Passes, Receptions, Possession and Confidence: An Interpretable Approach to Soccer Event Detection

Marc Peral[1][0000−0001−6521−6476], Guillem Capellera[2][0009−0006−7266−078X], Antonio Rubio[2][0000−0002−6771−8645], Luis Ferraz[2][0000−0001−7851−9193], and Antonio Agudo[1][0000−0001−6845−4998]

[1] Institut de Robòtica i Informàtica Industrial CSIC-UPC, Barcelona, Spain
{mperal, aagudo}@iri.upc.edu
[2] Kognia Sports Intelligence, Barcelona, Spain {guillem.capellera, antonio.rubio, luis.ferraz}@kogniasports.com

Abstract. We present a modular and interpretable system for detecting and assigning passes and receptions in professional soccer matches using tactical camera footage. To this end, our method processes per-player video tubes to estimate possession likelihoods, identifies the ball possessor at each frame through temporal modeling, and detects touch events based on transitions in possession. This architecture achieves high temporal precision and player attribution without relying on tracking data or positional annotations. Compared to existing baselines, it improves performance on pass detection and accurately captures receptions, which are often underrepresented in prior work.
Beyond performance, we introduce and exploit epistemic uncertainty to assess the reliability of our model's predictions. Using deep ensembles and a certainty-aware classification loss, we quantify model confidence and identify ambiguous or out-of-distribution cases. We further integrate this uncertainty into the final stage of the pipeline, where it serves as a dynamic filtering mechanism that suppresses low-confidence detections and reduces false positives. This integration demonstrates that epistemic uncertainty can be leveraged not only for interpretability but also for improving robustness and decision quality.

Keywords: Events Detection · Action Spotting · Sports Analytics.

1 Introduction

The ability to detect and assign fine-grained events to specific players in soccer is essential for performance analysis, scouting, and tactical breakdown. This growth comes hand-in-hand with the multiplication of data acquisition in terms of players and ball positional information, video footage of games and gathering of statistics [5–8, 18, 45, 47]. Among the most frequently requested insights by coaching teams are what we call touch events as those happen when a player touches the ball, mainly passes and receptions. These interactions are requested

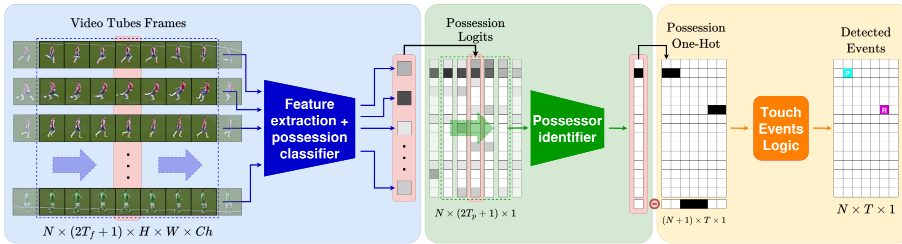

Fig. 1: **Overview of our event detection pipeline.** The model processes tactical video through three stages: video tubes are extracted and encoded to produce possession logits per player and frame (blue); the possessor is selected globally for each frame (green); passes and receptions are detected based on transitions in the possession sequence over a window of T frames (orange). Based on [34].

with precise temporal accuracy and clear player attribution, and occurring hundreds of times per match, form the backbone of most professional assessments. While sparse events such as goals or fouls are well-covered in existing benchmarks [13, 17, 50], dense and frequent actions remain less explored, despite being more tactically informative, due to their complexity of spotting automation.

Detecting such events presents several challenges. First, the task requires a temporal precision under a second, as the difference between a successful pass and a turnover can be a matter of milliseconds [31]. Second, the visual conditions are often imperfect. Tactical camera footage, while excellent for team-level spatial coverage, suffers from player occlusions or blurs, non-standardized lighting, and occasional tracking gaps [9]. Third, robust detection must go beyond recognizing when a pass occurs, also requiring the identification of the correct player involved [14, 36, 41, 45].

To address these challenges, we propose a modular pipeline that operates on video tubes, short spatio-temporal sequences centered on each player. Our method consists of three stages: 1) predicting possession likelihood from each tube (see Fig. 1, blue box), 2) identifying the per-frame possessor through temporal modeling (green box in the figure), and 3) detecting pass and reception events based on transitions in possession (orange box in the figure). This architecture enables both player localization and precise timing, while being robust to missing data and noisy detections. We train and evaluate our model using a newly annotated dataset of professional matches captured with tactical cameras, including thousands of touch events and frame-accurate annotations, and outperforming competing approaches.

In addition to performance, we explore the epistemic uncertainty of our system [2, 15, 24, 26, 27, 32, 38], that is, how confident it is in its own predictions. This aspect is particularly relevant in real-world deployments, where ambiguous inputs and distributional shifts are common. We estimate epistemic uncertainty via deep ensembles [27], and complement this with an uncertainty-aware clas-

Table 1: **Modalities of events detection methods.** Inputs are depicted as: ● - previous events class and their position in field, ■ - players and ball trajectories, ▲ - broadcast video, ▼ - tactical camera video, ◆ - method uses video tubes. Missing data column illustrates: ✓ - can work when missing players information either visual or positional, ✗ - needs dense information of all players or events. Player identification is represented as: ✓ - outputs which player is performing the event, □ - outputs region of the field where the event happens but not the player, ✗ - does not output player involved. Adapted from [34].

Paper	**Input**	**Missing data**	**Player identification ↓**	**Window [s]**
NetVLAD++ [17]	▲	✓	✗	5-60
Zhou *et al.* [50]	▲	✓	✗	5-60
Comedian [13]	▲	✓	✗	1-5
NMSTPP [48]	●	✗	□	-
Seq2Event [39]	●	✗	□	-
Vidal-Codina *et al.* [43]	● ■	✗	✓	20
BallRadar [25]	■	✗	✓	2
PassNet [40]	▲ ◆	✓	✗	1-4
Sanford *et al.* [37]	■ / ▼ ◆	✓	✗	1.7
Baikulov [1]	▼	✓	✗	1
Singer *et al.* [35]	▼	✓	✗	0.15-0.7
Ours	▼ ◆	✓	✓	0.6

sification inspired by evidential deep learning [38]. These methods allow us to quantify the model's confidence and identify samples that are visually ambiguous or underrepresented in training. Beyond analysis, we leverage this uncertainty signal to improve reliability: by using it as a dynamic filtering criterion in the final event detection stage, our system adaptively suppresses low-confidence detections, reducing false positives while preserving temporal precision.

An early version of this work was presented in [34], where we introduced our modular pipeline for temporally precise detection and attribution of touch events in soccer using tactical video. In this paper, we expand that contribution by providing a more detailed technical description of the architecture and extending the evaluation to include an in-depth study of epistemic uncertainty. We further demonstrate how this uncertainty can be operationalized within the pipeline itself, using it to refine event detection and enhance the robustness and interpretability of the overall system. Together, these extensions advance our framework toward a more reliable and trustworthy solution for large-scale, vision-based soccer analytics.

2 Related Work

The task of action detection in soccer has evolved rapidly in recent years, driven by increasing access to video data and the growing demand for fine-grained tactical analysis [33, 45]. Early approaches focused on segmenting sparse, high-level events such as goals or fouls [13, 17, 50], but more recent efforts have shifted toward identifying dense, temporally precise actions like passes and receptions [1, 25, 33, 35, 37, 39, 40, 43, 48], referred in this work as touch events. As stated in [16], events are defined as instantaneous in the soccer rulesbook. For example, a goal happens at the moment the ball crosses the goal line between the goalposts and the crossbar. For this reason, we consider events as occurrences in a certain frame.

Table 1 provides a comparative overview of existing methods across several key dimensions: the input modality, their robustness to missing data, and whether they identify the player responsible for the action. The table also includes a comparison of their temporal accuracy, expressed in terms of the event detection acceptance window considered in each work. As depicted, most approaches make trade-offs between interpretability, temporal resolution, and robustness. Our method aims to bridge these gaps yielding temporally sharp detections and player-specific assignments even under occlusion or missing data.

Data Observability. A central limitation of many action detection systems in soccer is their dependency on complete and densely annotated data. Several methods rely on prior events [39, 48] or trajectory information [25, 43] to make predictions. These approaches often assume continuous, high-quality tracking of all players and the ball, a condition rarely met in practical settings, especially when using broadcast video or tactical footage with occlusions.

In real-world scenarios, tracking failures are common: players may fall outside the field of view or become momentarily indistinguishable in dense scenes. Some models correct manually annotated data and achieve better synchronization [3]. Vision-based methods [1, 13, 17, 21, 35, 37, 40, 50] have generally been more resilient to such issues. These models often include mechanisms to handle partial visibility, such as frame downsampling, video padding, or aggregating global frame-level features. However, their outputs typically lack player-level resolution, which limits their utility for tactical evaluation. Trajectory-based approaches, by contrast, struggle with incomplete input. As noted by [19], trajectory-based pipelines are vulnerable to occlusion and require full visibility to infer possession or detect transitions. Models like BallRadar [25] or the work by Vidal-Codina *et al.* [43] achieve relatively coarse temporal resolution (2–20 seconds) and often fail to identify which player executed the action, especially under data loss.

Our method is designed with these challenges in mind. By focusing on dense video tubes centered on individual players, we mitigate the effect of missing or occluded data, allowing the model to operate robustly even when tracking is incomplete.

Sparse vs. Touch Events. In soccer event detection one can distinguish between sparse and touch events, depending on their frequency and tactical relevance. Sparse events refer to rare, game-defining moments such as goals, fouls, or penalties. These are widely studied because they are easier to annotate and often available in large-scale broadcast datasets such as SoccerNet-v2 [10]. Methods targeting these events [13, 17, 50] typically aim to provide a high-level summary of the game, focusing on semantic interpretation over temporal precision.

In contrast, touch events occur with much higher frequency and carry essential information for tactical evaluation. For example, during the 2020–2021 Premier League season, matches averaged just 2.7 goals and 2.9 yellow cards, compared to 940 passes and 696 receptions per game. These actions are fundamental for analyzing team formation, transitions, and player roles, especially in professional performance departments. Touch events are more challenging to detect, as they are tightly packed in time and require precise frame-level annotations. This complexity has moved the focus from sparse to touch events detection, leading to the emergence of dedicated tasks such as the Ball Action Spotting Challenge in SoccerNet [11]. As these new actions are closer in time, one cannot use broadcast video with changes in scenes, zooms, and replays, so the new dedicated dataset contained only tactical camera footage. Our approach works on the mentioned tactical camera videos to detect those touch events that provide the crucial data that soccer experts need.

Prediction with Past Events Sequences. Some previous work has approached event detection through sequence modeling, attempting to predict the next likely action based on the historical context of prior events. This is typically formulated using recurrent or transformer-based architectures trained on sequences of previously detected actions, often accompanied by positional or contextual metadata [39, 48].

While conceptually appealing, these approaches suffer from two major limitations. First, they often require dense annotations of prior events to construct the input sequence. In real scenarios, however, such data may be missing or inconsistent. Second, these models are prone to positional bias: since the input includes the spatial location of past actions, the predicted next event is heavily influenced by field zones. As a result, they tend to over-predict high-impact events like shots while underestimating the frequency of low-profile but tactically crucial actions like receptions.

Additionally, methods in this category do not usually identify the player responsible for the predicted event. Moreover, they do not offer precise timing, treating events as nodes in a sequence rather than properly finding the time of occurence. In contrast, our model is not dependent on previous event sequences. Instead, it makes local, frame-level decisions based on visual evidence. This makes it more flexible and deployable in real-world environments, where event history may be incomplete or noisy.

Detection with Players and Ball Location. Several works have approached event detection in soccer by leveraging the trajectories of players and the ball, relying on accurate spatial positioning to infer interaction dynamics [25, 43]. This line of research assumes access to high-resolution tracking data, often obtained through GPS sensors or multi-camera systems installed throughout the stadium [28]. These methods are attractive due to the structured nature of the data and the potential to encode movement patterns with interpretable features. However, trajectory-based models face significant limitations in both data availability and temporal accuracy. First, most datasets that include synchronized player and ball coordinates are proprietary and not publicly accessible. This restricts reproducibility and broader experimentation. Second, obtaining accurate trajectories from video alone typically requires detailed camera calibration and homography estimation, a process that is error-prone, especially when relying on tactical or broadcast footage [20].

Moreover, models based on trajectories often fail to meet the strict temporal resolution required for touch event detection. For example, BallRadar [25] achieves a temporal precision of roughly 2 seconds, while the method from Vidal-Codina *et al.* [43] operates with a 20-second tolerance. These delays are too coarse to support frame-level analysis and make it difficult to pinpoint the exact moment of contact or assign the action to a specific player. Our approach avoids the dependency on spatial coordinates altogether. By directly operating on video data and focusing on per-player regions, we maintain fine-grained temporal resolution without requiring any positional supervision or camera calibration.

Spotting with Full Frame from Video Footage. An alternative to coordinate-based approaches is to detect actions directly from video footage using image-based features extracted from full frames. These methods benefit from the dense visual context of the scene and can learn complex spatial patterns using convolutional networks. For example, Sanford *et al.* [37] conducted an ablation study comparing trajectory-based and image-based features, showing that visual information could outperform positional data when fine-tuned appropriately. Their model achieved a temporal accuracy of 1.7 seconds, with many correct detections clustering within 0.5 seconds of the ground truth.

A common drawback of full-frame methods is their reliance on downsampled video resolution and frame rate. To reduce computational cost, some approaches lower the frame rate to 5 Hz [40] or stack grayscale neighbor frames into artificial RGB inputs [1, 35], which can reduce motion sensitivity. Moreover, these models typically operate at the scene level, making it difficult to determine which player executed the action. Even if the timing is reasonably accurate, the lack of player-level granularity limits the practical use of these systems for tactical feedback or player evaluation.

Locating with Video Tubes. To address the limitations of full-frame analysis, previous work has introduced video tubes as a mechanism for extracting localized spatio-temporal features around specific players [49]. A video tube captures a sequence of cropped bounding boxes following the same player across several frames, effectively isolating their motion and interactions from the rest of the scene. This approach has been shown to improve dynamic action recognition, particularly under moving camera conditions.

In the context of soccer, Honda *et al.* [23] employed video tubes combined with trajectory data to predict pass receivers. While effective in controlled settings, their method was limited to sequences where all players were visible and considered only successful passes. Other works [46] used video tube crops centered on the ball region, applying transfer learning from a ball detection pretask. This focus helped improve detection performance but did not yield player identities. PassNet [40] partially addressed this by applying object detection to identify players near the ball and extract their features. However, their method reduces input quality through aggressive downsampling and does not localize the action to specific players in a structured way. Similarly [37] included bounding box information in their model but aggregated features across all visible players without resolving which individual performed the action.

Our approach advances this direction by generating a video tube for every visible player in each frame and processing them independently to compute possession likelihoods. This not only enables high temporal resolution but also provides a structured framework for identifying which player executed each action.

Uncertainty in Vision-Based Models. In recent years, the quantification of epistemic uncertainty in deep learning has garnered increasing attention [2, 15, 24, 26, 27, 32, 38]. Often referred to as model or knowledge confidence, epistemic uncertainty captures the lack of knowledge about the correct model parameters due to limited data. Unlike aleatoric uncertainty which arises from the inherent noise in the data and is irreducible, epistemic uncertainty can be reduced with more informative or diverse training data.

A foundational and widely adopted approach is the use of deep ensembles, as introduced by [27]. In this method, multiple independently trained models are used to form an ensemble whose output variance is interpreted as a proxy for epistemic uncertainty. This technique is robust, scalable, and empirically effective in both regression and classification settings, and despite being quite costly, it avoids some of the computational intractabilities associated with full Bayesian neural networks. Another line of work leverages Monte Carlo dropout, originally proposed by [15], where dropout is applied at test time to simulate an ensemble-like behavior in a single model. Although more efficient, it tends to underestimate uncertainty in high-dimensional and highly confident regions [2].

Moving beyond approximate Bayesian methods, Sensoy *et al.* [38] introduced a principled Evidential Deep Learning model that interprets the softmax outputs as parameters of a Dirichlet distribution, enabling it to explicitly represent uncertainty about class probabilities. The model outputs evidence for each class, and the concentration of the Dirichlet distribution reflects the confidence of predictions, leading to high epistemic uncertainty when there is low evidence. This allows the model abstention in ambiguous or out-of-distribution scenarios.

These methods are highly relevant to our domain, where ambiguous visual situations and rare gameplay configurations are frequent. In such settings, quantifying epistemic uncertainty is not only crucial for understanding when and why a model may fail, but also for improving the reliability of its decisions. In our work, we build upon these ideas by integrating epistemic uncertainty into both the analysis and operation of our system: beyond identifying failure modes, we use uncertainty as a dynamic filtering criterion in the final event detection stage. This enables the model to suppress low-confidence detections adaptively, thereby improving the precision and trustworthiness of the resulting touch event predictions.

3 Method

Our goal is to detect temporally accurate touch events in soccer matches, specifically passes and receptions, and to identify the player responsible for each. To achieve this, we design a three-stage architecture that processes tactical video footage and player bounding boxes as input. Each stage in the pipeline addresses a core component of the problem: estimating which player is in possession of the ball at each frame, determining when the ball transitions between players, and ultimately producing temporally accurate touch event detections.

The first stage operates at the level of individual players. For each detected player in a frame, we extract a video tube and predict the likelihood that they are in possession of the ball (see the blue block in Fig. 1). In the second stage, we aggregate all per-player predictions to determine the single possessor (or the absence of one) at each frame, using contextual and temporal cues from adjacent frames (green block in the figure). Finally, the third stage detects passes and receptions by analyzing changes in possession across consecutive frames (orange block).

This modular design enables robust performance in the presence of occlusions, missing detections, and visual ambiguity. Moreover, it naturally supports the integration of epistemic uncertainty estimation within the first stage, allowing the system to quantify its confidence at the tube level. These uncertainty estimates are then exploited downstream to refine the final event detections, dynamically filtering low-confidence predictions and improving overall reliability without altering the underlying pipeline structure.

3.1 Video Tubes Extraction

The first step in our pipeline focuses on extracting consistent visual representations for each player across time. Given a video frame and the corresponding set of player bounding boxes, we crop a short spatio-temporal sequence around each visible player, which we refer to as a video tube. This sequence captures both the appearance of the player and their short-term motion [49], which is often informative for possession-related actions.

To build a video tube, we define a temporal collar of T_f frames before and after the current one. For each target frame, we gather the bounding boxes for the same player across this $2T_f + 1$ frames window. To standardize the region of interest and capture useful context, we enlarge the bounding box to form

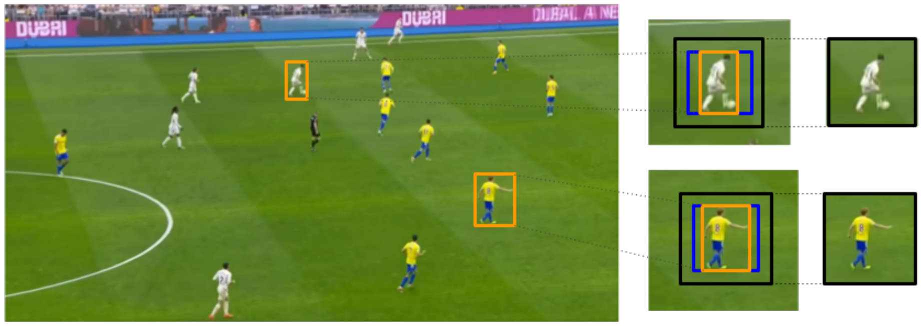

Fig. 2: **Video tube extraction. Left:** Bounding boxes are extracted for each player in the frame (orange boxes). **Middle:** The selected player's box is enlarged to a square and padded for normalization (dark blue and black squares, respectively). **Right:** Final cropped region used to construct the video tube. Redrawn from [34].

a square, add a 20% margin around it, and resize the result to 128×128 pixels. Figure 2 illustrates this process. The temporal window size T_f is set to 3, based on an empirical observation that ball movement typically enters or exits the player's bounding box within three frames when touch events happen. This aligns with prior work [23], which shows that longer visual sequences can dilute discriminative cues due to excessive context, while shorter sequences may fail to capture the relevant interaction. This stage yields one video tube per visible player in the frame, forming the input for the possession likelihood classifier.

3.2 Possession Likelihood from Video Tubes

With a video tube extracted for each player, the next step is to estimate the likelihood that the tube depicts a player in possession of the ball. This is framed as a binary classification task: for each tube, the model predicts whether the corresponding player is in control of the ball or not.

We use a ResNet-50 backbone [22] to extract features from each frame within the video tube. To capture subtle temporal variations such as slight differences in body posture or leg motion, we integrate a Temporal Shift Module (TSM) [29] into the network. This lightweight module shifts feature channels across adjacent time steps, enabling the model to learn temporal dynamics without increasing computational complexity. The inclusion of TSM is particularly beneficial here, as our videos are recorded at full frame rate and thus contain fine-grained motion cues.

The features from all tube frames are concatenated and passed through a fully connected layer to aggregate temporal information into a single 256-dimensional embedding. A final dense layer then projects this embedding to a two-dimensional output: one score for the player being in possession and one for not being in possession. These logits are interpreted as class probabilities after applying softmax.

To improve discriminative power, we replace the standard softmax cross-entropy with the Additive Angular Margin (AAM) loss [12], which was originally proposed for face recognition tasks. AAM encourages better separation between classes by enforcing both inter-class separation and intra-class compactness. This is particularly useful in our setting, where the visual appearance of players can vary greatly within the same class across teams wearing diverse color clothes and soccer fields lighting conditions.

Formally, the AAM loss for a sample x_i belonging to class y_i is defined as:

$$\mathcal{L}_f = -\log \frac{e^{s \cdot cos\left(\theta_{y_i} + m\right)}}{e^{s \cdot cos\left(\theta_{y_i} + m\right)} + \sum_{j=1, j \neq y_i}^{N} e^{s \cdot cos \theta_j}}, \quad (1)$$

where θ_j denotes the angle between the feature vector x_i and class center W_j, s is a scaling factor set to 1 and m the angular margin set to 0.5. This formulation distributes the learned embeddings on a hypersphere, where confident predictions correspond to well-separated clusters.

This stage yields a possession likelihood for every visible player in the frame, forming the foundation for the next component, which determines the actual possessor, if any, from all the players in each frame.

3.3 Per-Frame Possessor Identification

Once possession likelihoods are computed independently for each video tube, we move from isolated player assessments to the full-frame context. This is a structured prediction problem where the input consists of a set of possession scores, one per player, for each frame. Crucially, these scores are not interpreted in isolation. Instead, we leverage a short temporal window around each frame to capture the dynamics of possession change and to account for occasional fluctuations in model confidence. Specifically, for each frame t we consider a temporal collar from $t - T_p$ to $t + T_p$, forming a matrix of possession likelihoods with shape $(2T_p + 1) \times N$, where N is the number of players. Checking the slope that possession likelihoods show when changing from possession to not possession or vice versa, we observe that it takes five frames to completely toggle, so we adopt two frames as the value for our parameter T_P, which balances responsiveness with stability. In practice, some players may be temporarily occluded or out of view. For frames in which a player's tube is unavailable, we assign a default likelihood of zero. To reduce the effect of spurious spikes, we smooth the likelihood signals using a Gaussian filter across time before feeding it into the model stage.

To infer the possessor from these scores, we design a network that takes into account both temporal continuity and inter-player competition. We build on Conv-TasNet [30], originally introduced for speech separation, which models temporal dependencies via 1D convolutions. This choice is particularly appropriate here, as it enables the network to disentangle the temporal possession patterns of individual players in a way analogous to separating overlapping voices.

To enhance temporal context further, we include a Time-Delay Neural Network (TDNN) [44], which processes the smoothed likelihoods with shift-invariant filters over time. This architecture improves the model's ability to detect consistent possession transitions, even in noisy or rapidly changing scenes.

The network concludes with two fully connected layers. The first expands the feature dimension to allow for modeling player-to-player interactions. The second compresses the output to a one-hot vector of size $N+1$, where each entry corresponds to one of the N players or to a background class indicating that no player has possession. We train this component using standard cross-entropy loss between the predicted and ground truth possessor class. The final output is a dense, frame-level sequence that identifies the player in control of the ball, or detects when the ball is unclaimed. This refined possession signal enables precise event detection in the final stage of our pipeline.

3.4 Touch Events Detection

With a reliable per-frame estimate of the ball possessor in place, the final stage of our pipeline detects the precise timing of touch events, that are essential building blocks for downstream tactical analysis. The detection logic operates on the sequence of predicted possessors across frames. A reception is triggered when a player begins possessing the ball, while a pass is registered when that same player relinquishes it. However, in practice, directly applying this rule often leads to false positives. For instance, a brief or noisy change in the predicted possessor may create spurious, non-existent events. To mitigate such artifacts, we incorporate two temporal filters designed to refine the detection. The first filter, controlled by a parameter T_s suppresses cases where a player gains and loses possession within a very short time interval. This is common in quick first-touch passes, where expert annotations typically record only the pass but not the reception. By omitting receptions shorter than $T_s = 7$ frames, we better align with annotation conventions and reduce redundant detections.

The second filter was originally governed by the temporal threshold $T_e = 3$ frames, rejecting possessions too brief to be meaningful. Following our analysis of epistemic uncertainty, we replace this fixed temporal constraint with an uncertainty-based adaptive filter. Instead of relying solely on duration, the new mechanism discards events associated with high uncertainty in the first-stage predictions. In practice, possessions corresponding to samples with uncertainty above 90% are filtered out, effectively removing low-confidence detections that would otherwise appear as false positives. This dynamic filtering strategy adapts to each situation, balancing precision and recall according to the model's own confidence.

The result is a list of timestamped events, each labeled by its type (pass or reception) and the player responsible. This rule-based stage remains deliberately lightweight and interpretable, relying on the structure of possession transitions while integrating uncertainty as a reliability signal. The addition of the uncertainty-aware filter enhances robustness against ambiguous or noisy inputs, improving the temporal and semantic consistency of the detected events.

3.5 Epistemic Uncertainty

While our pipeline achieves strong results in terms of temporal precision and player identification, its reliability in ambiguous scenarios remains difficult to interpret through accuracy metrics alone. Soccer footage often contains complex or occluded situations where even human observers may hesitate. To better characterize model reliability in such cases, we incorporate an analysis of epistemic uncertainty, which reflects the model's lack of knowledge or familiarity with a given input. Unlike aleatoric uncertainty, which arises from irreducible noise in the data (e.g., motion blur or poor lighting), epistemic uncertainty can be reduced through better model generalization or more representative training samples. In our case, it reveals whether uncertainty stems from inherent visual ambiguity or from unfamiliar situations beyond the model's training distribution.

To estimate this uncertainty, we first adopt a deep ensemble approach [27]. We train multiple instances of the stage one classifier, each initialized with different random seeds and exposed to shuffled data, and compute the variance of their predicted possession likelihoods for each video tube. This variance acts as a direct measure of model's confidence: low variance indicates consistent predictions across ensemble members (high confidence), while high variance signals disagreement and thus uncertainty about the input.

We further explore a parametric alternative to capture uncertainty directly within the model. Following the formulatio in [38], we replace the standard classification head with an uncertainty-aware head, that produces a vector of class scores alongside a scalar uncertainty value. These components are constrained to sum to one as:

$$u + \sum_{k=1}^{K} b_k = 1, \tag{2}$$

where K is the number of classes, $u \geq 0$ is the total uncertainty and $b_k \geq 0$ the belief mass for class k, computed as:

$$b_k = \frac{e_k}{\sum_{i=1}^{K}(e_i + 1)} \text{ and } u = \frac{K}{\sum_{i=1}^{K}(e_i + 1)}, \tag{3}$$

where $e_i \geq 0$ is the evidence on class i. During training, this formulation encourages the model to assign most of the probability mass to a single class for confident samples, while allowing ambiguous cases to allocate weight to the uncertainty term. As a result, the model learns to express hesitation in a principled way rather than overcommitting to noisy predictions.

Incorporating epistemic uncertainty estimation at the first stage of our pipeline allows both downstream components and human analysts to act based on the model's self-assessed confidence, not just its predicted class. This integration adds a critical layer of interpretability and robustness, especially under real-world variability. Furthermore, we exploit these uncertainty estimates in the final stage of our pipeline, replacing the fixed temporal filtering by an uncertainty-based adaptive filter (see Section 3.4). By filtering out predictions with high uncertainty the system dynamically suppresses false detections while maintaining temporal accuracy. This mechanism demonstrates how epistemic uncertainty can evolve from an interpretability tool into a functional component that improves the reliability and trustworthiness of the overall event detection process.

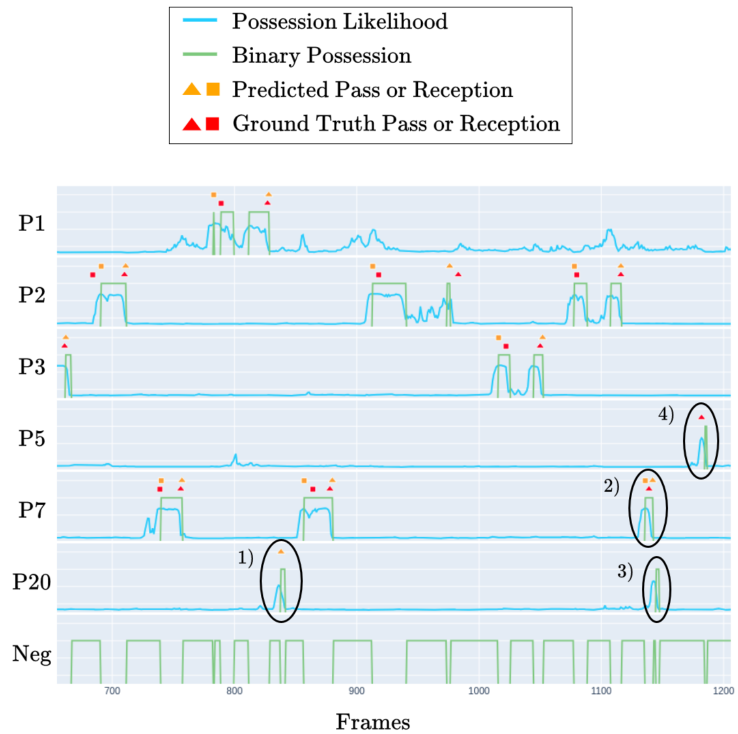

Fig. 3: **Model output and limitations.** The graph shows the model per-frame predictions against the ground truth only for relevant players in a chunk of 22 seconds. A false positive can be observed in 1), a correctly filtered spurious detection in 3) and a false negative missed in 4). In 2) a first touch pass was detected as a multiple contact reception-pass. Modified from [34].

4 Experiments

In this section, we analyze our system's performance and behavior in detail. We begin by examining its limitations and exploring the role of epistemic uncertainty in understanding model confidence and failure modes. This analysis provides valuable insight into the conditions under which the model struggles and how uncertainty can be leveraged to improve downstream predictions. Following this, we evaluate each stage of our architecture independently to quantify its contribution to the overall pipeline. Finally, we compare our complete system against state-of-the-art methods on touch event detection, demonstrating competitive or superior temporal precision and interpretability, even when other approaches do not explicitly model player identity or event granularity.

4.1 Model Limitations and Epistemic Uncertainty

To better understand the limitations of our system, we perform a qualitative analysis of the model's output before adding uncertainty strategies, highlighting both typical errors and ambiguous scenarios. Figure 3 shows a 550-frame segment (22 seconds) from a test match, annotated with the full pipeline output across several players. The chunk starts with a pass (specifically, a throw-in) performed by the player P3 and follows a possession of the players in the same team. Some players are clearly not in possession of the ball during this chunk, so by removing those we focus only on relevant players. This example includes correct predictions as well as various failure modes. In Fig. 3.1), a false positive reception is detected, triggered by a momentary shift in possession scores, but in Fig. 3.3) a spurious detection is correctly discarded. Figure 3.2) shows a common edge case: a first-touch pass that the model erroneously splits into a reception followed by a pass. Finally, Fig. 3.4) illustrates a false negative pass that is missed entirely, as the signal was too brief and uncertain for the model to register confidently.

These error cases raise an important question: are these predictions incorrect because the model is confident and wrong, or because the input is ambiguous and the model is appropriately uncertain? To answer that, we estimate the epistemic uncertainty of the first stage of the model by training multiple instances of the video tube classifier (each with different weight initializations and shuffling seeds) and measuring the variance in their outputs. For each video tube, we compute the variance of the predicted possession likelihood in the ensemble. Tubes with low variance correspond to well-understood, high-confidence samples, while high variance reveals ambiguity or unfamiliar visual patterns. We then sort the test set tubes by uncertainty and examine the extremes.

The lowest-variance samples are typically clean views of players, for example for negative samples in Fig. 4,b2-4). For positive samples, Fig. 4,b1) shows how midday shades help with the contrast of the white ball on the shadowed dark green grass. On the other hand, high-variance samples tend to fall into one of the following categories: goalkeepers Fig. 4,a1-2), players controlling the ball in strange postures Fig. 4,a3) or multiple players overlapping in the same tube Fig. 4,a4-5). These situations have clearly a low representation on the dataset, therefore it makes sense that those become the most challenging samples to certainly predict.

To further explore this, we modify the classification head in stage one to use an uncertainty-aware loss, where the network learns not only possession class scores but also a dedicated uncertainty score. Inspired by [38], this formulation forces the model to distribute its output between known classes and an explicit

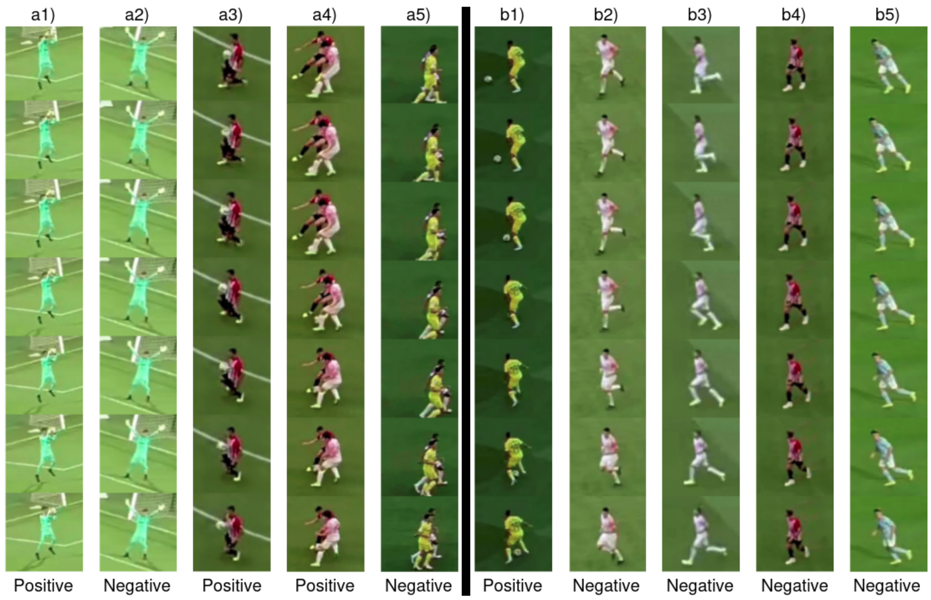

Fig. 4: **Qualitative analysis of epistemic uncertainty** Samples with highest variance across ensemble predictions are shown in the left (a1–a5), while those with the lowest variance appear in the right (b1–b4). The ground truth label is shown at the bottom of each tube. High-uncertainty cases include goalkeepers (a1–a2), unusual postures during ball control (a3), and tubes containing multiple players overlapping (a4–a5). In contrast, low-uncertainty examples typically show clear, unobstructed player views (b2–b4) or visually distinct ball conditions (b1), even if such configurations are underrepresented in the training set.

uncertainty channel, which we constrain to sum to one. Although this variant achieves slightly lower classification accuracy in first stage task as we will show in the next subsection, it enables direct access to epistemic uncertainty as part of the output. When applied to the test set, as depicted in Fig. 5, the majority of samples show moderate uncertainty around 24%, while a small but significant fraction exhibit near-total uncertainty. Interestingly, many of these 100% uncertain samples coincide with the false positives and false negatives shown earlier in Fig. 4b), confirming that model confidence is a strong indicator of prediction reliability.

Previous analysis open the door to incorporating uncertainty thresholds for downstream filtering or analyst review, especially in production settings where trust in the system is critical. We leverage this insight by integrating the estimated epistemic uncertainty into the final stage of our pipeline, replacing the fixed temporal filtering with a dynamic, uncertainty-based mechanism that adapts to each sample. In this setup, uncertainty acts as a confidence regulator, allowing the system to automatically suppress low-confidence detections while

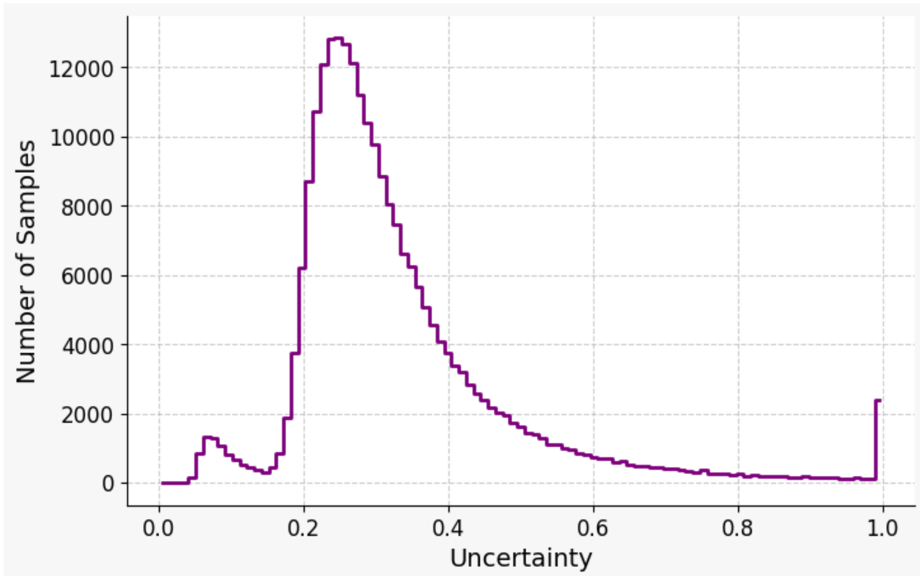

Fig. 5: **Distribution of epistemic uncertainty.** Most test samples cluster around a moderate uncertainty level (25%), while a smaller subset exhibits extreme uncertainty (close to 100%). These high-uncertainty cases often align with incorrect predictions, highlighting the potential of this signal for downstream filtering.

retaining reliable ones. By filtering out predictions associated with high uncertainty scores (above 90%), the overall precision of event detection improves, as shown at the end of this section. This demonstrates the practical value of epistemic uncertainty not only as an interpretability tool but also as a functional signal for improving robustness and decision quality in downstream tasks.

4.2 Possession Likelihood from Video Tubes

The first stage of the pipeline has the goal of classifying whether a spatio-temporal video tube corresponds to a player in possession of the ball. We construct a dedicated dataset by sampling tubes from frames containing annotated touch events. For each such frame, we crop a video tube for every visible player, labeling the one holding the ball as positive and the rest as negative. We exclude tubes where the player is not fully visible in all context frames.

The resulting dataset is highly imbalanced, with approximately 20 negative tubes for every positive one, totaling 75,683 positive and 1,516,586 negative samples. To address this imbalance, we adopt a balanced sampling strategy during training: each batch is composed of 50% positives and 50% randomly sampled negatives. At the end of each epoch, unused negative samples are discarded.

Table 2: **Possession likelihood from video tubes analysis.** Precision, recall and AUROC curve scores are reported for multiple variants using ResNet-18 and ResNet-50 backbones [22]. Adapted from [34], it includes our certainty-aware estimates which maintain strong performance while enabling epistemic uncertainty analysis.

Model	**Shift Module**	**Uncertainty**	**ResNet**	**Precision**	**Recall**	**AUROC**
Timeless	-	No	18	73.12	92.80	99.10
		No	50	77.58	92.87	99.14
Ours	-	No	18	82.84	97.90	99.78
		No	50	85.85	97.59	99.78
	TSM	No	18	80.73	97.67	99.76
		No	50	<u>86.29</u>	**98.01**	**99.81**
		Yes	50	78.79	97.42	99.56
	GSM	No	18	84.41	97.77	99.79
		No	50	**88.12**	<u>97.85</u>	**99.81**

To validate the importance of temporal context and the benefit of motion-aware feature extraction, we train several variants of our model:

- **Timeless**: Only the central frame of the tube, that is, no temporal context.
- **Ours without shift module**: Employs the full video tube with standard ResNet (18 or 50) [22] feature extractor.
- **Ours + TSM** [29]: Adds a Temporal Shift Module (TSM) to the baseline for efficient temporal reasoning.
- **Ours + GSM** [42]: Uses a Gate-Shift Module (GSM) instead of TSM to modulate temporal features.

First stage classification task results using precision, recall, and Area Under Receiver Operating Characteristic (AUROC) as metrics are summarized in Table 2. We observe the following:

- The Timeless variant underperforms compared to full tube-based models. Therefore, temporal context is crucial.
- Deeper backbones (ResNet50 [22]) consistently outperform ResNet18 [22] in all settings.
- Incorporating temporal shift modules improves performance, with TSM [29] offering the best recall, which is particularly desirable at this stage to avoid missing true possession events downstream.
- We applied the modification explained in Section 3.5 to the best performing configuration to obtain uncertainty values. This achieves inferior precision and recall in this first task classification but provides uncertainty values to be used in downstream tasks.

4.3 Per-Frame Possessor Identification

The second stage in our pipeline aims to determine which player, if any, has possession of the ball at every frame of the match. Unlike the previous task, where it was estimated from isolated video tubes, this step leverages the possession likelihoods from all players in a frame to produce a single per-frame prediction.

To train and evaluate this component, we construct a new dataset where ground truth labels indicate the possessor in each frame. These labels are derived from the annotated touch events: a player is considered in possession from their first touch until their last touch. For the frames when the ball is in transition between players, and therefore does not have a possessor, we assign a background or negative class. We reduce the number of matches because now, unlike for the previous task, we can use all the frames when the ball is in play, that is, when the game is not stopped. For this task, we use a subset of 5 matches: 3 for training, 1 for validation, and 1 for testing.

We employ multiple metrics to assess this component. While overall accuracy provides a basic performance measure, it fails to reflect how well the model captures continuous segments of possession. For this reason, we adopt two additional metrics widely used in speaker diarization: purity and coverage. These segment-level metrics allow us to assess how precisely the predicted possession intervals align with the reference. According to [4], given $\mathcal{R}$ the set of reference possessor segments for our task and $\mathcal{H}$ the set of hypothesized segments, coverage is defined as:

$$\text{coverage}(\mathcal{R}, \mathcal{H}) = \frac{\sum_{r \in \mathcal{R}} \max_{h \in \mathcal{H}} |r \cap h|}{\sum_{r \in \mathcal{R}} |r|}, \tag{4}$$

where $|r|$ is the duration of segment r and $r \cap h$ is the intersection of segments r and h. Purity is computed analogously by interchanging $\mathcal{R}$ and $\mathcal{H}$ in Eq. (4). Coverage assesses how much of each ground truth segment is correctly recovered, while purity measures the degree to which predicted segments are consistent with the true ones. Together, they offer insight into whether the model tends to under- or over-segment possession phases.

In Table 3 we compare several variants of our model to understand the impact of temporal modeling. A simple baseline chooses the player with the highest possession likelihood, provided it exceeds a fixed threshold. This approach is fast and intuitive but ignores temporal consistency. Our initial attempt at a neural network classifier without temporal context, relying only on dense layers over likelihoods, performed worse than this baseline, showing that the problem cannot be solved by pointwise classification alone. We then incorporate temporal reasoning by adding two modules inspired by audio signal processing: Conv-TasNet, which uses 1D convolutions to model interactions over time, and a Time-Delay Neural Network (TDNN), which enhances the model's ability to detect time-shifted possession transitions. The combination of these two modules results in a significant improvement, particularly in purity, indicating that the model predicts cleaner, more coherent segments of possession.

To further stabilize predictions, we apply a Gaussian smoothing filter to the raw likelihoods before feeding them into the network. This simple addition re-

Table 3: **Evaluation of per-frame possessor identification.** We report purity and coverage of predicted possessor sequences, as well as frame-level accuracy. Results are shown with and without Gaussian filtering. Our full model with CTN+TDNN outperforms the baselines across all metrics, particularly when combined with temporal smoothing. Adapted from [34].

Model	**Audio Modules**	**Gaussian Filter**	**Purity**	**Coverage**	**Accuracy**
Max + Threshold	✗	✗	49.87	66.96	70.12
		✓	52.65	<u>67.50</u>	70.60
Ours	✗	✗	44.08	59.83	60.90
		✓	44.19	59.74	60.86
Ours	CTN+TDNN	✗	**57.03**	66.87	<u>70.95</u>
		✓	<u>56.61</u>	**67.79**	**71.88**

Table 4: **Frame-wise pass detection comparison.** Precision, recall, and F1-score are reported for identifying whether each frame belongs to a pass segment (Pass) or not (No Pass), treating passes as temporal intervals defined by start and end frames. Reproduced from [34].

Model	**Pass**			**No Pass**		
	Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**
PassNet [40]	43.13	**72.24**	54.01	70.57	41.22	52.04
Ours	**69.95**	52.74	**60.14**	**74.46**	**86.10**	**79.86**

duces spurious peaks and improves both accuracy and alignment with the ground truth as depicted in Table 3. Altogether, this experiment highlights the importance of temporal and contextual awareness in accurately identifying possession. The task is not merely a matter of comparing likelihoods frame by frame, but of understanding the underlying dynamics across time and players.

4.4 Touch Events Detection

Having determined the ball possessor in each frame, we now aim to identify the precise moments when touch events, passes and receptions, occur. This stage leverages the per-frame possessor sequence and applies a rule-based logic to detect changes in control that correspond to meaningful interactions with the ball. Evaluations on this subsection are carried out on a new test set consisting of six matches from the 2023–2024 LaLiga season, never before seen by any of the algorithms, covering various stadiums, field sizes, and weather conditions.

First, we compare our approach with PassNet [40], a model that detects passes based on full-frame features and object detection. Their dataset is not available, but the code for their method is, so we train it in nine of our dataset matches (225% times the data in their original train set). However, PassNet [40] defines passes as extended segments with a start and end frame, rather than

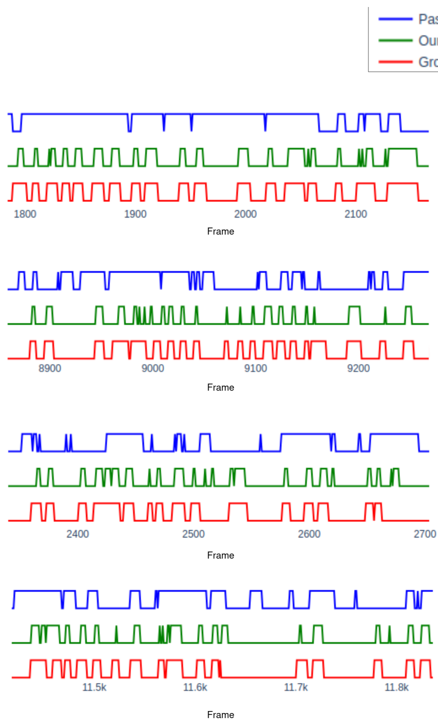

Fig. 6: **Qualitative comparison with PassNet. Left:** Frame-wise pass predictions from PassNet [40] and our model compared to ground truth across several match segments. **Right:** Zoomed-in views of selected intervals highlight our model's improved alignment with ground truth, producing sharper and more accurate event boundaries.

discrete events. To enable comparison, we try to train their algorithm on instantaneous events to use our metrics for comparison, but PassNet [40] only learns to put every frame to negative due to the imbalance in frames. Therefore, we adapt our predictions to match their segment definition: frames following a pass detection are labeled as belonging to the "pass" class, while frames after receptions are treated as negatives. Even under this setup, one can observe in Table 4 how our model outperforms Sorano *et al.* in all relevant metrics, except for recall on positive frames, where PassNet [40] tends to over-predict pass segments and therefore gets a higher value. This behavior results in high recall but low precision, and more importantly, makes it impossible to count the number of actual passes, an essential requirement for soccer analysts.

Qualitative examples in Fig. 6-left illustrate how our model aligns closely with ground truth in both timing and frequency of events. On the right we include several zoomed-in examples where our model captures timing and transitions of ball movement more faithfully, while PassNet [40] tends to over-segment

Table 5: **Comparison on touch events detection.** We reproduce Baikulov *et al.*'s [1] 7-fold evaluation setup, reporting precision, recall, F1 at 0.6s and 1s acceptance windows and mAP@1 for pass detection only. Fold-wise scores and their mean are shown. Our methods are evaluated jointly for both passes and receptions on the same test set. Receptions are highlighted as the evaluation is done separately for each event type.

Model	Fold	Event	0.6s			1s			
			Prec	**Rec**	**F1**	**Prec**	**Rec**	**F1**	**mAP@1**
Baikulov [1]	0	Pass	71.40	64.71	67.89	76.50	70.01	73.11	33.37
	1		86.32	37.37	52.16	90.08	39.98	55.38	20.53
	2		79.65	47.66	59.64	83.74	51.62	63.86	27.06
	3		78.15	54.70	64.36	82.16	58.98	68.67	31.66
	4		79.01	54.42	64.45	82.03	57.88	67.87	29.54
	5		74.19	53.93	62.46	78.32	58.30	66.84	30.38
	6		71.36	53.74	61.31	75.73	58.34	65.91	31.06
	Mean	Pass	77.15	52.36	62.38	81.22	56.44	66.60	29.09
Ours w/o Uncertainty	-	Recep	54.23	**57.53**	55.83	64.96	**67.12**	66.02	**38.36**
		Pass	70.12	**64.67**	67.28	77.62	**71.33**	74.34	**44.15**
Ours	-	Recep	**57.68**	55.49	**56.56**	**66.83**	65.96	**66.39**	35.67
		Pass	**78.67**	61.73	**69.18**	**83.14**	69.85	**75.92**	40.74

event boundaries. These visualizations confirm that our system, trained with instantaneous annotations and explicit possessor tracking, captures not only the occurrence but also the identity of each event more reliably.

To further demonstrate that, we do a final comparison of our model with the approach by Baikulov [1], the winner of the SoccerNet Ball Action Spotting Challenge [11], which is trained to detect passes and drives defined as in our main task as instantaneous events. As their model does not explicitly identify player identities, and there is not a semantical mapping between receptions and drives, we evaluate only pass detection. Their method employs a 7-fold cross-validation scheme during its initial pretraining stage, resulting in a collection of seven distinct models. To ensure a fair and consistent comparison, we replicate their evaluation strategy: each model is run independently on our six-match test set, and we report both the individual results and the arithmetic mean of their outputs, following their original protocol. Considering that events are temporally sharp, we assess prediction accuracy using acceptance windows of 0.6 and 1.0 seconds. A detection is considered correct if it falls within the specified window around the ground truth frame. In addition to precision and recall, we report the mean Average Precision at 1s window (mAP@1), the main evaluation metric used in the state of the art. Under this formulation, a predicted event is considered correct if it falls within a window of δ seconds around the ground truth. A precision-recall curve is computed for varying δ, and the average across event types yields the final mAP score. While the original challenge [11] evaluated windows ranging from 1 to 5 seconds, Baikulov [1] argues for a tighter window of 1 second, better suited to densely occurring ball actions. This tighter temporal constraint aligns well with the emphasis of our work on frame-accurate detection.

Table 5 presents the performance of each Baikulov's [1] fold separately, along with the average detection metrics, which we treat as the final ensemble output. Both of our methods outperform Baikulov's ensemble on most evaluation axes, but with different trade-offs. The original (non-uncertainty) system achieves the highest recall and mAP@1, indicating stronger coverage and temporal alignment with ground truth. By contrast, the uncertainty-filtered variant, where stage-1 epistemic uncertainty is used to remove low-confidence detections, yields higher precision and F1 than the original model, effectively reducing false positives through confidence-based filtering. It is worth noting that Baikulov's ensemble still reports higher precision than our original, non-filtered system, although it is not the top performer once the uncertainty filter is applied. This pattern highlights a practical precision–recall trade-off: the uncertainty filter improves reliability and reduces spurious detections (useful for analysts who prefer fewer false alarms), whereas the non-filtered pipeline retains higher recall and mean average precision (preferable when exhaustive retrieval is required).

5 Conclusions

We present a modular and interpretable pipeline for detecting and assigning passes and receptions, called touch events, in professional soccer matches from tactical camera video. Our method decomposes the task into three stages: per-player video tube analysis to estimate possession likelihood, temporal modeling to determine the possessor at each frame, and rule-based detection of touch events from possession transitions. This design enables fine-grained temporal resolution, robust handling of occlusions and missing data, and explicit identification of the player responsible for each event. Unlike many existing approaches that either focus on sparse events or lack player-level attribution, our model delivers accurate, dense touch events detection grounded in visual evidence. Evaluation on a multi-match test set with thousands of frame-level annotations demonstrates that it outperforms prior state-of-the-art methods.

Beyond performance, we have extended the model with an analysis and practical use of epistemic uncertainty. Through a deep ensemble and an uncertainty-aware classification head, our system learns to quantify confidence in its possession predictions. More importantly, we show how this uncertainty can be leveraged downstream: by replacing the fixed temporal filtering in the final event detection stage with an uncertainty-based criterion, the system dynamically suppresses low-confidence detections and reduces false positives. This not only enhances precision but also increases the model's interpretability and trustworthiness in real-world analytical workflows.

Acknowledgments. This work has been partially supported by the project GRAVATAR PID2023-151184OB-I00 funded by MCIU/AEI/10.13039/501100011033 and by ERDF, UE and by the Government of Catalonia under 2020 DI 00106.

References

1. Baikulov, R.: Winning solution for soccernet ball action spotting challenge 2023. https://github.com/lRomul/ball-action-spotting (2023)
2. Berry, L., Meger, D.: Efficient epistemic uncertainty estimation in regression ensemble models using pairwise-distance estimators. arXiv preprint arXiv:2308.13498 (2023)
3. Biermann, H., Komitova, R., Raabe, D., Müller-Budack, E., Ewerth, R., Memmert, D.: Synchronization of passes in event and spatiotemporal soccer data. Scientific Reports **13**(1), 15878 (2023)
4. Bredin, H.: Tristounet: triplet loss for speaker turn embedding. In: ICASSP. pp. 5430–5434 (2017)
5. Capellera, G., Ferraz, L., Rubio, A., Agudo, A., Moreno-Noguer, F.: Footbots: A transformer-based architecture for motion prediction in soccer. In: International Conference on Image Processing. pp. 2313–2319 (2024)
6. Capellera, G., Ferraz, L., Rubio, A., Agudo, A., Moreno-Noguer, F.: Transportmer: A holistic approach to trajectory understanding in multi-agent sports. In: Asian Conference on Computer Vision (2024)
7. Capellera, G., Rubio, A., Ferraz, L., Agudo, A.: Unified uncertainty-aware diffusion for multi-agent trajectory modeling. In: Conference on Computer Vision and Pattern Recognition. pp. 22476–22486 (2025)
8. Cossich, V.R., Carlgren, D., Holash, R.J., Katz, L.: Technological breakthroughs in sport: Current practice and future potential of artificial intelligence, virtual reality, augmented reality, and modern data visualization in performance analysis. Applied Sciences **13**(23), 12965 (2023)
9. Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G., Wang, L.: Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In: International Conference on Computer Vision. pp. 9921–9931 (2023)
10. Deliege, A., Cioppa, A., Giancola, S., Seikavandi, M.J., Dueholm, J.V., Nasrollahi, K., Ghanem, B., Moeslund, T.B., Van Droogenbroeck, M.: Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In: Conference on Computer Vision and Pattern Recognition Workshops. pp. 4508–4519 (2021)
11. Deliège, A., Cioppa, A., Giancola, S., Seikavandi, M.J., Dueholm, J.V., Nasrollahi, K., Ghanem, B., Moeslund, T.B., Droogenbroeck, M.V.: Ball action data and labels for soccernet ball action spotting challenge. https://www.soccer-net.org/datah.ykgf675j127d (2023)
12. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
13. Denize, J., Liashuha, M., Rabarisoa, J., Orcesi, A., Hérault, R.: Comedian: Self-supervised learning and knowledge distillation for action spotting using transformers. In: Winter Conference on Applications of Computer Vision. pp. 530–540 (2024)
14. Dick, U., Link, D., Brefeld, U.: Who can receive the pass? a computational model for quantifying availability in soccer. Data mining and knowledge discovery **36**(3), 987–1014 (2022)
15. Gal, Y., Ghahramani, Z.: Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In: international conference on machine learning. pp. 1050–1059. PMLR (2016)

16. Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: Soccernet: A scalable dataset for action spotting in soccer videos. In: Conference on Computer Vision and Pattern Recognition Workshops. pp. 1711–1721 (2018)
17. Giancola, S., Ghanem, B.: Temporally-aware feature pooling for action spotting in soccer broadcasts. In: Conference on Computer Vision and Pattern Recognition. pp. 4490–4499 (2021)
18. Goes, F., Meerhoff, L., Bueno, M., Rodrigues, D., Moura, F., Brink, M., Elferink-Gemser, M., Knobbe, A., Cunha, S., Torres, R., et al.: Unlocking the potential of big data to support tactical performance analysis in professional soccer: A systematic review. European Journal of Sport Science **21**(4), 481–496 (2021)
19. Gutiérrez-Pérez, M., Agudo, A.: No bells just whistles: Sports field registration by leveraging geometric properties. In: Conference on Computer Vision and Pattern Recognition Workshops (2024)
20. Gutiérrez-Pérez, M., Agudo, A.: Pnlcalib: Sports field registration via points and lines optimization. SSRN (2024)
21. Gutiérrez-Pérez, M., Agudo, A.: Soccernet-v3D: Leveraging sports broadcast replays for 3D scene understanding. In: Conference on Computer Vision and Pattern Recognition Workshops (2025)
22. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Conference on Computer Vision and Pattern Recognition. pp. 770–778 (2016)
23. Honda, Y., Kawakami, R., Yoshihashi, R., Kato, K., Naemura, T.: Pass receiver prediction in soccer using video and players' trajectories. In: Conference on Computer Vision and Pattern Recognition. pp. 3503–3512 (2022)
24. Huang, Z., Lam, H., Zhang, H.: Quantifying epistemic uncertainty in deep learning. arXiv preprint arXiv:2110.12122 (2021)
25. Kim, H., Choi, H.J., Kim, C.J., Yoon, J., Ko, S.K.: Ball trajectory inference from multi-agent sports contexts using set transformer and hierarchical bi-lstm. In: ACM SIGKDD. pp. 4296–4307 (2023)
26. Kirsch, A.: (implicit) ensembles of ensembles: Epistemic uncertainty collapse in large models. arXiv preprint arXiv:2409.02628 (2024)
27. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in neural information processing systems **30** (2017)
28. Liga, D.F.: Positional tracking takes a big leap forward as latest generation is installed at bundesliga and bundesliga 2 stadiums. https://www.dfl.de/en/innovation/positional-tracking-takes-a-big-leap-forward-as-latest-generation-is-installed-at-bundesliga-and-bundesliga-2-stadiums/ (2020)
29. Lin, J., Gan, C., Han, S.: Tsm: Temporal shift module for efficient video understanding. In: International Conference on Computer Vision. pp. 7083–7093 (2019)
30. Luo, Y., Mesgarani, N.: Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. T AUDIO SPE **27**(8), 1256–1266 (2019)
31. Majeed, F., Nazir, M., Swart, K., Agus, M., Schneider, J.: Real-time analysis of soccer ball–player interactions using graph convolutional networks for enhanced game insights. Scientific Reports **15**(1), 21859 (2025)
32. Morales, G., Sheppard, J.W.: Adaptive sampling to reduce epistemic uncertainty using prediction interval-generation neural networks. In: Conference on Artificial Intelligence. vol. 39, pp. 19546–19553 (2025)
33. Ochin, J., Devineau, G., Stanciulescu, B., Manitsaris, S.: Game state and spatio-temporal action detection in soccer using graph neural networks and 3d convolutional networks. International Conference on Pattern Recognition Applications and Methods (2025)

34. Peral, M., Capellera, G., Rubio, A., Ferraz, L., Moreno-Noguer, F., Agudo, A.: Temporally accurate events detection through ball possessor recognition in soccer. International Conference on Computer Vision Theory and Applications **221**, 231 (2025)
35. Philipp Singer, Yauhen Babakhin, P.P.: Winning solution for bundesliga data shootout. https://www.kaggle.com/competitions/dfl-bundesliga-data-shootout/discussion/359932 (2022)
36. Rahimian, P., Kim, H., Schmid, M., Toka, L.: Pass receiver and outcome prediction in soccer using temporal graph networks. In: International Workshop on Machine Learning and Data Mining for Sports Analytics. pp. 52–63. Springer (2023)
37. Sanford, R., Gorji, S., Hafemann, L.G., Pourbabaee, B., Javan, M.: Group activity detection from trajectory and video data in soccer. In: Conference on Computer Vision and Pattern Recognition Workshops. pp. 898–899 (2020)
38. Sensoy, M., Kaplan, L., Kandemir, M.: Evidential deep learning to quantify classification uncertainty. Advances in neural information processing systems **31** (2018)
39. Simpson, I., Beal, R.J., Locke, D., Norman, T.J.: Seq2event: Learning the language of soccer using transformer-based match event prediction. In: ACM SIGKDD. pp. 3898–3908 (2022)
40. Sorano, D., Carrara, F., Cintia, P., Falchi, F., Pappalardo, L.: Automatic pass annotation from soccer video streams based on object detection and lstm. In: European Conference on Machine Learning. pp. 475–490 (2021)
41. Spearman, W., Basye, A., Dick, G., Hotovy, R., Pop, P.: Physics-based modeling of pass probabilities in soccer. In: MIT Sloan Sports Analytics Conference. vol. 1. Boston, MA (2017)
42. Sudhakaran, S., Escalera, S., Lanz, O.: Gate-shift networks for video action recognition. In: Conference on Computer Vision and Pattern Recognition. pp. 1102–1111 (2020)
43. Vidal-Codina, F., Evans, N., El Fakir, B., Billingham, J.: Automatic event detection in football using tracking data. Sports Engineering **25**(1), 18 (2022)
44. Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., Lang, K.J.: Phoneme recognition using time-delay neural networks. In: Backpropagation, pp. 35–61. Psychology Press (2013)
45. Xie, X., Wang, J., Liang, H., Deng, D., Cheng, S., Zhang, H., Chen, W., Wu, Y.: Passvizor: Toward better understanding of the dynamics of soccer passes. IEEE Transactions on Visualization and Computer Graphics **27**(2), 1322–1331 (2020)
46. Yamamoto, D.: Third place solution for bundesliga data shootout. https://www.kaggle.com/competitions/dfl-bundesliga-data-shootout/discussion/360236 (2022)
47. Yamamoto, K., Uezu, S., Kagawa, K., Yamazaki, Y., Narizuka, T.: Theory and data analysis of player and team ball possession time in football. Physical Review E **109**(1), 014305 (2024)
48. Yeung, C.C., Sit, T., Fujii, K.: Transformer-based neural marked spatio temporal point process model for football match events analysis. arXiv preprint arXiv:2302.09276 (2023)
49. Yu, G., Yuan, J.: Fast action proposals for human action detection and search. In: Conference on Computer Vision and Pattern Recognition. pp. 1302–1311 (2015)
50. Zhou, X., Kang, L., Cheng, Z., He, B., Xin, J.: Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection. arXiv preprint arXiv:2106.14447 (2021)