481. Association for the Advancement of Artificial Intelligence (AAAI), 2020. 6

[51] Hengyue Liu and Bir Bhanu. Pose-guided R-CNN for jersey number recognition in sports. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 2457–2466, Long Beach, CA, USA, 2019. 2

[52] Hongshan Liu, Colin Adreon, Noah Wagnon, Abdul Latif Bamba, Xueshen Li, Huapu Liu, Steven MacCall, and Yu Gan. Automated player identification and indexing using two-stage deep learning network. *Sci. Reports*, 13(1), 2023. 2

[53] Katja Ludwig, Julian Lorenz, Robin Schön, and Rainer Lienhart. All keypoints you need: Detecting arbitrary keypoints on the body of triple, high, and long jump athletes. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 5179–5187, Vancouver, Can., 2023. Inst. Electr. Electron. Eng. (IEEE). 2

[54] 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. *Int. J. Comput. Vis.*, 129(2):548–578, 2020. 2, 4, 1

[55] Floriane Magera. SoccerNet camera calibration challenge. https://github.com/SoccerNet/sn-calibration, 2022. 4

[56] Adrien Maglo, Astrid Orcesi, Julien Denize, and Quoc Cuong Pham. Individual locating of soccer players

from a single moving view. *Sensors*, 23(18):1–28, 2023. 3, 1

[57] Amir M. Mansourian, Vladimir Somers, Christophe De Vleeschouwer, and Shohreh Kasaei. Multi-task learning for joint re-identification, team affiliation, and role classification for sports visual tracking. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, page 103–112, Ottawa, Ontario, Can., 2023. ACM. 2, 5, 6, 7

[58] Cise Midoglu, Steven Hicks, Vajira Thambawita, Tomas Kupka, and Pål Halvorsen. MMSys'22 grand challenge on AI-based video production for soccer. In *ACM Multimedia Systems Conference (MMSys)*, pages 1–6, Athlone, Ireland, 2022. 2

[59] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. SoccerNet-caption: Dense video captioning for soccer broadcasts commentaries. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 5074–5085, Vancouver, Can., 2023. Inst. Electr. Electron. Eng. (IEEE). 2

[60] Thomas B. Moeslund, Graham Thomas, and Adrian Hilton. *Computer vision in sports*. Springer, 2014. 2

[61] Ahmed Nady and Elsayed Hemayed. Player identification in different sports. In *Comput. Vis. Imaging Comput. Graph. Theory Appl. (VISIGRAPP)*, pages 1–8, Vienna, Austria, 2021. SCITEPRESS - Science and Technology Publications. 2

[62] Banoth Thulasya Naik, Mohammad Farukh Hashmi, Neeraj Dhanraj Bokde, and Zaher Mundher Yaseen. A comprehensive review of computer vision in sports: Open issues, future trends and research directions. *Appl. Sci.*, 12(9):1–49, 2022. 2

[63] Luca Pappalardo, Paolo Cintia, Paolo Ferragina, Emanuele Massucco, Dino Pedreschi, and Fosca Giannotti. PlayeRank: Data-driven performance evaluation and player ranking in soccer via a machine learning approach. *ACM Trans. Intell. Syst. Technol.*, 10(5):1–27, 2019. 2

[64] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. A public data set of spatio-temporal match events in soccer competitions. *Sci. Data*, 6(1):1–15, 2019. 2

[65] Pascaline Parisot and Christophe De Vleeschouwer. Scene-specific classifier for effective and efficient team sport players detection from a single calibrated camera. *Comput. Vis. Image Underst.*, 159:74–88, 2017. 2

[66] Charles Perin, Romain Vuillemot, and Jean-Daniel Fekete. SoccerStories: A kick-off for visual soccer analysis. *IEEE Trans. Vis. Comput. Graph.*, 19(12):2506–2515, 2013. 2

[67] Reza Pourreza, Morteza Khademi, Hamidreza Pourreza, and Habib Rajabi Mashhadi. Robust camera calibration of soccer video using genetic algorithm. In *IEEE Int. Conf. Intell. Comput. Commun. Process. (ICCP)*, pages 123–127, Cluj-Napoca, Romania, 2008. Inst. Electr. Electron. Eng. (IEEE). 2

[68] S. Kanaga Suba Raja, K. Kausalya, B. Sandhiya, K. Abdul Waseem Nihaal W., A. Abiya Feba Mary, and J. Afra Thahseen. Tracking of multi athlete and action recognition in soccer sports video using deep learning techniques. *AIP Conference Proceedings*, 2802(1), 2024. 2

[69] Upendra M. Rao and Umesh C. Pati. A novel algorithm for detection of soccer ball and player. In *Int. Conf. Commun. Signal Process. (ICCSP)*, pages 344–348, Melmaruvathur, India, 2015. 2

[70] D. Sacha, F. Al-Masoudi, M. Stein, T. Schreck, D. A. Keim, G. Andrienko, and H. Janetzko. Dynamic visual abstraction of soccer movement. *Computer Graphics Forum*, 36(3):305–315, 2017. 2

[71] Miguel Santos Marques, Ricardo Gomes Faria, and José Henrique Brito. Hierarchical line extremity segmentation U-Net for the SoccerNet 2022 calibration challenge - pitch localization. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 442–453. Springer Nat. Switz., 2023. 2

[72] Atom Scott, Ikuma Uchida, Masaki Onishi, Yoshinari Kameda, Kazuhiro Fukui, and Keisuke Fujii. SoccerTrack: A dataset and tracking algorithm for soccer with fish-eye and drone videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3568–3578, New Orleans, LA, USA, 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[73] Karolina Seweryn, Gabriel Cheć, Szymon Łukasik, and Anna Wróblewska. Improving object detection quality in football through super-resolution techniques. *arXiv*, abs/2402.00163, 2024. 2

[74] Long Sha, Jennifier Hobbs, Panna Felsen, Winyu Wei, Patrick Lucey, and Sujoy Ganguly. End-to-end camera calibration for broadcast videos. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 13627–13636, Seattle, WA, USA, 2020. Inst. Electr. Electron. Eng. (IEEE). 2

[75] Gal Shitrit, Ishay Be'ery, and Ido Yerhushalmy. SoccerNet 2023 tracking challenge – 3rd place MOT4MOT team technical report. *arXiv*, abs/2308.16651, 2023. 2

[76] João V. B. Soares, Avijit Shah, and Topojoy Biswas. Temporally precise action spotting in soccer videos using dense detection anchors. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 2796–2800, Bordeaux, France, 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[77] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body part-based representation learning for occluded person Re-Identification. In *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pages 1613–1623, Waikoloa, HI, USA, 2023. Inst. Electr. Electron. Eng. (IEEE). 6

[78] Manuel Stein, Halldor Janetzko, Andreas Lamprecht, Thorsten Breitkreutz, Philipp Zimmermann, Bastian Goldlucke, Tobias Schreck, Gennady Andrienko, Michael Grossniklaus, and Daniel A. Keim. Bring it to the pitch: Combining video and movement data to enhance team sport analysis. *IEEE Trans. Vis. Comput. Graph.*, 24(1):13–22, 2018. 2

[79] ShiJie Sun, Naveed Akhtar, HuanSheng Song, Ajmal S. Mian, and Mubarak Shah. Deep affinity network for multiple object tracking. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43 (1):104–119, 2019. 2

[80] Jonas Theiner and Ralph Ewerth. TVCalib: Camera calibration for sports field registration in soccer. In *IEEE/CVF*

Winter Conf. Appl. Comput. Vis. (WACV), pages 1166–1175, Waikoloa, HI, USA, 2023. Inst. Electr. Electron. Eng. (IEEE). 2, 6, 7

[81] Jonas Theiner, Wolfgang Gritz, Eric Müller-Budack, Robert Rein, Daniel Memmert, and Ralph Ewerth. Extraction of positional player data from broadcast soccer videos. *arXiv*, abs/2110.11107, 2021. 2, 3

[82] Graham Thomas, Rikke Gade, Thomas B. Moeslund, Peter Carr, and Adrian Hilton. Computer vision for sports: current applications and research topics. *Comput. Vis. Image Underst.*, 159:3–18, 2017. 2

[83] Gabriel Van Zandycke, Vladimir Somers, Maxime Istasse, Carlo Del Don, and Davide Zambrano. DeepSportradar-v1: Computer vision dataset for sports understanding with high quality annotations. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 1–8, Lisbon, Port., 2022. ACM. 2

[84] Renaud Vandeghen, Anthony Cioppa, and Marc Van Droogenbroeck. Semi-supervised training to improve player and ball detection in soccer. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 3480–3489, New Orleans, LA, USA, 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[85] Kanav Vats, Mehrnaz Fani, David A. Clausi, and John Zelek. Multi-task learning for jersey number recognition in ice hockey. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, page 11–15. ACM, 2021. 2

[86] Balaji Veeramani, John W. Raymond, and Pritam Chanda. DeepSort: deep convolutional networks for sorting haploid maize seeds. *BMC Bioinformatics*, 19(S9), 2018. 2

[87] Luping Wang, Hao Guo, and Bin Liu. A boosted model ensembling approach to ball action spotting in videos: The runner-up solution to CVPR'23 SoccerNet challenge. *arXiv*, abs/2306.05772, 2023. 2

[88] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 3645–3649, Beijing, China, 2017. Inst. Electr. Electron. Eng. (IEEE). 1

[89] Fan Yang, Shigeyuki Odashima, Shoichi Masui, and Shan Jiang. The second-place solution for CVPR 2022 SoccerNet tracking challenge. *arXiv*, abs/2211.13481, 2022. 2

[90] Qixiang Ye, Qingming Huang, Shuqiang Jiang, Yang Liu, and Wen Gao. Jersey number detection in sports video for athlete identification. In *Visual Communications and Image Processing*, Beijing, China, 2005. SPIE. 2

[91] Junqing Yu, Aiping Lei, Zikai Song, Tingting Wang, Hengyou Cai, and Na Feng. Comprehensive dataset of broadcast soccer videos. In *IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, pages 418–423, Miami, FL, USA, 2018. Inst. Electr. Electron. Eng. (IEEE). 2

[92] Guiwei Zhang, Yongfei Zhang, Tianyu Zhang, Bo Li, and Shiliang Pu. PHA: Patch-wise high-frequency augmentation for transformer-based person re-identification. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 14133–14142, Vancouver, Can., 2023. Inst. Electr. Electron. Eng. (IEEE). 2

[93] Pengyi Zhang, Huanzhang Dou, Yunlong Yu, and Xi Li. Adaptive cross-domain learning for generalizable person re-identification. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 215–232. Springer Nat. Switz., 2022. 2

[94] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the fairness of detection and re-identification in multiple object tracking. *Int. J. Comput. Vis.*, 129(11):3069–3087, 2021. 2, 1

[95] 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 *Eur. Conf. Comput. Vis. (ECCV)*, pages 1–21. Springer Nat. Switz., 2022. 2, 1

[96] Xin Zhou, Le Kang, Zhiyu Cheng, Bo He, and Jingyu Xin. Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection. *arXiv*, abs/2106.14447, 2021. 2

[97] He Zhu, Junwei Liang, Chengzhi Lin, Jun Zhang, and Jianming Hu. A transformer-based system for action spotting in soccer videos. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 103–109, Lisbon, Port., 2022. ACM. 2

[98] Chen Zhu-Tian, Qisen Yang, Xiao Xie, Johanna Beyer, Haijun Xia, Yingcai Wu, and Hanspeter Pfister. Sporthesia: Augmenting sports videos using natural language. *IEEE Trans. Vis. Comput. Graph.*, 29(1):918–928, 2023. 2