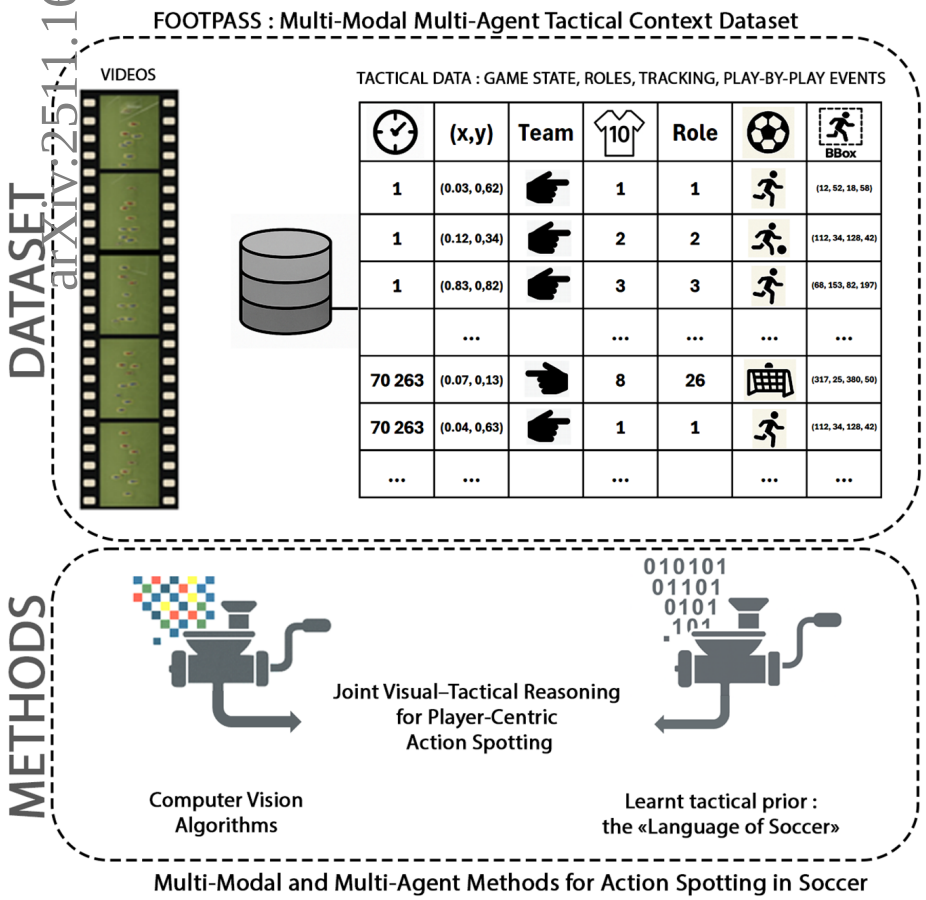
Graphical Abstract

FOOTPASS: A Multi-Modal Multi-Agent Tactical Context Dataset for Play-by-Play Action Spotting in Soccer Broadcast Videos

Jeremie Ochin, Raphael Chekroun, Bogdan Stanciulescu, Sotiris Manitsaris



FOOTPASS: A Multi-Modal Multi-Agent Tactical Context Dataset for Play-by-Play Action Spotting in Soccer Broadcast Videos

Jeremie Ochin[a,b,*], Raphael Chekroun[b], Bogdan Stanciulescu[a], Sotiris Manitsaris[a]

[a] *Center for Robotics, Mines Paris - PSL, 60 Bd Saint-Michel, 75006 Paris, France*
[b] *Footovision, 17 Rue Saint-Augustin, 75002 Paris, France*

Abstract

Soccer video understanding has motivated the creation of datasets for tasks such as temporal action localization, spatiotemporal action detection (STAD), or multi-object tracking (MOT). The annotation of structured sequences of events (who does what, when, and where) used for soccer analytics requires a holistic approach that integrates both STAD and MOT. However, current action recognition methods remain insufficient for constructing reliable play-by-play data and are typically used to assist rather than fully automate annotation. Parallel research has advanced tactical modeling, trajectory forecasting, and performance analysis, all grounded in game-state and play-by-play data. This motivates leveraging tactical knowledge as a prior to support computer-vision-based predictions, enabling more automated and reliable extraction of play-by-play data.

We introduce *Footovision Play-by-Play Action Spotting in Soccer Dataset* (FOOTPASS), the first benchmark for play-by-play action spotting over entire soccer matches in a multi-modal, multi-agent tactical context. It enables the development of methods for player-centric action spotting that exploit both outputs from computer-vision tasks (e.g., tracking, identification) and prior knowledge of soccer, including its tactical regularities over long time horizons, to generate reliable play-by-play data streams. These streams form an essential input for data-driven sports analytics.

Keywords: Dataset, Multimodal, Sports video understanding, Action spotting, Spatiotemporal action detection, Soccer Analytics, Tactical prior

*Corresponding author.
Email addresses: jeremie.ochin@minesparis.psl.eu (Jeremie Ochin), raphael.chekroun@footovision.com (Raphael Chekroun), bogdan.stanciulescu@minesparis.psl.eu (Bogdan Stanciulescu), sotiris.manitsaris@minesparis.psl.eu (Sotiris Manitsaris)

1. Introduction

Soccer stands as the most popular sport in the world, followed by millions of fans and played in over 200 countries. Its cultural significance and global audience have also made it one of the most data-rich sports, as each competition generates hours of video and every match produces millions of positional data points. In recent years, technological progress has transformed how this wealth of information is analyzed, giving rise to what is broadly referred to as soccer analytics, a field that integrates computer vision, machine learning, and domain knowledge to extract meaningful insights from the game (Cuevas et al., 2020).

While data-driven methods have long supported broadcasters and fans through visual enhancements and post-match statistics, the most profound shift has occurred within professional clubs themselves. Elite teams increasingly use advanced data analytics to evaluate player performance, optimize training sessions, and model tactical systems. Clubs now rely on video-based analytics not only to study their opponents' strategies but also to refine their own collective play, analyzing recurring tactical patterns, spatial occupation, and phase transitions across matches. As highlighted by Wang et al. (2024), the advent of AI-assisted tactical modeling and forecasting tools, such as TacticAI, has accelerated this transformation by enabling the interpretation of spatiotemporal data at a level of granularity previously achievable only by expert analysts.

This evolution reflects a broader trend: soccer is no longer observed solely through the lens of visual intuition but increasingly through structured, data-centric representations that quantify who does what, when, and where on the pitch. These structured descriptions, commonly known as play-by-play or event data, constitute the foundation of tactical analysis and performance evaluation. Yet, constructing them at scale still requires extensive manual annotation. This motivates research at the intersection of computer vision and soccer analytics, with the goal of automatically generating reliable, player-centric play-by-play data streams from broadcast videos

1.1. Definitions

In the sports analytics literature, the term *play-by-play data* refers to a structured chronological record of atomic match events, each specifying who performed an action, what the action was, when it occurred, and where on the field it took place (Davis et al., 2024). In soccer, two main types of events can be distinguished. *On-ball events* are instances where a player

makes contact with the ball: ball drives, passes, crosses, headers, throw-ins, shots, tackles, or blocks. They occur frequently, represent the vast majority of actions in a match, and their sequence captures the flow of play and reveals team tactics. In contrast, *sparse events* are rarer and typically correspond to decisive highlights (e.g., goals), infractions (e.g., fouls, offsides), or referee interventions (e.g., yellow/red cards, VAR checks). Set pieces such as corners or free kicks are usually encoded as passes or shots within on-ball sequences, but are annotated explicitly in sparse event datasets. This work focuses on on-ball play-by-play data.

Play-by-play data thus describe the match as a chain of discrete, player-centric events, enabling the reconstruction of full match dynamics. Such representations form the standard input for tactical analysis, outcome modeling, and performance evaluation. Play-by-play data are used in *soccer* (Aalbers and Van Haaren, 2019; Korte et al., 2019; Simpson et al., 2022; Mendes-Neves et al., 2024; Anzer et al., 2025; Yeung et al., 2025); *basketball* (Vračar et al., 2016; Damoulaki et al., 2025; Sun et al., 2025); *handball* (Mortelier et al., 2024); and *American football* (Ötting, 2021). This widespread use underscores their centrality in sports analytics.

Alongside play-by-play records, *game-state data* refers to the spatiotemporal localization of players on the pitch together with their identities. This spatiotemporal representation of the game's dynamics is a standard input for tasks such as tactical analysis and multi-agent trajectory modeling. It is used, for example, in *soccer* (Bialkowski et al., 2014; Martens et al., 2021; Forcher et al., 2022; Raabe et al., 2022; Capellera et al., 2024; Ogawa et al., 2025); *basketball* (Sha et al., 2017; Felsen et al., 2018); *handball* (Mortelier et al., 2024); and *hockey* (Lucey et al., 2013). In soccer, play-by-play data can be regarded as a subset of game-state data, filtered to the players currently performing a ball-related action and augmented with an action label.

Given their central role in soccer analytics, both play-by-play and game-state data are essential representations to be extracted through automated video understanding.

1.2. Bridging Perception and Tactical Reasoning for Action Spotting through Multi-Modal, Multi-Agent Representations

Thanks to dedicated perception datasets, the computer vision subtasks required for soccer game-state reconstruction, such as multi-object tracking, player (re-)identification, jersey number recognition, field localization, and camera calibration, are improving steadily, making game-state recovery

increasingly automated. This trend is reflected in the increasing metrics reported for these tasks in the last four editions of the SoccerNet Challenges (Giancola et al., 2022; Cioppa et al., 2024b,a; Giancola et al., 2025). In contrast, the precise and fine-grained annotation of play-by-play data remains largely a manual, time-consuming process carried out by trained expert operators, who must scrub through the footage to label actions accurately (Cartas et al., 2022; Bassek et al., 2025).

Current state-of-the-art Spatiotemporal Action Detection (STAD) methods provide a natural starting point for reconstructing play-by-play data from raw soccer broadcast video, as they jointly perform action detection and actor localization. However, although promising, these methods still fall short of the performance required for exhaustive event coverage in soccer analytics, since achieving high recall would produce numerous false positives that must be filtered out by human annotators (Singh et al., 2023; Wang et al., 2023; Ochin et al., 2025b).

Several factors account for these limitations. In particular, the visual conditions of broadcast footage often make reliable detection difficult, as variations in weather, illumination, and shadows combine with occlusions, motion blur, rapid camera movements, frequent visual ambiguities, and replays, to obscure certain actions. Second, existing STAD models often lack contextual understanding, treating detections almost in isolation: they are usually trained on short clips without explicitly reasoning about the tactical intentions of players. This leaves them blind to the structured dependencies that govern soccer. The problem is especially acute in high-recall settings, where many false positives could be avoided by incorporating long-range temporal and game-state contexts (Ochin et al., 2025a).

Although these approaches may continue to improve, Ochin et al. (2025a) demonstrated that operating at a more abstract modeling level, centered on players as entities in an evolving system, makes it possible to leverage the *language of soccer*: the tactical and temporal regularities and dynamics that structure the game over time spans much longer than the short clips typically used in conventional video understanding. In this way, even imperfect sequences of action predictions can be refined by exploiting such *tactical priors*.

Therefore, a path toward reliable player-centric action spotting lies in approaches that bridge perception with tactical reasoning. Yet, no existing public dataset provides both: broadcast video annotated at the play-by-play level together with the tactical state information necessary to integrate higher-level reasoning.

To support research at the intersection of computer vision and tactical

modeling, we introduce *Footovision Play-by-Play Action Spotting in Soccer Dataset* (FOOTPASS). The dataset provides human-annotated play-by-play data together with multi-modal, multi-agent tactical data aligned with full-length broadcast videos of soccer matches. This includes extended game-state information (player positions and velocities on the pitch, team memberships, jersey numbers, and roles), as well as single player tracking data in screen space.

FOOTPASS is designed to serve both communities: researchers working on fundamental computer vision subtasks and those exploring methods to fuse vision-based data with tactical priors to enhance sequences of spotted events. It establishes a shared foundation for advancing methods that either strengthen individual components of the pipeline or integrate perception with tactical modeling, with evaluation centered on the core benchmark: reliable player-centric play-by-play action spotting.

1.3. Contributions

This paper makes two main contributions. First, it introduces FOOTPASS, the first dataset for play-by-play action spotting in full-length soccer broadcast videos, designed within a multi-modal and multi-agent tactical context. Second, it benchmarks several existing methods to evaluate their ability to generate precise play-by-play records.

FOOTPASS is thus released as a resource for the community, offering a basis for future research on both improved perception modules and the fusion of vision-based data with tactical priors to deliver reliable play-by-play streams for scalable and data-driven soccer analytics.

2. Related Work

2.1. Soccer Video Action Recognition Datasets

This section review the available public datasets for action recognition in Soccer. Their main characteristics are compared in Table 1.

Early large-scale soccer datasets primarily focused on the detection of *sparse events* such as goals, fouls, cards, and offsides. These annotations captured key highlights of the match but did not provide the dense sequence of on-ball actions that reflects tactical organization and team play.

Yu et al. (2018) introduced the Comprehensive Soccer dataset, later extended by Feng et al. (2020) to encompass 350 broadcast videos totaling 282 hours. Its annotations include shot boundary detection (types of field of

Dataset	Task	Event type	#Classes	Modality	#Events
Comprehensive Soccer	✓	Sparse	11	▲	6,850
SoccerNet v1	•	Sparse	4	▲	6,637
SoccerNet v2	•	Sparse	17	▲	110,458
SoccerDB	✓	Sparse	11	▲	37,715
SoccerReplay-1988	✓	Sparse	24	▲	∼150,000
Ball Action Spotting	•	On-ball	12	▲	12,433
STAD Multisports	■	On-ball	15	▲	12,254
Integrated Dataset	•	On-ball	45	⧫	11,137

Table 1: Comparison of public soccer datasets across key dimensions. **Task markers:** ✓ = Temporal Action Localization (TAL), ■ = Spatiotemporal Action Detection (STAD), • = Action Spotting (AS). **Modality markers:** ▲ = Video, ⧫ = Spatiotemporal data (position on the pitch)

view, replays, and transition types), 11 classes of sparse events with temporal boundaries (start and end frames), and limited tracking annotations covering 40 shots (approximately 13 minutes of play). However, no information is provided about the active players performing the actions.

Giancola et al. (2018) released the SoccerNet dataset and proposed the task of action spotting in soccer, where the goal is to locate the anchor time of an event rather than its temporal boundaries (the latter being the focus of Temporal Action Detection or Temporal Action Localization). The original dataset contained 500 complete matches, amounting to 764 hours of video. With SoccerNet v2, published by Deliège et al. (2021), annotations expanded from 4 to 17 sparse event classes, with the addition of shot segmentation annotations and a replay grounding task. Subsequent works (Cioppa et al., 2022a,b; Gutiérrez-Pérez and Agudo, 2025) further extended the dataset by incorporating cross-view correspondences between main shots and replays, field line and jersey number annotations, tracking data for all visible players across 200 sequences of 30 seconds, and three-dimensional ball localization annotations on replay sequences.

Similarly, Jiang et al. (2020) introduced the SoccerDB dataset, comprising 343 matches divided into clips ranging from 3 to 30 seconds. The annotations cover 11 classes of sparse events with temporal boundaries, alongside bounding boxes for players and the ball. Nevertheless, no tracking data are included, and event annotations do not identify the active player.

More recently, Rao et al. (2025) released the SoccerReplay-1988 dataset, a large-scale multi-modal resource comprising 1,988 broadcast matches from six European leagues (2014/15–2023/24 seasons). Its annotations include 24 classes of sparse events, automatically aligned with around 150,000 broadcast commentaries, as well as rich metadata about matches, teams, coaches, and referees. While the dataset is notable for its scale and the integration of natural language commentaries, it does not provide tracking data, player positions, or game-state annotations. Consequently, it is primarily suited for sparse event spotting and video-language research, rather than dense on-ball STAD, play-by-play action spotting or tactical modeling.

In 2023, a Ball Action Spotting dataset was incorporated into SoccerNet, with 5 complete matches initially annotated with 2 classes of ball-related events (drive and pass). This later evolved into 12 classes of on-ball events and was further extended into a Team Ball Action Spotting task in 2025 (Cioppa et al., 2024b,a; Giancola et al., 2025). Despite these advances, the existing on-ball event annotations do not include the localization or identity of the active players.

To our knowledge, the only public dataset that includes both action classes and active player localization in soccer is the STAD Multisports dataset Li et al. (2021). This dataset spans four sports, basketball, volleyball, soccer, and aerobic gymnastics, and covers 66 action categories with 800 clips per sport, each averaging 750 frames. Actions are annotated through a class label and an action tube, i.e., a sequence of bounding boxes tracking the active player during the temporal extent of the action. For soccer, this amounts to 12,254 action instances across 15 classes and 225,000 bounding boxes. However, it lacks broader tracking data beyond the active player and does not provide player identities.

Finally, the Integrated Dataset of Spatiotemporal and Event Data in Elite Soccer published by Bassek et al. (2025) addresses the scarcity of public resources that combine play-by-play and game-state data. This dataset includes frame-by-frame game-state information for seven Bundesliga matches, together with hierarchically structured discrete events categorized into player, team, and referee actions. It also provides tactical information such as team formations and jersey numbers. Importantly, the positional data represents a true ground truth, captured by an expensive multi-view camera system that tracks all players simultaneously across the full pitch. However, the corresponding broadcast video footage is not publicly available, which limits the possibility of directly synchronizing the annotations with other video sources. Moreover, the authors note that, with only seven matches, the dataset provides a limited sample size for deriving robust conclusions in match analysis,

as representative studies typically require much larger datasets.

2.2. Spatiotemporal Action Detection Methods

Spatiotemporal Action Detection (STAD) has received increasing attention in recent years, driven by its broad range of real-world applications in surveillance, sports analysis, and autonomous driving (Wang et al., 2023). This growing interest has led to the development of deep learning approaches that can be broadly grouped into two families: frame-level and clip-level methods (Li et al., 2021). Frame-level models assign bounding boxes and action labels independently at each frame and subsequently integrate predictions over time. In contrast, clip-level models, often referred to as *action tubelet detectors*, jointly capture temporal context and localize actions across sequences of frames. A more detailed overview of these families is provided in the survey by Wang et al. (2023).

Within clip-level approaches, the Track-Aware Action Detector (TAAD) proposed by Singh et al. (2023) generates per-actor, per-frame action predictions by first detecting and tracking actors, then aggregating features along their trajectories with a fine-tuned 3D CNN and ROI Align (He et al., 2017), followed by a Temporal Convolutional Network. This design enables TAAD to achieve state-of-the-art performance on public STAD benchmarks, while showing robustness to camera motion.

In contrast, Peral et al. (2025) introduced a soccer-specific approach focused on temporally accurate detection of passes and receptions. Their method also begins with tracking potential acting players, but instead of sampling features from the global feature map along the track, it constructs player-centric video tubes and extracts features directly from these cropped sequences to estimate per-frame ball possession.

These approaches are well suited to automated soccer analytics, where reconstructing the game-state typically precedes event annotation, enabling actions to be directly associated with player identities. However, their precision in high-recall settings remains low, primarily because they lack contextual understanding.

2.3. Multi-Modal and Multi-Agent Action Spotting in Soccer

Recent research has explored moving beyond pixel-based detections toward representations that integrate structured, multi-agent, and multi-modal information, enabling richer contextual reasoning over soccer dynamics. These methods treat the game as an evolving system of interdependent agents, in

contrast to conventional spatiotemporal action detection approaches that focus on per-player feature aggregation along motion paths, and therefore capture only limited structured inter-player context.

Ochin et al. (2025b) proposed a Graph Neural Network (GNN)-based extension of the Track-Aware Action Detector (TAAD) that explicitly incorporates game-state information, such as player positions, velocities, and team memberships, alongside visual features extracted by a 3D CNN. By encoding local inter-player relationships as a spatio-temporal graph, the model captures important contextual information and jointly learns visual and tactical representations, which improve its action-spotting performance. Experiments on a dedicated private dataset demonstrated that fusing visual and structured features leads to substantial precision gains, particularly in the high-recall regimes required for exhaustive event coverage. While this method has the advantage of being trainable end-to-end, it still lacks long-range temporal context, operating on short clips of 50 frames, and has not yet been tested on full-length broadcast matches and public datasets.

To address these temporal limitations, Ochin et al. (2025a) introduced the *Denoising Sequence Transduction* (DST) model, which extends STAD by integrating additional structured information specific to coordinated multi-agent domains, such as soccer, and by operating on long sequences of actions. The approach uses noisy, context-free, player-centric STAD predictions as a pixel-based prior and then denoises these sequences using role-based structured features, including player positions, velocities, and team context. In this setting, DST produces coherent and tactically plausible action sequences and achieves significant improvements in both precision and recall in high-recall regimes compared to purely visual baselines, while remaining computationally efficient on a single GPU.

Together, these works outline a multi-modal, multi-agent paradigm for action spotting in soccer, bridging the gap between computer vision and tactical modeling. They demonstrate how the integration of game-state reasoning with vision-based detections can lead to more reliable and interpretable play-by-play predictions, an approach directly aligned with the motivation of the FOOTPASS benchmark.

Implementations of both models, retrained on the FOOTPASS dataset, are used as baseline benchmarks, providing reproducible reference points and enabling future comparisons by the research community.

3. The FOOTPASS Dataset

At a glance, FOOTPASS provides broadcast video aligned with human-annotated play-by-play data of on-ball events, player identities (via jersey numbers, team memberships, and roles), single-player tracklets, and game-state variables (positions and velocities). The dataset spans 54 full matches and is designed to benchmark reliable player-centric action spotting in a multi-modal, multi-agent tactical context, where the core task is to predict spatiotemporal sequences of actions, *i.e.* identifying *who* performs *what*, *where*, and *when*, directly from broadcast video and auxiliary information.

A distinctive feature of FOOTPASS is the joint availability of broadcast video and tactical context. Ground-truth annotations are provided for play-by-play and game-state data, while tracking data consists of single-player tracklets. Importantly, the play-by-play ground truth always specifies the acting player's identity (via jersey number), even if no corresponding bounding box is available, since these annotations are provided manually. This setup reflects the natural difficulties of broadcast analysis, where occlusions, replays, and motion blur often obscure parts of the match.

FOOTPASS thus provides the means to bootstrap research at the intersection of perception and tactical modeling, with improvements in perception modules evaluated through their contribution to the core benchmark of reliable play-by-play action spotting.

The remainder of this section details the dataset construction (Section 3.1) and its global statistics (Section 3.2).

3.1. Dataset Construction

The dataset was curated with an emphasis on size, quality, and diversity, enabling researchers to test hypotheses and train deep neural networks under conditions that closely mirror practical applications such as automated play-by-play annotation.

Size of the dataset. We provide data from 54 full-length soccer matches. While this number is smaller than the number of matches in the datasets of Jiang et al. (2020) and Deliège et al. (2021), which comprise 347 and 500 matches respectively, it is comparable in terms of the total number of annotated events. Specifically, our dataset contains 102,992 events, compared to 37,715 and 110,458 for the two aforementioned datasets. In addition, our dataset addresses the scarcity of joint spatiotemporal and event data in soccer (Bassek et al., 2025), although the positions of players not visible in the broadcast footage are inferred, as explained in the next section.

Quality of the data. We provide broadcast videos recorded natively in Full HD (1920×1080), at 25 fps, and properly synchronized with the released play-by-play and game-state annotations.

Diversity of the data. The matches were selected from major European leagues and competitions of the season 2023/24 (in alphabetical order): the French Ligue 1, the German Bundesliga, the Italian Serie A, the Spanish La Liga, and the UEFA Champions League. In total, 50 different teams are represented.

3.1.1. Game-State Reconstruction

The game-state reconstruction is performed sequentially: field lines detection and camera calibration, players detection, localization and tracking, and imputation of missing values. The overall goal of this step of the dataset construction process is to determine, for each frame of the broadcast video, *where* is each *player* on the pitch.

Field line detection and camera calibration. The first step of the game-state reconstruction process is to detect specific field lines and landmarks in the video using well-established deep learning approaches to image segmentation (Minaee et al., 2022), trained on a private dataset of field lines and landmarks. Given the known dimensions of these field lines and landmarks, a field-to-image homography is then estimated for each frame of the broadcast video, excluding replays. Camera intrinsics (including focal length) and extrinsics are recovered by decomposing the homography under planar constraints (Hartley and Zisserman, 2004), and camera pose and lens distortion parameters are jointly optimized via nonlinear refinement on all visible landmarks and line-straightness constraints.

Player detection, localization, and tracking. Well-established deep learning–based object detection methods are used to detect players (Sun et al., 2024), trained on a private dataset. At this stage, an initial triage of bounding boxes is performed to retain only single-player detections, since the presence of multiple players in one box can bias its dimensions and negatively affect subsequent localization. In addition, a pitch mask obtained via segmentation is used to eliminate bounding boxes outside the playing field or its immediate surroundings. Similarly to Cartas et al. (2022), it is assumed that a player contacts the ground at the midpoint between the bottom coordinates of the bounding box, which is then projected from the image plane onto the pitch using the camera model. Tracking is carried out in multiple

passes using a custom matching algorithm that combines: team membership prediction, recognized jersey numbers and the roster of numbers present in the match, segmentation-based color profiles of different body regions, deep feature similarity, inter-frame bounding boxes intersection-over-union and distances, and statistical priors on player positions and roles. Overall, this production process of positional data is *FIFA EPTS certified*.

Imputation of missing values. In broadcast videos of soccer, on average 9 to 12 players are simultaneously visible on screen. Since this dataset is constructed from single-player tracking data, there are missing values in the positions of players extracted through the process described above. To facilitate the use of the dataset, missing values are imputed and the trajectories of invisible players are inferred using a custom algorithm that interpolates positions between visible timesteps, subject to constraints on speed and acceleration, as well as statistical priors on relative positioning based on player roles.

Player role annotation. Another important type of tactical data provided in this dataset concerns the *player roles*. These roles depend on the team formation adopted during the game and may vary dynamically as players adjust their positioning and responsibilities. Although teams can adopt numerous formations and role configurations depending on their strategy, we reduced the total number of possible roles to 13 in order to facilitate the use of learning algorithms under limited data conditions. Increasing the number of role categories would lead to a higher-dimensional and more sparsely populated feature space, thereby requiring substantially more annotated data for robust training.

Player roles were determined through a combination of trajectory analysis, clustering, and expert annotation. The defined roles are: *Goalkeeper, Left Back, Left Central Back, Mid Central Back, Right Central Back, Left Midfielder, Right Midfielder, Defensive Midfielder, Attacking Midfielder, Left Winger, Right Winger, Central Forward, and Right Back*.

3.1.2. Play-by-play Data Annotation

Action classes. The dataset comprises on-ball events, annotated by a professional team using custom-built annotation software designed to streamline the process and ensure both high-quality temporal alignment and accurate identification of the acting player. Events are categorized into eight classes: *drive, pass, cross, shot, header, throw-in, tackle,* and *block*. Samples of those classes can be viewed in Figure 1.

- *Drive:* Occurs when a player receives the ball and maintains possession until passing it to a teammate or losing it to an opponent. The temporal anchor corresponds to the frame in which the player first receives the ball. If the ball is played immediately (without being controlled), a pass is annotated instead, with its anchor defined at the frame in which the ball is struck.

- *Pass:* Defined when a player strikes the ball with the intent of transferring possession to a teammate. Its temporal anchor is the frame of ball contact.

- *Cross:* A pass made from outside the penalty area toward a teammate located inside the penalty area, regardless of whether the pass is completed. Its temporal anchor is the frame when the ball is struck.

- *Shot:* Annotated when the ball is directed toward the goal with the clear intent to score. The temporal anchor is the frame in which the ball is struck.

- *Header:* Occurs when a player intentionally strikes the ball with the head. The temporal anchor is the frame of head–ball contact.

- *Throw-in:* A manual throw performed by a player after the ball has completely crossed the touchline. Its temporal anchor is the frame when the ball leaves the player's hands.

- *Tackle:* Annotated when a player legally dispossesses an opponent of the ball. The temporal anchor is the frame in which ball contact occurs.

- *Block:* Occurs when a player intercepts an opponent's pass or shot. The temporal anchor is the frame when the ball is first touched by the intercepting player.

Figure 1: Samples of the dataset classes. From left to right : ball-drive, pass, cross, shot, header, throw-in, tackle and ball-block

Temporal grounding of events. Given the instantaneous nature of on-ball events—ball touches—the spotting framework introduced by Giancola et al. (2018) is adopted. Each event is temporally grounded by a single anchor frame, defined as the frame in which the ball touch occurs. Consequently, events are treated as temporally instantaneous and have no duration.

Event representation. Each event is represented as a tuple *(frame, class, team, jersey)*, where:

- *frame* is the index of the video frame in which the event occurs (indices start at 0),
- *class* is an integer corresponding to the event class (from 0 to 9),
- *team* is an integer indicating the team of the acting player (0 for the left team, 1 for the right team), and
- *jersey* is an integer corresponding to the jersey number of the acting player.

The play-by-play annotations are integrated with the game-state data by augmenting each player's state with class labels, including a background class for non-action frames.

3.2. Dataset Statistics

Class distribution. The dataset exhibits a strong class imbalance (Fig. 2). The vast majority of events are *passes* (49.9%) and *drives* (39.0%), together accounting for nearly 90% of all annotations. Less frequent actions include *headers* (4%), *crosses* (2.3%), *throw-ins* (1.9%), *blocks* (1.4%), *shots* (1.2%), and *tackles* (0.3%). This distribution mirrors the natural statistics of soccer, where passes and carries dominate ball interactions, while decisive actions such as shots and tackles are comparatively rare.

Bounding-box coverage. STAD approaches either predict bounding boxes for actors during inference or require bounding boxes as input to perform action detection on pre-identified players. To support the latter case, FOOTPASS provides single-player tracking bounding boxes. An event is considered covered if the acting player has an associated bounding box after tracklet interpolation to fill gaps shorter than 50 frames (Fig. 3). On average, 81.5% of events are covered. Certain classes, such as *blocks* (78.4%), *crosses* (70.7%), *headers* (66.7%), *tackles* (62.1%), and *throw-ins* (43.8%), show lower coverage due to their frequent occurrence in crowded areas (e.g., close to penalty

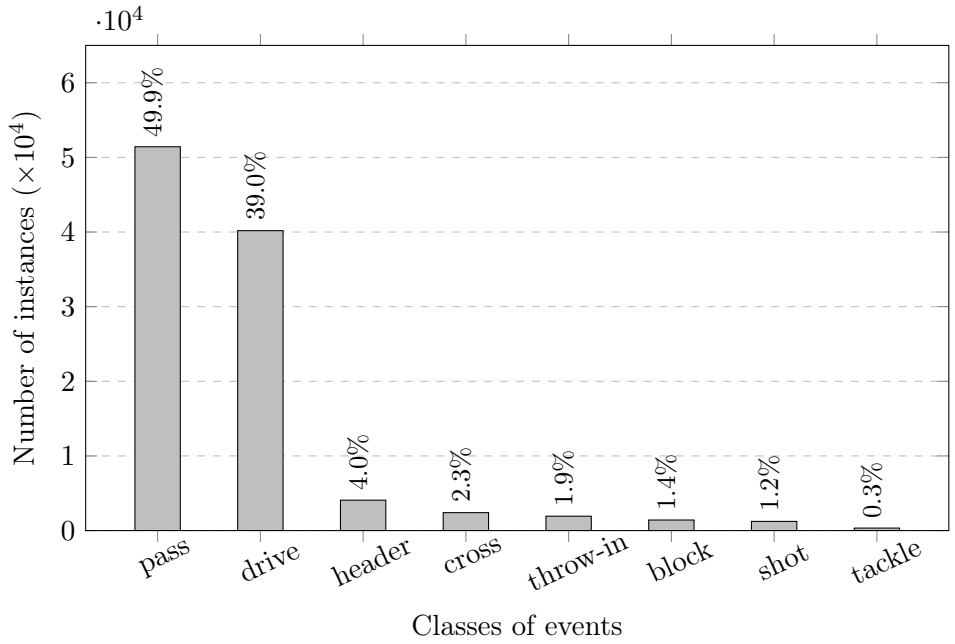


Figure 2: Distribution of annotated events across the 8 on-ball classes. Passes and drives dominate, while decisive actions such as shots and tackles are comparatively rare.

boxes), their involvement in duels or player contacts that lead to occlusion, or broadcast practices that cut to replays. By contrast *drives* (85.7%), *shots* (81.6%) and *passes* (81.5%) achieve higher coverage rates, reflecting their greater observability in broadcast footage.

Replay vs. live broadcast. Fig. 4 reports the proportion of events whose anchor frame occurs during live broadcast segments versus replay segments. As expected, most events occur during live play. However, 46.8% of *throw-ins* fall into replay segments because broadcast directors typically cut to replays of the preceding action after the ball goes out of play and return to live coverage only once the throw-in has already been executed. By contrast, actions such as *headers* (95.4% live), *blocks* (95% live), and especially *shots* (99.5% live) almost always occur during live play. These editing practices disrupts temporal continuity, complicating temporal localization.

Summary. In summary, the dataset provides a realistic distribution of soccer events characterized by strong class imbalance, varying levels of visual

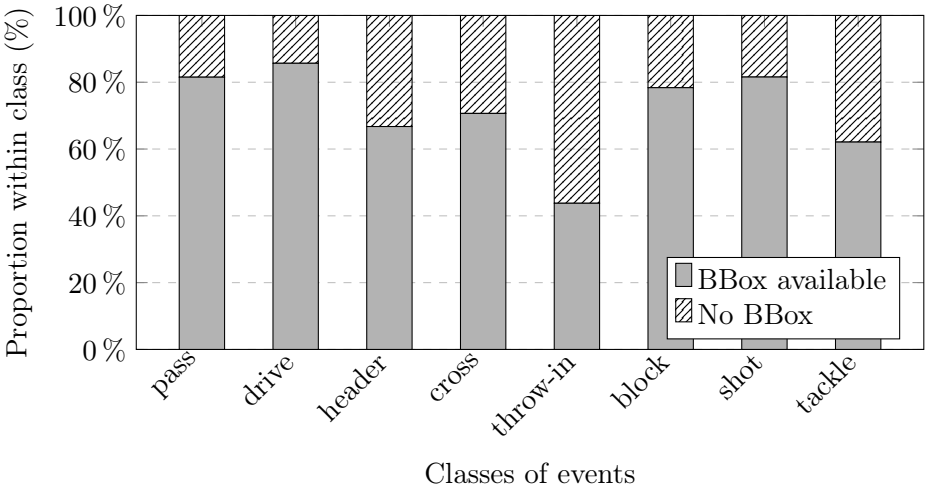


Figure 3: Proportion of events per class with bounding-box annotation after interpolation and extrapolation. Coverage is highest for drives, passes, and shots, while headers, tackles, and throw-ins are more affected by occlusion and broadcast editing practices.

observability, and broadcast-specific artifacts such as replays. These properties capture the inherent complexity of real-world soccer video and underscore the importance of approaches that combine low-level perception with contextual reasoning for reliable play-by-play reconstruction.

4. Experiments

4.1. The FOOTPASS benchmark

Benchmark setup. The 54 matches in FOOTPASS were manually divided into training, validation, and test sets. The validation and test sets were selected to ensure that even the least-represented classes contain at least 15 visible instances in each split. All videos were downsampled from Full HD (1920×1080) to a resolution of 352×640. The resulting split comprises 91,327 training instances from 48 matches, 6,070 validation instances from 3 matches (indices 18, 24, and 47), and 5,595 test instances from 3 matches (indices 0, 36, and 50).

Metrics. Following Ochin et al. (2025a), we report overall and per-class Precision and Recall at a low threshold $\tau = 0.15$, which is slightly above the score expected from a uniform distribution given the number of classes. This

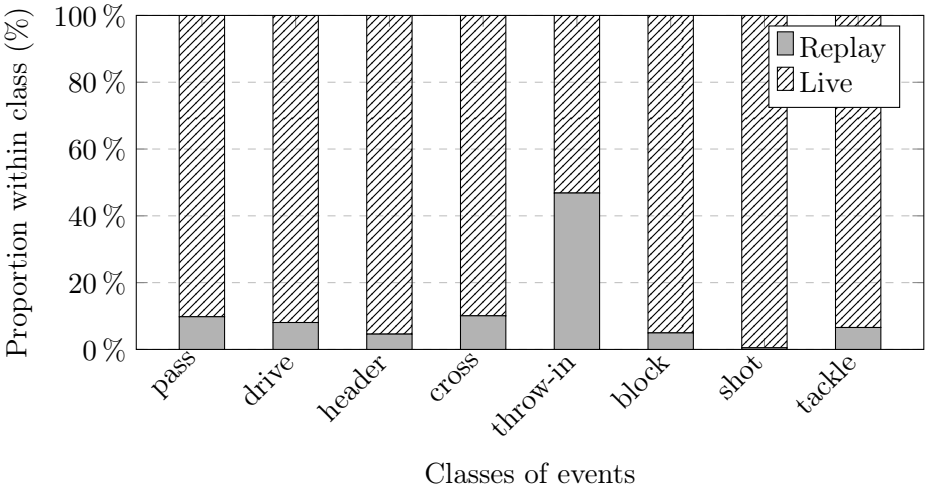


Figure 4: Distribution of events by broadcast mode (live vs. replay). Throw-ins frequently fall into replay segments because directors cut away to show the preceding sequence after the ball goes out of play, returning only after the throw-in has been executed. In contrast, headers, blocks, and shots occur almost exclusively during live play (>95% of the time).

configuration prioritizes high recall to minimize missed detections, while aiming to maximize precision within that regime. It is suitable in the context of assisted annotation, where it is faster to discard a false positive than to scrub through the video to find a missed false negative. Evaluation is performed by matching predicted and ground-truth actions of the same player and class within a fixed temporal tolerance of $\pm\delta$ frames around the annotation, with $\delta = 12$. A match is counted as a true positive (TP), unmatched predictions as false positives (FP), and unmatched ground-truth events as false negatives (FN). This procedure enables the computation of Precision and Recall both overall and per class. Additionally, we report per-class Average Precision (AP), computed using the 11-point interpolation method proposed in Everingham et al. (2010). All metrics are reported on the test set.

4.2. Benchmarked Methods

To evaluate the benefit of reasoning at the game level using tactical information, three methods are benchmarked on FOOTPASS: TAAD (Singh et al., 2023), a purely visual STAD approach that produces per-player predictions; TAAD+GNN, a graph-based extension of TAAD derived from Ochin et al. (2025b); and TAAD+DST, a post-processing method applied

to TAAD, proposed by Ochin et al. (2025a). The latter two methods exploit the multi-modal, multi-agent tactical context provided by FOOTPASS.

In addition, an ablation study is conducted on the DST model, extending that of the original paper, to demonstrate that the observed improvements stem from reasoning at the game level rather than from learning to process TAAD's raw predictions more efficiently than standard Non-Maximum Suppression (NMS) and thresholding.

Implementation Details – TAAD. A lightweight implementation of TAAD was trained on randomly selected short clips of 50 frames sampled at 25 frames per second (2 seconds of video), each containing at least one annotated action for which the acting player had a bounding box. To mitigate class imbalance during training, no more than 500 samples per class were selected in a given epoch, thereby reducing the ratio between the most and least represented classes. Moreover, since TAAD produces player-centric predictions and up to 22 players can be visible on the field during an action, 21 of whom are background, only 4 to 5 background tracklets were provided as negative samples to the network along with the tracklet containing the action.

Training used the AdamW optimizer (Loshchilov and Hutter, 2019) with a learning rate of 10^{-3} for the classification head and 5×10^{-5} for the pretrained X3D-L video backbone (Feichtenhofer, 2020), a weight decay of 10^{-4} on non-bias parameters, and a batch size of 6. The network was trained for 25 epochs on a single NVIDIA RTX A6000 GPU, with learning rates reduced by a factor of 10 at epochs 10 and 20. Data augmentation was applied to both videos and bounding boxes, including random scaling, rotation, translation, horizontal flipping, and color jittering, using the Albumentations library (Buslaev et al., 2020). Following the recommendations of Hong et al. (2022) for handling a dominant background class in action spotting, a label dilation of ± 1 frame was applied during training, and the cross-entropy loss weights of the foreground classes were boosted relative to the background.

For evaluation, the trained TAAD network was applied to the full matches of the test set, split into overlapping windows of 50 frames with a stride of 25 frames. Overlapping logits were averaged and stored for further processing. Per-role, per-frame predictions were obtained by applying a softmax and selecting the action with the highest score. As a post-processing step, temporal NMS with a 25-frame window was applied per player and per action class to handle successive actions occurring in quick succession. The resulting predictions were concatenated to produce play-by-play records and matched with the ground-truth annotations.

Implementation Details – TAAD+GNN. A variant of TAAD+GNN based on Ochin et al. (2025b) was implemented. The same pre-trained X3D-L video backbone was used, together with Edge Convolution layers employing an asymmetric edge function (Wang et al., 2019). To incorporate temporal modeling, Temporal Convolutional layers were inserted between Edge Convolution layers instead of creating explicit temporal edges between nodes representing the same player across adjacent time steps. Additionally, a Gated Recurrent Unit (Cho et al., 2014) layer was added before the final classification head.

The optimizer, training parameters, and data augmentation scheme were identical to those used for TAAD, with the addition that position and velocity data were also flipped when the corresponding video frame was flipped horizontally. A focal loss (Lin et al., 2020) was employed to mitigate the dominance of the background class among the 22 player instances. The same evaluation and post-processing procedures as for TAAD were applied.

Implementation Details – TAAD+DST. The DST model was trained using the averaged TAAD logits from the training set and the role-based structured game-state representation described in Ochin et al. (2025a). Training was performed on randomly sampled sequences of 750 frames, using the AdamW optimizer with a learning rate of 2.5×10^{-4}, a weight decay of 10^{-4} on non-bias parameters, and a batch size of 96. A single RTX A6000 GPU was used. The network was trained for 10 epochs, with the learning rate reduced by a factor of 10 at epochs 3, 6, and 8. For each epoch, 2000 sequences of 750 frames were randomly sampled from each of the 48 full-length matches in the training set. To mitigate overfitting, data augmentation was performed by randomly mirroring players along the X-axis of the pitch, the Y-axis, or both. Team memberships and roles were adjusted accordingly (e.g., a left winger from team "right" becomes a right winger from team "left" when mirroring along the X-axis). Since the averaged logits are ordered by role on the pitch, they were reshuffled to reflect the applied symmetries.

For evaluation, the trained DST network was applied to the full matches of the test set, split into contiguous windows of 750 frames. No post-processing was applied, except for thresholding the confidence scores of the predictions. The resulting predictions were concatenated to produce play-by-play predictions and matched with the ground-truth annotations.

Implementation Details – TAAD+DST Ablation. To rule out the possibility that the DST merely learns to post-process the continuous per-player signals produced by TAAD (the logits) into event sequences, the same

training procedure as above was applied to *unstructured* data. In this setting, the DST was trained on batches of individual player sequences of logits, together with corresponding tactical features (player position, velocity, and role, the latter provided as a one-hot encoded vector). These features were encoded and decoded by the DST, enabling it to learn how to transform each player's sequence of logits and tactical data into an event sequence.

While the original DST is trained using three separate cross-entropy losses (one for the frame, one for the player role, and one for the action), this ablated version employs only two: one for the frame and one for the action. The optimizer, training parameters, and data augmentation scheme were identical to those used for TAAD+DST, and the same evaluation procedure was applied.

4.3. Results and discussion

Overall performance. At a low confidence threshold suited for assisted annotation (τ=0.15), the purely visual TAAD baseline attains high recall but low precision, generating many spurious detections (Table 2). Incorporating spatiotemporal inter-player relationship modeling (TAAD+GNN) improves both recall (+3 points) and precision (+19 points), lifting F1 from 35.9% to 52.1%. Finally, the model that reasons over long temporal horizons and at the game level (TAAD+DST) achieves the best overall performance, with 68% precision and 67% recall, almost doubling the F1 score relative to TAAD (67.5% vs. 35.9%). These results indicate that modeling tactical structure and temporal context provides a promising and effective direction for reliable play-by-play extraction in soccer analytics.

Class	**TAAD** (Singh et al., 2023)			**TAAD+GNN** (Ochin et al., 2025b)			**TAAD+DST** (Ochin et al., 2025a)			**TAAD+DST (Abl.)** (Ochin et al., 2025a)		
	PR	**REC**	**F1**	**PR**	**REC**	**F1**	**PR**	**REC**	**F1**	**PR**	**REC**	**F1**
Pass	43.8	61.0	51.0	56.7	65.1	60.6	**72.6**	**72.1**	**72.3**	43.9	57.9	49.9
Drive	23.9	60.0	34.2	51.6	64.3	57.3	**68.2**	**68.9**	**68.5**	52.1	55.1	53.6
Header	10.6	**54.1**	17.7	15.7	45.9	23.4	**30.7**	26.3	**28.3**	24.8	38.5	30.2
Cross	16.6	**66.4**	26.6	32.0	59.1	41.5	**62.8**	51.8	**56.8**	47.7	38.7	42.7
Throw-in	23.8	44.4	31.0	42.9	39.4	41.1	**65.9**	**58.6**	**62.0**	40.0	30.3	34.5
Block	3.2	**47.1**	6.0	4.9	33.8	8.6	**23.3**	14.7	**18.0**	10.8	14.7	12.5
Shot	12.3	**74.6**	21.1	28.3	61.2	38.7	**63.4**	67.2	**65.2**	36.6	50.7	42.5
Tackle	1.5	16.7	2.8	**2.4**	**25.0**	**4.4**	0.0	0.0	0.0	0.0	0.0	0.0
Overall	25.6	59.9	35.9	44.5	62.7	52.1	**68.2**	**66.8**	**67.5**	45.0	54.0	49.1

Table 2: Comparison of the Precision (**PR**), Recall (**REC**), and F1-score (**F1**) across benchmarked methods at a confidence threshold of 15% and $\delta = 12$ frames. Metrics are $\times 10^2$. The best performance per class is highlighted in bold.

Class-specific behavior. Performance gains are not uniform across classes (Fig. 5, Table 2). *Drive* and *Pass* show the largest recall improvements, reflecting how both actions depend on player trajectories and inter-player interactions, patterns best captured by TAAD+GNN and TAAD+DST.

The *drive* class is particularly difficult because its annotation depends explicitly on temporal context. The action begins at the first ball control after reception, and without memory, every contact could be misinterpreted as a new drive. Models lacking sequence-level reasoning may over-segment these actions, while TAAD+DST's temporal modeling correctly anchors the initial touch, explaining its superior performance for this class.

Sparse and spatially constrained events such as *Cross*, *Throw-in* and *Block* show the largest relative precision improvements. These actions often occur near the sidelines or in crowded areas, where context about player positions and team organization may help reduce false detections. *Shot* also benefits strongly (F1 increasing from 21.1% to 65.2%), likely reflecting its higher contextual predictability and visibility compared with other action types.

Interestingly, *Header* shows the highest AP with the TAAD baseline, even though TAAD+DST achieves better F1 at low confidence thresholds. This indicates that while DST enhances recall, it slightly degrades the high-precision region of TAAD's signal for this class. A likely explanation lies in the nature of headers, which are often less controlled or tactically structured than passes or drives. The ball may deflect unpredictably or head in non-strategic directions, making long-range tactical reasoning less effective and sometimes mildly detrimental when denoising. A similar observation applies to *Block*, where the action outcome often reflects reaction rather than intention.

Finally, the *tackle* class remains challenging across methods given its scarcity and frequent occlusions. TAAD+DST did not register true positives for *tackle*, suggesting insufficiently discriminative cues in current features and the need to adjust sequence sampling during training to mitigate class imbalance.

Comparison with previous experiments. Relative to the results reported in prior work, our numbers are lower in several classes, a discrepancy largely explained by dataset scale. The original TAAD+GNN from Ochin et al. (2025b) was trained on a curated clips dataset with at least 2,500 samples per class and evaluated under a temporal action localization protocol, where events have a start and an end, rather than action spotting as in FOOTPASS. Reported results reached a mean Average Precision (mAP) of 49.7%

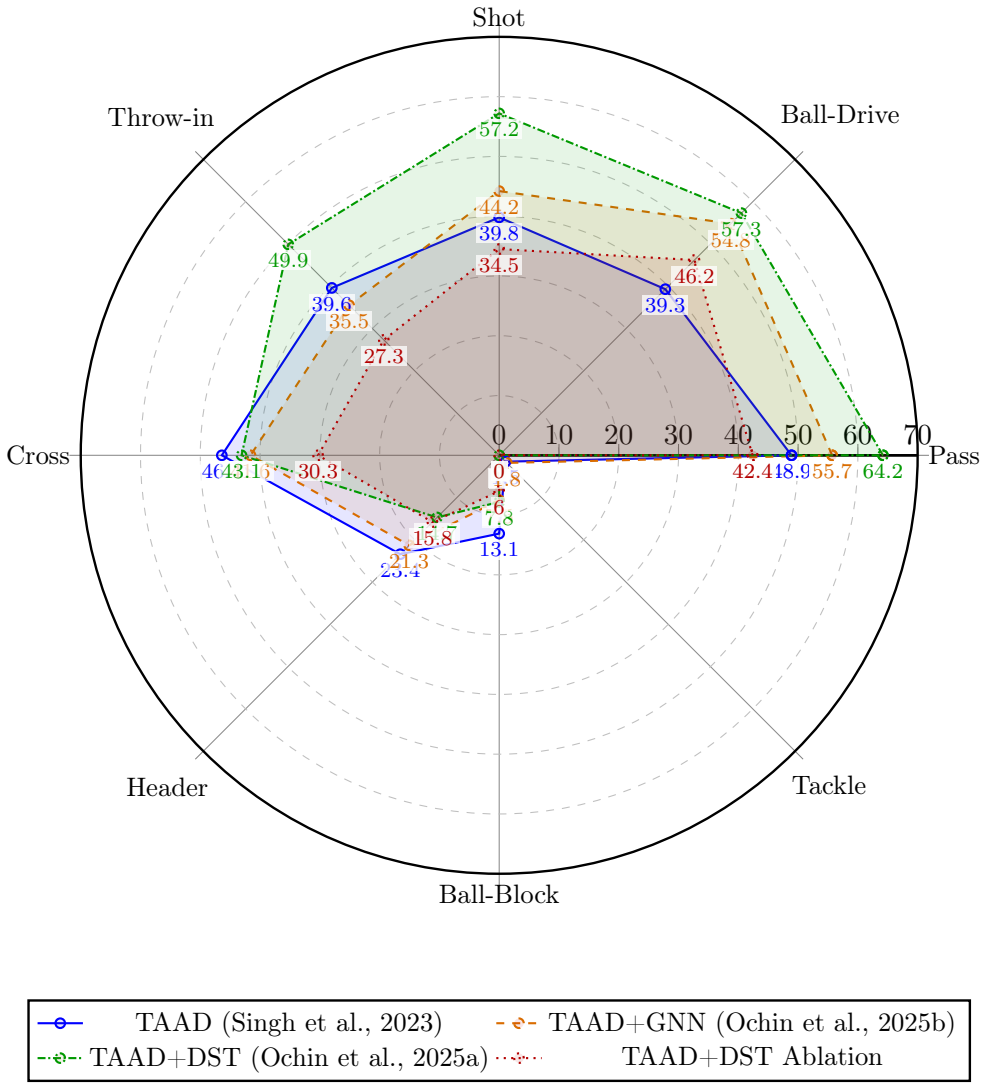


Figure 5: Comparison of Average Precision (AP, $\times 10^2$) per action class across benchmarked methods, with $\delta = 12$ frames.

at a temporal IoU threshold of 0.5, compared with 32.8% mAP with $\delta = 12$ frames for TAAD+GNN and 36.8% mAP for TAAD+DST, both trained on FOOTPASS. Although not directly comparable, these differences align with the stronger supervision available in the former dataset and its controlled clip-level context, which excluded full broadcast artifacts such as replays, occlusions, and long off-ball sequences. The sharper performance

drop observed here is therefore consistent with the increased difficulty of broadcast-level data and the stronger class imbalance in FOOTPASS.

The original TAAD from Ochin et al. (2025a) was trained on the same rich clips dataset as above, and the DST was trained on a much larger set of full matches (240 vs. 48 in FOOTPASS), providing both higher-quality prior detections and greater temporal diversity for sequence learning. Reported results on that private dataset reached 78.7% overall precision and 75.8% overall recall, higher than those observed on FOOTPASS, respectively 68.2% and 66.8%, confirming that the model performances scales with data quantity.

Despite using roughly four times less data than the private dataset of Ochin et al. (2025a), these experiments demonstrate that with appropriate data augmentation, such data-hungry methods can still be trained effectively and reproduce their relative improvements over visual baselines. This highlights the value of FOOTPASS as a public benchmark for studying how multi-modal, tactically grounded reasoning can be learned under realistic, resource-limited conditions.

Robustness to visibility and broadcast conditions. Roughly 80% of events occur with a bounding box available for the acting player and 20% without (with 9% due to replays). Focusing on recall, when a box is present, all methods perform reasonably well, with TAAD+GNN slightly ahead (see Figure 6). The crucial difference appears without boxes: TAAD and TAAD+GNN are almost entirely dependent on visual cues and recover very few events when bounding boxes are missing for the acting players (3.9% and 2.6% recall), while TAAD+DST maintains 33.4% recall. Overall, approximately 9 to 10% of TAAD+DST's true positives occur without bounding boxes (compared to about 1% for TAAD). This shows that TAAD+DST can infer actions even when the actor is not visible and highlights the value of tactical priors and long-range consistency. Even when the actor is visually missing, due to occlusion, off-screen positioning or replay cuts, game-state–aware denoising can still infer plausible and temporally consistent actions.

Ablation analysis and role of tactical context. The ablated TAAD+DST, trained without multi-agent reasoning, confirms that structured context is essential. It underperforms the full TAAD+DST across all aggregate metrics (F1: 49.1% vs. 67.5%) and in AP for most classes (Fig. 5). Yet, it exceeds the visual baselines in several cases and, notably, outperforms TAAD and TAAD+GNN in recall and F1 at low confidence thresholds when no bounding boxes are available, particularly for *Throw-in*, *Drive*, *Pass* and *Cross*.

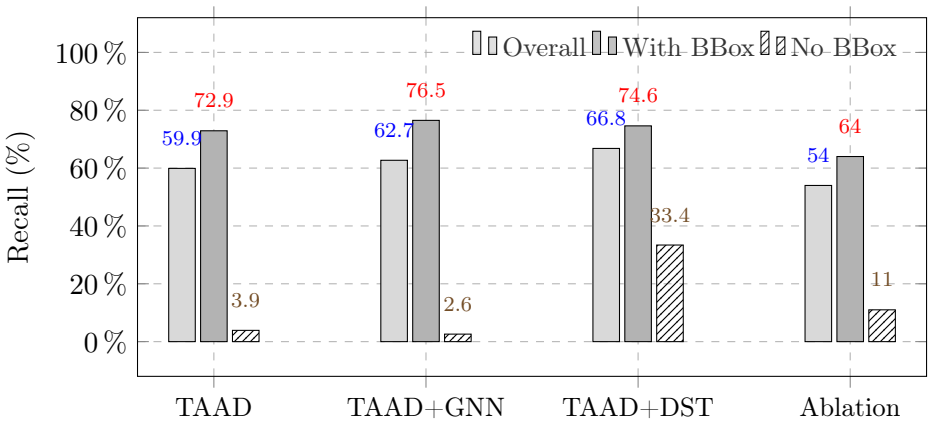


Figure 6: Overall recall and recall by visibility subset. Only TAAD+DST maintains substantial recall when the acting player has no bounding box (33.4%), highlighting robustness to missing visual localization.

This behavior suggests that the ablated version still leverages local spatiotemporal regularities, such as the recurrent spatial patterns of throw-ins near the sidelines and short-term temporal continuity, but lacks the global, role-aware reasoning that enables the full TAAD+DST to capture coordinated team behavior across longer sequences.

The comparison between the ablated and full TAAD+DST therefore shows that temporal modeling alone cannot explain the observed improvements. The full model's gain arises from integrating tactical priors and multi-agent relationships over longer time horizons. Two mechanisms likely underpin this effect. First, role and team conditioned priors encode who is likely to perform what and where; for example, wingers are more likely to cross, central roles to pass or shoot, and defenders to clear or block. Second, long-range temporal reasoning enforces ball possession continuity and smooths noisy clip-level detections, allowing the model to infer missing or off-screen actions and filter implausible events. Together, these mechanisms yield precision gains at fixed high recall and enable action recovery under occlusion or replay conditions.

5. Conclusion

This work presents FOOTPASS, a public benchmark for player-centric action spotting that couples full-match broadcast video with multi-modal, multi-agent tactical context. Using FOOTPASS, we provide reproducible

baselines confirming the previously reported benefits of tactical structure and long-range temporal reasoning on broadcast-level data, while quantifying their robustness under occlusion and replay conditions. The benchmark establishes a solid foundation for evaluating methods on realistic soccer footage that reflects the challenges of professional match analysis.

On the dataset side, future work includes extending annotations with acting-player and referee boxes, hierarchical action taxonomies, and sparse events such as infractions, referee interventions, and set pieces. On the methods side, we plan to explore end-to-end learning that unifies perception, tracking, and sequence reasoning; leverage audio and commentary to recover off-screen or occluded actions; and evaluate direct spotting systems that predict (frame, team, jersey, class) from video without explicit tactical inputs.

FOOTPASS thus opens new avenues for research at the intersection of computer vision and tactical modeling, providing a realistic and publicly accessible foundation to study how visual cues and tactical structure, such as spatiotemporal data and roles, can be jointly leveraged for player-centric action understanding.

Data availability. The FOOTPASS annotation files supporting this study will be released on Hugging Face and the baselines on Github[1], where the final license and README will specify terms of use for research and evaluation.

The associated broadcast videos are distributed via the SoccerNet dataset repository on Hugging Face and remain subject to SoccerNet's current non-disclosure agreement (NDA). Researchers wishing to access the footage must request it directly from SoccerNet and agree to the NDA; our experiments can be reproduced by combining those videos and annotations with the code provided in the Github release.

Any additional materials required to reproduce the results are available from the corresponding author upon reasonable request.

Declaration of generative AI and AI-assisted technologies in the writing process. During the preparation of this work, the authors used ChatGPT, developed by OpenAI, exclusively in order to improve grammar, spelling, and LaTeX formatting. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

[1] `https://github.com/JeremieOchin/FOOTPASS`

References

Aalbers, B., Van Haaren, J., 2019. Distinguishing between roles of football players in play-by-play match event data, in: Brefeld, U., Davis, J., Van Haaren, J., Zimmermann, A. (Eds.), Machine Learning and Data Mining for Sports Analytics, Springer International Publishing, Cham. pp. 31–41.

Anzer, G., Arnsmeyer, K., Bauer, P., Bekkers, J., Brefeld, U., Davis, J., Evans, N., Kempe, M., Robertson, S.J., Smith, J.W., Haaren, J.V., 2025. Common data format (cdf): A standardized format for match-data in football (soccer). URL: https://arxiv.org/abs/2505.15820, arXiv:2505.15820.

Bassek, M., Rein, R., Weber, H., Memmert, D., 2025. An integrated dataset of spatiotemporal and event data in elite soccer. Scientific Data 12, 195. URL: https://doi.org/10.1038/s41597-025-04505-y, doi:10.1038/s41597-025-04505-y.

Bialkowski, A., Lucey, P., Carr, P., Yue, Y., Sridharan, S., Matthews, I., 2014. Large-scale analysis of soccer matches using spatiotemporal tracking data, in: 2014 IEEE International Conference on Data Mining, pp. 725–730. doi:10.1109/ICDM.2014.133.

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A., 2020. Albumentations: Fast and flexible image augmentations. Information 11. URL: https://www.mdpi.com/2078-2489/11/2/125, doi:10.3390/info11020125.

Capellera, G., Ferraz, L., Rubio, A., Agudo, A., Moreno-Noguer, F., 2024. Footbots: A transformer-based architecture for motion prediction in soccer, in: IEEE International Conference on Image Processing, ICIP 2024, Abu Dhabi, United Arab Emirates, october 27-30, 2024, IEEE. pp. 2313–2319. URL: https://doi.org/10.1109/ICIP51287.2024.10647396, doi:10.1109/ICIP51287.2024.10647396.

Cartas, A., Ballester, C., Haro, G., 2022. A graph-based method for soccer action spotting using unsupervised player classification, in: Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports, Association for Computing Machinery, New York, NY, USA. p. 93–102. URL: https://doi.org/10.1145/3552437.3555691, doi:10.1145/3552437.3555691.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y., 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: Moschitti, A., Pang, B., Daelemans, W. (Eds.), Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Doha, Qatar. pp. 1724–1734. URL: https://aclanthology.org/D14-1179/, doi:10.3115/v1/D14-1179.

Cioppa, A., Deliège, A., Giancola, S., Ghanem, B., Van Droogenbroeck, M., 2022a. Scaling up soccernet with multi-view spatial localization and re-identification. Scientific Data 9, 355. URL: https://doi.org/10.1038/s41597-022-01469-1, doi:10.1038/s41597-022-01469-1.

Cioppa, A., Giancola, S., Deliege, A., Kang, L., Zhou, X., Cheng, Z., Ghanem, B., Van Droogenbroeck, M., 2022b. SoccerNet-Tracking: Multiple Object Tracking Dataset and Benchmark in Soccer Videos , in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Computer Society, Los Alamitos, CA, USA. pp. 3490–3501. URL: https://doi.ieeecomputersociety.org/10.1109/CVPRW56347.2022.00393, doi:10.1109/CVPRW56347.2022.00393.

Cioppa, A., Giancola, S., Somers, V., Joos, V., Magera, F., Held, J., Ghasemzadeh, S.A., Zhou, X., Seweryn, K., Kowalczyk, M., Mróz, Z., Łukasik, S., Hałoń, M., Mkhallati, H., Deliège, A., Hinojosa, C., Sanchez, K., Mansourian, A.M., Miralles, P., Barnich, O., Vleeschouwer, C.D., Alahi, A., Ghanem, B., Droogenbroeck, M.V., Gorski, A., Clapés, A., Boiarov, A., Afanasiev, A., Xarles, A., Scott, A., Lim, B., Yeung, C., Gonzalez, C., Rüfenacht, D., Pacilio, E., Deuser, F., Altawijri, F.S., Cachón, F., Kim, H., Wang, H., Choe, H., Kim, H.J., Kim, I.M., Kang, J.M., Tursunboev, J., Yang, J., Hong, J., Lee, J., Zhang, J., Lee, J., Zhang, K., Habel, K., Jiao, L., Li, L., Gutiérrez-Pérez, M., Ortega, M., Li, M., Lopatto, M., Kasatkin, N., Nemtsev, N., Oswald, N., Udin, O., Kononov, P., Geng, P., Alotaibi, S.G., Kim, S., Ulasen, S., Escalera, S., Zhang, S., Yang, S., Moon, S., Moeslund, T.B., Shandyba, V., Golovkin, V., Dai, W., Chung, W., Liu, X., Zhu, Y., Kim, Y., Li, Y., Yang, Y., Xiao, Y., Cheng, Z., Li, Z., 2024a. Soccernet 2024 challenges results. URL: https://arxiv.org/abs/2409.10587, arXiv:2409.10587.

Cioppa, A., Giancola, S., Somers, V., Magera, F., Zhou, X., Mkhallati, H., Deliège, A., Held, J., Hinojosa, C., Mansourian, A.M., Miralles, P.,

Barnich, O., De Vleeschouwer, C., Alahi, A., Ghanem, B., Van Droogenbroeck, M., Kamal, A., Maglo, A., Clapés, A., Abdelaziz, A., Xarles, A., Orcesi, A., Scott, A., Liu, B., Lim, B., Chen, C., Deuser, F., Yan, F., Yu, F., Shitrit, G., Wang, G., Choi, G., Kim, H., Guo, H., Fahrudin, H., Koguchi, H., Ardö, H., Salah, I., Yerushalmy, I., Muhammad, I., Uchida, I., Be'ery, I., Rabarisoa, J., Lee, J., Fu, J., Yin, J., Xu, J., Nang, J., Denize, J., Li, J., Zhang, J., Kim, J., Synowiec, K., Kobayashi, K., Zhang, K., Habel, K., Nakajima, K., Jiao, L., Ma, L., Wang, L., Wang, L., Li, M., Zhou, M., Nasr, M., Abdelwahed, M., Liashuha, M., Falaleev, N., Oswald, N., Jia, Q., Pham, Q.C., Song, R., Hérault, R., Peng, R., Chen, R., Liu, R., Baikulov, R., Fukushima, R., Escalera, S., Lee, S., Chen, S., Ding, S., Someya, T., Moeslund, T.B., Li, T., Shen, W., Zhang, W., Li, W., Dai, W., Luo, W., Zhao, W., Zhang, W., Yang, X., Ma, Y., Joo, Y., Zeng, Y., Gan, Y., Zhu, Y., Zhong, Y., Ruan, Z., Li, Z., Huang, Z., Meng, Z., 2024b. Soccernet 2023 challenges results. Sports Engineering 27, 24. URL: https://doi.org/10.1007/s12283-024-00466-4, doi:10.1007/s12283-024-00466-4.

Cuevas, C., Quilón, D., García, N., 2020. Techniques and applications for soccer video analysis: A survey. Multimedia Tools and Applications 79, 29685–29721. URL: https://doi.org/10.1007/s11042-020-09409-0, doi:10.1007/s11042-020-09409-0.

Damoulaki, A., Ntzoufras, I., Pelechrinis, K., 2025. Lasso multinomial performance indicators for in-play basketball data: Lasso multinomial performance indicators for in-play... Comput. Stat. 40, 2157–2181. URL: https://doi.org/10.1007/s00180-025-01604-7, doi:10.1007/s00180-025-01604-7.

Davis, J., Bransen, L., Devos, L., Jaspers, A., Meert, W., Robberechts, P., Van Haaren, J., Van Roy, M., 2024. Methodology and evaluation in sports analytics: challenges, approaches, and lessons learned. Mach. Learn. 113, 6977–7010. URL: https://doi.org/10.1007/s10994-024-06585-0, doi:10.1007/s10994-024-06585-0.

Deliège, A., Cioppa, A., Giancola, S., Seikavandi, M.J., Dueholm, J.V., Nasrollahi, K., Ghanem, B., Moeslund, T.B., Van Droogenbroeck, M., 2021. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos, in: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 4503–4514. doi:10.1109/CVPRW53098.2021.00508.

Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A., 2010. The pascal visual object classes (voc) challenge. International Journal of Computer Vision 88, 303–338. URL: https://doi.org/10.1007/s11263-009-0275-4, doi:10.1007/s11263-009-0275-4.

Feichtenhofer, C., 2020. X3d: Expanding architectures for efficient video recognition, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 203–213.

Felsen, P., Lucey, P., Ganguly, S., 2018. Where will they go? predicting fine-grained adversarial multi-agent motion using conditional variational autoencoders, in: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (Eds.), Computer Vision – ECCV 2018, Springer International Publishing, Cham. pp. 761–776.

Feng, N., Song, Z., Yu, J., Chen, Y.P.P., Zhao, Y., He, Y., Guan, T., 2020. Sset: a dataset for shot segmentation, event detection, player tracking in soccer videos. Multimedia Tools Appl. 79, 28971–28992. URL: https://doi.org/10.1007/s11042-020-09414-3, doi:10.1007/s11042-020-09414-3.

Forcher, L., Altmann, S., Forcher, L., Jekauc, D., Kempe, M., 2022. The use of player tracking data to analyze defensive play in professional soccer - a scoping review. International Journal of Sports Science & Coaching 17, 1567–1592. URL: https://doi.org/10.1177/17479541221075734, doi:10.1177/17479541221075734, arXiv:https://doi.org/10.1177/17479541221075734.

Giancola, S., Amine, M., Dghaily, T., Ghanem, B., 2018. SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos , in: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE Computer Society, Los Alamitos, CA, USA. pp. 1792–179210. URL: https://doi.ieeecomputersociety.org/10.1109/CVPRW.2018.00223, doi:10.1109/CVPRW.2018.00223.

Giancola, S., Cioppa, A., Deliège, A., Magera, F., Somers, V., Kang, L., Zhou, X., Barnich, O., De Vleeschouwer, C., Alahi, A., Ghanem, B., Van Droogenbroeck, M., Darwish, A., Maglo, A., Clapés, A., Luyts, A., Boiarov, A., Xarles, A., Orcesi, A., Shah, A., Fan, B., Comandur, B., Chen, C., Zhang, C., Zhao, C., Lin, C., Chan, C.Y., Hui, C.C., Li, D., Yang, F., Liang, F., Da, F., Yan, F., Yu, F., Wang, G., Chan, H.A., Zhu, H., Kan, H., Chu, J., Hu, J., Gu, J., Chen, J., Soares, J.a.V.B., Theiner,

J., De Corte, J., Brito, J.H., Zhang, J., Li, J., Liang, J., Shen, L., Ma, L., Chen, L., Santos Marques, M., Azatov, M., Kasatkin, N., Wang, N., Jia, Q., Pham, Q.C., Ewerth, R., Song, R., Li, R., Gade, R., Debien, R., Zhang, R., Lee, S., Escalera, S., Jiang, S., Odashima, S., Chen, S., Masui, S., Ding, S., Chan, S.w., Chen, S., El-Shabrawy, T., He, T., Moeslund, T.B., Siu, W.C., Zhang, W., Li, W., Wang, X., Tan, X., Li, X., Wei, X., Ye, X., Liu, X., Wang, X., Guo, Y., Zhao, Y., Yu, Y., Li, Y., He, Y., Zhong, Y., Guo, Z., Li, Z., 2022. Soccernet 2022 challenges results, in: Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports, Association for Computing Machinery, New York, NY, USA. p. 75–86. URL: https://doi.org/10.1145/3552437.3558545, doi:10.1145/3552437.3558545.

Giancola, S., Cioppa, A., Gutiérrez-Pérez, M., Held, J., Hinojosa, C., Joos, V., Leduc, A., Magera, F., Sanchez, K., Somers, V., Xarles, A., Agudo, A., Alahi, A., Barnich, O., Clapés, A., Vleeschouwer, C.D., Escalera, S., Ghanem, B., Moeslund, T.B., Droogenbroeck, M.V., Abe, T., Alotaibi, S., Altawijri, F., Araujo, S., Bai, X., Bi, X., Cao, J., Chao, V., Czarnogórski, K., Deuser, F., Du, M., Feng, T., Frenzel, P., Fuchs, M., García, J., Habel, K., Hashiguchi, T., Hirose, S., Hu, X., Hwang, Y., Inoue, R., Itsuji, R., Iwai, K., Ji, H., Ji, Y., Jiao, L., Kageyama, Y., Kamikawa, Y., Kanasugi, Y., Kim, H., Kim, J., Kurihara, T., Li, B., Li, L., Li, X., Lian, Y., Liang, D., Lin, H., Lin, J., Liu, J., Liu, L., Liu, S., Liu, Z., Lu, Y., Méndez, F., Ma, H., Ma, W., Maksymiuk, J., Mantilla, H., Mathkour, I., Matthes, D., Motomochi, A., Muhammad, A.R., Nakayama, H., Oh, J., Oo, Y.M., Ortega, M., Oswald, N., Otsubo, R., Perez, F., Qi, M., Rey, C., Reyes-Angulo, A., Rose, O., Rueda-Chacón, H., Saito, H., Sarmiento, J., Sawafuji, K., Scott, A., Shen, X., Shrestha, P., Sim, J.Y., Sun, L., Sun, Y., Suzuki, T., Tang, L., Tonouchi, M., Uchida, I., Velesaca, H.O., Wang, T., Watanabe, R., Wu, J., Wu, Y., Yamagishi, S., Yang, D., Yang, X., Yang, Y., Ye, H., Ye, X., Yeung, C., Yu, X., Zhang, C., Zhang, D., Zhang, K., Zhao, Z., Zhou, X., Zhu, W., Ziegler, J., 2025. Soccernet 2025 challenges results. URL: https://arxiv.org/abs/2508.19182, arXiv:2508.19182.

Gutiérrez-Pérez, M., Agudo, A., 2025. Soccernet-v3d: Leveraging sports broadcast replays for 3d scene understanding, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, pp. 5977–5986.

Hartley, R., Zisserman, A., 2004. Multiple View Geometry in Computer Vision. 2 ed., Cambridge University Press, Cambridge.

He, K., Gkioxari, G., Dollár, P., Girshick, R., 2017. Mask r-cnn, in: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 2980–2988. doi:10.1109/ICCV.2017.322.

Hong, J., Zhang, H., Gharbi, M., Fisher, M., Fatahalian, K., 2022. Spotting temporally precise, fine-grained events in video, in: Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXV, Springer-Verlag, Berlin, Heidelberg. p. 33–51. URL: https://doi.org/10.1007/978-3-031-19833-5_3, doi:10.1007/978-3-031-19833-5_3.

Jiang, Y., Cui, K., Chen, L., Wang, C., Xu, C., 2020. Soccerdb: A large-scale database for comprehensive video understanding, in: Proceedings of the 3rd International Workshop on Multimedia Content Analysis in Sports, Association for Computing Machinery, New York, NY, USA. p. 1–8. URL: https://doi.org/10.1145/3422844.3423051, doi:10.1145/3422844.3423051.

Korte, F., Link, D., Groll, J., Lames, M., 2019. Play-by-play network analysis in football. Frontiers in Psychology Volume 10 - 2019. URL: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.01738, doi:10.3389/fpsyg.2019.01738.

Li, Y., Chen, L., He, R., Wang, Z., Wu, G., Wang, L., 2021. MultiSports: A Multi-Person Video Dataset of Spatio-Temporally Localized Sports Actions , in: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), IEEE Computer Society, Los Alamitos, CA, USA. pp. 13516–13525. URL: https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.01328, doi:10.1109/ICCV48922.2021.01328.

Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P., 2020. Focal loss for dense object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 42, 318–327. doi:10.1109/TPAMI.2018.2858826.

Loshchilov, I., Hutter, F., 2019. Decoupled weight decay regularization, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

Lucey, P., Bialkowski, A., Carr, P., Morgan, S., Matthews, I., Sheikh, Y., 2013. Representing and discovering adversarial team behaviors using player roles, in: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2706–2713. doi:10.1109/CVPR.2013.349.

Martens, F., Dick, U., Brefeld, U., 2021. Space and control in soccer. Frontiers in Sports and Active Living Volume 3 - 2021. URL: https://www.frontiersin.org/journals/sports-and-active-living/articles/10.3389/fspor.2021.676179, doi:10.3389/fspor.2021.676179.

Mendes-Neves, T., Meireles, L., Mendes-Moreira, J., 2024. Towards a foundation large events model for soccer. Mach. Learn. 113, 8687–8709. URL: https://doi.org/10.1007/s10994-024-06606-y, doi:10.1007/s10994-024-06606-y.

Minaee, S., Boykov, Y., Porikli, F., Plaza, A., Kehtarnavaz, N., Terzopoulos, D., 2022. Image segmentation using deep learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence 44, 3523–3542. doi:10.1109/TPAMI.2021.3059968.

Mortelier, A., Rioult, F., Komar, J., 2024. Design of a handball tactics observatory based on dynamic sub-graphs, in: Dong, J.S., Izadi, M., Hou, Z. (Eds.), Sports Analytics, Springer Nature Switzerland, Cham. pp. 149–166.

Ochin, J., Chekroun, R., Stanciulescu, B., Manitsaris, S., 2025a. Beyond pixels: Leveraging the language of soccer to improve spatio-temporal action detection in broadcast videos. arXiv preprint arXiv:2505.09455 .

Ochin, J., Devineau, G., Stanciulescu, B., Manitsaris, S., 2025b. Game state and spatio-temporal action detection in soccer using graph neural networks and 3d convolutional networks, in: Proceedings of the 14th International Conference on Pattern Recognition Applications and Methods - Volume 1: ICPRAM, INSTICC. SciTePress. pp. 636–646. doi:10.5220/0013161100003905.

Ogawa, Y., Umemoto, R., Fujii, K., 2025. Space evaluation at the starting point of soccer transitions. doi:10.48550/arXiv.2505.14711.

Peral, M., Capellera, G., Rubio, A., Ferraz, L., Moreno-Noguer, F., Agudo, A., 2025. Temporally accurate events detection through ball possessor recognition in soccer, in: Bashford-Rogers, T., Meneveaux, D., Ammi, M., Ziat, M., Jänicke, S., Purchase, H.C., Radeva, P., Furnari, A., Bouatouch, K., de Sousa, A.A. (Eds.), Proceedings of the 20th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, VISIGRAPP 2025 - Volume 2: VISAPP, Porto, Portugal,

February 26-28, 2025, SCITEPRESS. pp. 221–231. URL: https://doi.org/10.5220/0013317700003912, doi:10.5220/0013317700003912.

Raabe, D., Nabben, R., Memmert, D., 2022. Graph representations for the analysis of multi-agent spatiotemporal sports data. Applied Intelligence 53, 3783–3803. URL: https://doi.org/10.1007/s10489-022-03631-z, doi:10.1007/s10489-022-03631-z.

Rao, J., Wu, H., Jiang, H., Zhang, Y., Wang, Y., Xie, W., 2025. Towards Universal Soccer Video Understanding , in: 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society, Los Alamitos, CA, USA. pp. 8384–8394. URL: https://doi.ieeecomputersociety.org/10.1109/CVPR52734.2025.00785, doi:10.1109/CVPR52734.2025.00785.

Sha, L., Lucey, P., Zheng, S., Kim, T., Yue, Y., Sridharan, S., 2017. Fine-grained retrieval of sports plays using tree-based alignment of trajectories. ArXiv abs/1710.02255. URL: https://api.semanticscholar.org/CorpusID:6443784.

Simpson, I., Beal, R.J., Locke, D., Norman, T.J., 2022. Seq2event: Learning the language of soccer using transformer-based match event prediction, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, New York, NY, USA. p. 3898–3908. URL: https://doi.org/10.1145/3534678.3539138, doi:10.1145/3534678.3539138.

Singh, G., Choutas, V., Saha, S., Yu, F., Van Gool, L., 2023. Spatio-temporal action detection under large motion, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 6009–6018.

Sun, H., Xiong, Y., Wu, R., Wang, K., Zhang, L., Fan, C., Tang, S., Li, X.Y., 2025. Mvp-shapley: Feature-based modeling for evaluating the most valuable player in basketball. URL: https://arxiv.org/abs/2506.04602, arXiv:2506.04602.

Sun, Y., Sun, Z., Chen, W., 2024. The evolution of object detection methods. Engineering Applications of Artificial Intelligence 133, 108458. URL: https://www.sciencedirect.com/science/article/pii/S095219762400616X, doi:https://doi.org/10.1016/j.engappai.2024.108458.

Vračar, P., Štrumbelj, E., Kononenko, I., 2016. Modeling basketball play-by-play data. Expert Syst. Appl. 44, 58–66. URL: https://doi.org/10.1016/j.eswa.2015.09.004, doi:10.1016/j.eswa.2015.09.004.

Wang, P., Zeng, F., Qian, Y., 2023. A survey on deep learning-based spatio-temporal action detection. URL: https://arxiv.org/abs/2308.01618, arXiv:2308.01618.

Wang, Y., Sun, Y., Liu, Z., Sarma, S.E., Bronstein, M.M., Solomon, J.M., 2019. Dynamic graph cnn for learning on point clouds. ACM Trans. Graph. 38. URL: https://doi.org/10.1145/3326362, doi:10.1145/3326362.

Wang, Z., Veličković, P., Hennes, D., Tomašev, N., Prince, L., Kaisers, M., Bachrach, Y., Elie, R., Wenliang, L.K., Piccinini, F., Spearman, W., Graham, I., Connor, J., Yang, Y., Recasens, A., Khan, M., Beauguerlange, N., Sprechmann, P., Moreno, P., Heess, N., Bowling, M., Hassabis, D., Tuyls, K., 2024. Tacticai: an ai assistant for football tactics. Nature Communications 15, 1906. URL: https://doi.org/10.1038/s41467-024-45965-x, doi:10.1038/s41467-024-45965-x.

Yeung, C., Sit, T., Fujii, K., 2025. Transformer-based neural marked spatio temporal point process model for analyzing football match events: Transformer-based neural marked spatio temporal point process model... Applied Intelligence 55. URL: https://doi.org/10.1007/s10489-024-05996-9, doi:10.1007/s10489-024-05996-9.

Yu, J., Lei, A., Song, Z., Wang, T., Cai, H., Feng, N., 2018. Comprehensive dataset of broadcast soccer videos, in: 2018 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 418–423. doi:10.1109/MIPR.2018.00090.

Ötting, M., 2021. Predicting play calls in the national football league using hidden markov models. IMA Journal of Management Mathematics 32, 535–545. doi:10.1093/imaman/dpab005.