AI for Sport

Event-Recognition and Highlight Clipping

Project for
KAUST x Stats Perform

Submitted by:

Bader Alshamrani
Abdulelah Khalaf
Heba Omer
Omar Hejazi
Khadijah Baothman

August 25, 2025

Abstract

This project introduces a comprehensive system designed to significantly enhance the sports viewing experience by automating the identification and extraction of pivotal moments from extensive, untrimmed broadcast soccer videos. Leveraging the SoccerNet dataset, our methodology focuses on detecting six key event types: goals, penalties, fouls, shots on target, red and yellow cards, and direct free-kicks.

The core of our action spotting pipeline employs the Context-Aware Loss Function (CALF) model, as proposed by Cioppa et al. (CVPR 2020). This advanced deep learning approach utilizes a pre-trained convolutional neural network for feature extraction, followed by a 1D temporal convolutional neural network for robust temporal modeling, effectively addressing the challenges of localizing sparse events in long video sequences. CALF's innovative loss function integrates temporal context, improving the accuracy and robustness of event detection.

Upon detection, the system automatically retrieves and extracts short highlight clips for each identified event, providing viewers with instant access to crucial game moments. A novel component of our system is the integration of dynamic, tactical mini-map visualizations alongside these clips. These mini-maps offer real-time positional tracking of players and the ball, providing a clear, simplified tactical overview of how each play unfolded on the pitch.

This dual-modality approach, combining precise highlight extraction with rich, visualized tactical insights, aims to deliver a more immersive, analytical, and insightful way for viewers to engage with and understand sports content.

1 Introduction

The landscape of sports consumption has been profoundly reshaped by the proliferation of broadcast media, offering fans unprecedented access to live and archived matches. However, the sheer volume of content in full-length untrimmed videos often makes it challenging for viewers to quickly identify and re-engage with pivotal moments. While traditional highlight reels exist, they frequently lack the contextual and tactical depth that can truly enhance a viewer's understanding of the game. Automating the precise identification of significant in-game events and presenting them in an enriched format remains a key challenge and a significant opportunity for innovation in sports analytics and media.

This project addresses these challenges by developing a comprehensive system for automated event recognition and dynamic highlight clipping within broadcast soccer videos. Our primary objective is to move beyond simple clip extraction to provide a more insightful and immersive viewing experience. We focus on automatically detecting six critical event types: goals, penalties, fouls, shots on target, red and yellow cards, and direct free-kicks, which are widely recognized as game-changing moments.

Our technical approach is founded on state-of-the-art deep learning methodologies, specifically employing the Context-Aware Loss Function (CALF) model from Cioppa et al. (CVPR 2020) for robust temporal action spotting. This model is adept at pinpointing the exact occurrences of sparse events within long video sequences, overcoming common issues associated with traditional event localization techniques.

Beyond merely extracting video segments, a core innovation of this work lies in the integration of dynamic, tactical mini-map visualizations. For each identified highlight clip, a corresponding mini-map is generated to illustrate player and ball positions, offering a clear spatial and strategic context of the event as it unfolded. This dual-modality presentation—combining the raw visual thrill of the highlight with an analytical tactical overview—aims to provide a richer understanding for fans, analysts, and coaches alike.

The remainder of this report is structured as follows: Section 2 details the data collection process using the SoccerNet dataset. Section 3 elaborates on the Action Spotting methodology, providing an in-depth look at the CALF pipeline. Section 4 describes the event selection and video clip extraction processes. Section 5 introduces the novel mini-map generation technique, followed by Section 6, which outlines the creation of the final highlight reel. Finally, Section 7 discusses our evaluation strategies.

2 Related Work

The automatic analysis of sports videos is a well-established field in computer vision, with applications spanning highlight generation, tactical analysis, and enhanced viewer engagement. Our project builds upon existing research in temporal action spotting, video summarization, and sports data visualization.

Temporal action spotting, which involves pinpointing specific events within untrimmed video, has evolved from methods using hand-crafted features to sophisticated deep learning models like Convolutional Neural Networks (CNNs) and Temporal Convolutional Networks (TCNs). A persistent challenge in sports, especially with sparse events like goals, is the precise localization due to their infrequent occurrence and the inherent variability in human annotations. The Context-Aware Loss Function (CALF) model, central to our work, represents a significant advancement by integrating temporal context into its learning process. This approach enhances robustness against annotation uncertainty and improves the accurate spotting of rare events, making it highly effective for our task.

In parallel, sports video summarization aims to condense lengthy broadcasts into engaging highlight reels. Early methods often relied on heuristic cues such as crowd noise or camera motion. More modern approaches employ learning-based techniques to identify exciting segments. Our project refines this by focusing on event-centric highlights, ensuring that the extracted clips are directly tied to objectively defined, game-altering moments rather than general excitement scores.

Finally, tactical analysis in sports traditionally uses specialized tracking data to visualize player and ball movements on a pitch. While deriving such detailed positional information from broadcast video presents its own set of challenges, our innovation lies in integrating dynamic mini-map visualizations directly with automatically extracted highlight clips. This provides viewers with an immediate, complementary tactical context for the visual event, thus bridging the gap between raw video footage and insightful game analysis.

3 Methodology

This section meticulously details the entire technical pipeline implemented for our project, from initial data handling to the final generation of enriched highlight packages. We describe the sequential stages involved in automatically detecting soccer events, extracting relevant video segments, and augmenting them with tactical visualizations.

3.1 Data Collection

Our project initiates with the acquisition of comprehensive match data. For this, we utilized the SoccerNet dataset, a rich repository providing full-length broadcast soccer videos. Each video within this collection is crucially accompanied by precise, time-stamped ground-truth annotations. These annotations accurately mark the occurrence of various key in-game events, which serve as the targets for our detection system. Specifically, the events we focused on from the SoccerNet annotations include:

- Goals
- Penalties
- Fouls
- Shots on Target
- Red and Yellow Cards
- Direct Free-Kicks

This meticulously labeled dataset forms the indispensable foundation for training and validating our event spotting models.

3.2 Action Spotting with Context-Aware Loss Function (CALF)

The precise and automated identification of specific events within the continuous flow of untrimmed broadcast sports videos represents a formidable challenge in computer vision. To address this, our project integrates the Context-Aware Loss Function (CALF) model, a highly effective deep learning architecture first presented by Cioppa et al. at CVPR 2020. CALF is particularly well-suited for the

task of temporal action spotting in soccer, as it adeptly navigates the complexities posed by the sparse and often subjectively annotated nature of events across long video durations. It moves beyond conventional methods by explicitly incorporating temporal context into its learning paradigm, leading to more robust and accurate event localization. The detailed technical pipeline for this action spotting mechanism is described below.

3.2.1 CALF Pipeline Overview

The CALF event spotting system processes raw broadcast video data through a sophisticated, multi-stage workflow, systematically transforming visual information into precise event detections:

1. **Visual Feature Extraction:** The initial stage involves converting the raw video frames into a rich, compact representation. This is achieved by feeding the input broadcast video into a pre-trained ResNet-152 convolutional neural network (CNN). ResNet-152, known for its deep architecture and strong representational capacity, has been extensively pre-trained on the vast ImageNet dataset. This pre-training endows the network with a powerful ability to discern generic visual patterns, which are highly transferable to understanding sports imagery. To manage the computational demands of processing long videos and to capture a balanced view of temporal dynamics, frames are sampled at a rate of 2 frames per second (fps). The high-level features extracted for each individual frame are then aggregated through a temporal pooling mechanism into contiguous 2-second chunks. This pooling operation serves to create a fixed-length feature vector for short temporal windows, effectively summarizing the visual information over these intervals and providing a structured sequence of feature vectors that encapsulate the evolving visual timeline of the match.

2. **Temporal Pattern Learning:** Following feature extraction, these sequences of 2-second feature vectors are input into a 1D Temporal Convolutional Neural Network (1D CNN). Unlike 2D CNNs which operate on spatial dimensions, or 3D CNNs which process both spatial and temporal dimensions simultaneously, a 1D CNN excels at identifying patterns exclusively along the temporal axis. This specialized network is tasked with learning intricate, localized temporal patterns that are highly characteristic of specific soccer events. For instance, it can learn to recognize the subtle sequence of player movements and ball trajectories that consistently precede a shot on target, or the distinct gestures and immediate reactions associated with a referee issuing a card. This module's ability to capture short-range temporal dependencies is crucial for distinguishing between background activity and genuine event occurrences.

3. **Event Classification Layer:** At each discrete temporal step within the processed match timeline, the model's final output layer, typically a softmax or sigmoid function, generates a probability distribution. This distribution quantifies the likelihood of the presence of each predefined event class (e.g., goal, penalty, foul, red card, yellow card, direct free-kick) at that precise moment. Crucially, an additional "background" class is included. This background class is vital for handling the vast majority of frames where no specific event of interest is occurring, allowing the model to discriminate effectively between event-bearing and non-event segments. The output probabilities for each class across the entire timeline form the raw predictions that will undergo further refinement.

4. **Context-Aware Loss Function (CALF):** The fundamental innovation and distinguishing feature of this pipeline is its proprietary Context-Aware Loss Function. Traditional loss functions, such as cross-entropy, impose a strict penalty only for incorrect predictions precisely at the single ground-truth timestamp. However, in real-world sports annotations, human labelers may slightly disagree on the exact second an event begins or ends, leading to inherent "annotation uncertainty." CALF addresses this by employing a sophisticated soft-labeling scheme. Instead of a binary "hit or miss" evaluation, CALF assigns varying weights to predictions based on their temporal proximity to the ground-truth event. Predictions occurring very close to the true event timestamp receive a higher weight, meaning any error there is more heavily penalized, while predictions further away still contribute to the loss but with diminished importance. This design actively encourages the model to learn not just the isolated event timestamp, but also the broader temporal context and the crucial preparatory or follow-through actions that surround it (e.g., the passing sequence before a goal, or the immediate aftermath of a foul). This contextual learning capacity makes the model significantly more robust to minor annotation discrepancies and improves its ability to generalize to nuanced event timings.

5. **Inference and Post-processing for Refinement:** During the inference phase, the trained model generates a continuous stream of probability scores for all event classes across the entire duration of the match video. To convert these raw, continuous predictions into discrete, actionable event detections, a series of post-processing steps are applied. Firstly, Non-Maximum Suppression (NMS) is utilized. NMS is a critical technique that addresses the common scenario where the model might output multiple, closely-spaced high-probability detections for what is, in reality, a single event. It works by iteratively selecting the detection with the highest confidence score and then suppressing (removing) all other overlapping detections within a defined temporal window. This ensures that each distinct event is represented by only one optimal prediction. Following NMS, a predefined confidence threshold is applied. This threshold filters out any remaining detections whose probability scores fall below a certain level, effectively pruning low-confidence, potentially spurious detections and thereby enhancing the precision of the final output. The culmination of this intricate process is a structured JSON file, which serves as the final output. This file precisely lists each detected event, its timestamp in milliseconds, the predicted event label (e.g., "goal," "foul"), and the associated confidence score, making the detections readily consumable for subsequent stages of analysis and highlight generation.

3.2.2 Key Advantages of CALF for Sports Event Spotting

The decision to implement the Context-Aware Loss Function (CALF) model is underpinned by its inherent advantages in overcoming several critical challenges pervasive in sports video analysis:

- **Enhanced Robustness to Annotation Uncertainty:** One of the most significant benefits of CALF is its ability to gracefully handle the inherent imprecision often found in human-labeled ground-truth data. By employing a soft-labeling scheme that weighs predictions based on their proximity to the actual event, the model becomes less sensitive to slight variations in annotation timestamps, which can occur due to human perception or the ambiguous nature of event boundaries. This significantly reduces the impact of minor annotation errors on the training process, leading to a more resilient and accurate model.
- **Superior Tolerance for Sparse Event Distributions:** Sports events like goals or penalties are, by their nature, infrequent within an hour-long match. This creates a highly imbalanced dataset where "background" frames vastly outnumber "event" frames. Traditional cross-entropy loss functions can struggle in such scenarios, often leading to models biased towards the majority class (background). CALF's contextual weighting helps to amplify the importance of the rare event moments and their surrounding temporal context, allowing the model to learn more effectively from these sparse occurrences and significantly improving its ability to detect them.
- **Achieving State-of-the-Art Performance:** The effectiveness of CALF is not merely theoretical; it has been empirically validated. The model has demonstrated a leading performance on the SoccerNet-v2 Action Spotting benchmark, which is a rigorous standard for this task. Its achievements, consistently measured by strong average-mAP (mean Average Precision) scores across various temporal tolerances, underscore its capability to accurately and reliably identify events under real-world conditions.
- **Effective Reduction of False Positives:** By explicitly learning the temporal context surrounding events, CALF develops a more nuanced understanding of what constitutes a true event. This contextual awareness helps the model to better discriminate between genuine event occurrences and visually similar, but non-event, activities. Consequently, this leads to a notable reduction in false positive detections, ensuring that the generated highlights are more accurate and relevant, thus improving the overall quality of the automated output.

4 Event Selection

Upon the successful completion of the temporal action spotting phase, the CALF model produces a comprehensive output detailing numerous detected events, each with an associated timestamp and confidence score, spanning the entirety of the broadcast soccer video. However, for the ultimate goal of generating engaging and impactful highlight reels, a direct, unfiltered utilization of this raw output is often suboptimal. Not every detected incident, regardless of its technical classification, holds equal significance or

appeal for a typical spectator. Moreover, while highly accurate, deep learning models can occasionally yield detections with lower confidence, which may not represent genuine events. Consequently, a crucial Event Selection stage is implemented to meticulously refine these CALF predictions, prioritizing both semantic importance for the audience and the inherent reliability of each detection.

4.0.1 Curating for Spectator Relevance and Semantic Weight

The initial and paramount criterion for event selection revolves around the intrinsic value and spectator appeal of the identified actions. Our objective is to construct highlight packages that resonate deeply with viewers, offering a concise yet compelling narrative of a match's most critical junctures. To achieve this, we selectively filter the CALF output to include only those event categories universally acknowledged as pivotal, emotionally charged, or strategically defining moments within a soccer game. These are the instances that profoundly influence the match's progression, evoke strong reactions, and are frequently replayed and discussed. The specific event types chosen for this curated highlight generation are:

- **Goals:** As the definitive scoring moments, goals represent the zenith of offensive play and are the most impactful events, directly altering the match score and atmosphere. Their inclusion is non-negotiable for any meaningful highlight package.
- **Penalties:** These are moments of extreme tension and high consequence, often resulting from critical fouls within the penalty area. A penalty kick can dramatically swing the momentum of a game or secure a vital score, making them intrinsically dramatic.
- **Fouls:** While not always leading directly to a goal, significant fouls can disrupt play, lead to free-kicks in dangerous positions, or result in disciplinary action. Their inclusion provides context to stoppages and tactical shifts.
- **Shots on Target:** These events signify a direct attempt at scoring, showcasing offensive prowess and often serving as the culmination of intricate attacking sequences. They capture moments of near-misses and goalkeeping heroics.
- **Red and Yellow Cards:** These disciplinary actions carry substantial weight, altering team dynamics by potentially reducing player numbers (red card) or influencing a player's future conduct (yellow card). They highlight flashpoints and crucial referee decisions.
- **Direct Free-Kicks:** Awarded for certain fouls, these present direct scoring opportunities or dangerous set-piece situations near the opponent's goal, often leading to intense attacking plays.

This stringent focus on semantically rich events ensures that the resulting highlight reel is not merely a collection of detected actions but a compelling summary of the game's defining narrative points, directly aligning with what a human spectator would deem exciting and essential to re-watch.

5 Video Clip Extraction

Following the rigorous event selection process, which yields a curated list of high-confidence, semantically important events, the subsequent crucial step in our methodology is the Video Clip Extraction. This stage is responsible for programmatically segmenting the full-length broadcast match video into individual, concise highlight clips, each centered around a detected event. The primary objective is to create standalone video assets that encapsulate the complete narrative of each pivotal moment, providing immediate visual access to the action.

The process for extracting these highlight clips is meticulously executed through a series of steps:

1. **Timestamp Retrieval from CALF Output:** For each event that successfully passed the confidence and semantic filtering criteria, its precise temporal location is retrieved. This critical piece of information, the event timestamp, is extracted directly from the JSON file generated during the inference and post-processing stage of the CALF model. These timestamps are typically recorded in milliseconds, ensuring a high degree of temporal accuracy for the subsequent cutting operation. This numerical precision is fundamental to isolating the exact moment of an event.

2. **Application of Contextual Time Window:** To ensure that each highlight clip is not just an isolated moment but a coherent and comprehensive segment of action, a fixed time window is applied around the detected event timestamp. Specifically, a window of ± 5 seconds is utilized. This means that the extracted clip begins 5 seconds before the detected event and concludes 5 seconds after it. The rationale behind this symmetric 10-second window (event at the 5-second mark) is multifaceted:

   - **Capturing Build-up and Aftermath:** The pre-event segment (e.g., the 5 seconds leading up to a goal) often contains vital contextual information, such as the strategic passes, dribbles, or tactical movements that directly contribute to the event. Similarly, the post-event segment captures immediate reactions, celebrations, or subsequent play that adds to the narrative completeness.
   - **Providing Narrative Flow:** A short, isolated 'snap' of an event can be jarring. A slightly longer window allows for a more natural viewing experience, enabling the spectator to understand the sequence of actions that led to and immediately followed the highlight.
   - **Standardization:** A consistent window length across all highlights provides uniformity, simplifying subsequent processing and presentation.

   This contextual framing is crucial for generating clips that are not only accurate in their event localization but also rich in narrative detail.

3. **Programmatic Video Segmentation:** With the start and end times for each highlight clip precisely defined, the actual video cutting is performed programmatically. For this task, a robust video processing library such as MoviePy is employed. MoviePy (or similar tools) provides the necessary functionalities to precisely trim video segments from the larger, untrimmed source file based on the calculated start and end timestamps. Each identified 10-second segment is then saved as a distinct, individual video file (e.g., MP4 format). These files are typically named and organized in a way that correlates them with their respective event types and original match context.

The culmination of this Video Clip Extraction process is the creation of a comprehensive set of short, self-contained highlight videos. Each video corresponds directly to a key match event, enriched with a surrounding temporal context, and is prepared for further augmentation and assembly into the final highlight reel. This systematic approach ensures that the output is both precise and readily usable for enhancing the spectator's viewing experience.

6 Mini-Map Generation for Highlights

While the extracted video clips offer a direct, high-fidelity visual replay of key events, they frequently lack the crucial tactical context that underpins a deeper understanding of the play. For spectators, analysts, and coaches alike, grasping the spatial dynamics—such as player positioning, ball movement, and the precise location of an event on the field—is essential for truly comprehending its significance. To bridge this analytical gap, a pivotal and innovative component of our methodology is the Mini-Map Generation for Highlights. This stage is built upon the SoccerNet Game State Reconstruction (GSR) framework (`sn-gamestate`) and is designed to enrich interpretability by providing dynamic, visual representations of the tactical landscape that accompany the raw video footage, transforming a simple highlight into a comprehensive analytical package.

The systematic generation of these mini-maps involves the following integrated steps:

1. **Establishment of a Standardized Pitch Model:** The foundation for all tactical visualizations is a consistent and geometrically accurate digital representation of the soccer field. We employ a standardized 2D pitch model, precisely adhering to official regulatory dimensions (105 meters in length by 68 meters in width). This virtual pitch serves as a canonical and undistorted canvas onto which all event locations and dynamic player positions are projected. The adoption of such a standardized model is crucial for ensuring uniformity across all highlights and mitigating potential distortions that could arise from varying broadcast camera angles or stadium-specific layouts, thereby providing an objective tactical overview.

2. **Event Mapping and Positional Data Integration:** Accurately transposing the detected events and player movements from video frames onto the 2D pitch model is a complex but essential process.

- **Event Location Mapping:** The exact coordinates corresponding to the detected events (e.g., the precise spot of a foul, the origin point of a shot) are mapped onto our standardized pitch model. When available, this mapping benefits from leveraging explicit in-game tracking data, such as that provided by optical tracking systems or GPS sensors, which offer highly precise real-world coordinates.
- **Approximation from Broadcast Video:** In cases where direct in-game coordinate data is absent—a common scenario for broadcast video feeds—we utilize computer vision techniques to approximate these spatial positions. This involves homography estimation, which transforms pixel coordinates from the video frame into corresponding coordinates on the 2D pitch plane, combined with broadcast camera tracking algorithms. These techniques reconstruct the approximate locations of the ball and key players on the field, providing a robust tactical layout.
- **Temporal Synchronization:** A critical aspect of this integration is the precise temporal synchronization of the mapped positional data points with the corresponding video frames of the highlight clip. This ensures that the mini-map dynamically updates, accurately reflecting player and ball movements in real-time as the associated video segment unfolds.

3. **Dynamic Visualization of Tactical Action:** For each extracted highlight video clip, a corresponding dynamic mini-map is generated. This visualization is not static but evolves in concert with the video, intuitively illustrating the unfolding action. The mini-map can render various tactical elements, customized according to the nature of the event:

   - **Player and Ball Positions:** Small, distinct icons or markers represent players from both teams, showing their real-time positions and trajectories on the pitch before, during, and after the event. The ball's position and movement are also tracked.
   - **Event-Specific Indicators:** For "Shots on Target" or "Direct Free-Kicks," the mini-map graphically depicts the ball's trajectory, offering valuable insight into its path, angle, and distance. For "Fouls," "Penalties," or "Cards," the precise location of the incident is highlighted on the pitch.
   - **Movement Patterns:** Over the duration of the clip, the mini-map highlights offensive runs, defensive formations, passing triangles, or pressing actions, thereby enriching the viewer's tactical understanding of the play's development.

   The visual design prioritizes clarity and minimal clutter, aiming to convey maximum tactical information in an easily digestible format.

4. **Integrated Hybrid Presentation:** The final step is the seamless integration of the dynamically generated mini-map with its corresponding highlight video. Rather than presenting the mini-map as a separate graphic, it is strategically overlaid alongside the main video clip. This dual-display approach, which can manifest as a Picture-in-Picture (PiP) window or a side-by-side split-screen format, provides viewers with an enriched experience. It simultaneously offers the direct visual replay of the event and an immediate tactical understanding of its spatial context. This enables viewers to:

   - **Enhance Situational Awareness:** Understand not just *what* critical action transpired, but also *where* it unfolded on the pitch relative to teammates, opponents, and landmarks.
   - **Gain Deeper Tactical Insight:** Observe player formations, defensive lines, offensive runs, and ball circulation patterns that may be difficult to fully appreciate in a single broadcast camera view.
   - **Foster Increased Engagement:** Provide a more comprehensive viewing experience that appeals to both casual fans and dedicated analysts.

   By combining these two modalities—raw video and tactical visualization—the mini-map integration transforms a standard highlight into a rich, informative, and analytically valuable package, significantly elevating the interpretability of sports content.

Detector Fine-Tuning

To support this mini-map generation pipeline, we employed a YOLOv12 detector, a more modern and efficient architecture compared to the baseline used in the original `sn-gamestate` framework. We fine-tuned the model on a SoccerNet-derived dataset available through Roboflow Universe. The dataset originally provides seven categories (players, goalkeepers, referees, and ball), which we harmonized into a simpler schema: *athlete* (all players and goalkeepers), *referee* (main and side referees), and optionally *ball*. This remapping avoids brittle orientation-based labels (e.g., team left/right) and ensures compatibility with the downstream modules of the GSR pipeline. By integrating the fine-tuned YOLOv12 detector, our system achieves more robust and accurate detection of athletes and referees in broadcast footage, thereby improving the reliability and fidelity of the tactical mini-map visualizations.

7 Highlight Reel Creation

The culmination of our entire methodology is the Highlight Reel Creation stage, where all previously processed components are synergistically integrated to form a final, enriched output. This is the ultimate step designed to deliver a comprehensive and insightful sports viewing experience. The objective is to assemble the precisely cut video clips of key events with their corresponding tactical mini-map visualizations into a coherent and engaging final product, suitable for both retrospective analysis and enhanced entertainment.

The assembly process is twofold, focusing on bringing together the visual dynamics of the event with its spatial context:

1. **Integration of Filtered Highlight Clips:** The first primary component integrated are the short, contextually rich video segments. These are the filtered highlight clips, such as those depicting goals, penalties, fouls, shots on target, red/yellow cards, and direct free-kicks, which were meticulously extracted in the preceding "Video Clip Extraction" stage. Each of these clips encapsulates a pivotal moment of the match, providing the raw, compelling visual content that forms the backbone of any highlight package.

2. **Synchronization of Mini-Map Visualizations:** Concurrently, the dynamic tactical mini-map visualizations, generated in the "Mini-Map Generation for Highlights" stage, are synchronized with their respective video clips. As previously detailed, these mini-maps offer real-time representations of player and ball positions on a standardized pitch model. The integration ensures that the tactical context dynamically updates alongside the video, providing a seamless and informative parallel viewing experience.

The output of this integration can manifest in various forms, tailored to different consumption platforms and user needs:

- **Summary Video Compilation:** One primary output is a traditional summary video. In this format, all the selected event clips, each paired with its dynamic mini-map overlay, are sequentially stitched together. This creates a cohesive, single video file that viewers can watch from start to finish, providing a chronological or event-categorized recap of the game's most important moments. This is ideal for broadcast, web platforms, or mobile viewing where a linear consumption experience is preferred.

- **Interactive Interface Integration:** Alternatively, these integrated highlight packages can be embedded within a more advanced, interactive user interface. Such an interface might allow users to:

  - Navigate between highlights on demand.
  - Pause, rewind, or fast-forward specific events.
  - Potentially toggle the mini-map visualization on or off based on their preference.
  - Access additional statistics or insights related to the event or player actions.

  This approach offers greater user control and a deeper exploratory experience, suitable for dedicated analytical platforms or sophisticated digital media applications.

The fundamental benefit of this hybrid visualization approach is its ability to profoundly enhance user understanding and engagement. Users are empowered to watch the thrilling visual sequence of an event unfold while simultaneously gaining a clear, spatial, and tactical understanding of its location on the field, the movement patterns that led to it, and the immediate consequences. This dual perspective transforms passive viewing into an active learning experience, allowing for both immediate emotional impact and deeper analytical insight, ultimately delivering a more immersive and insightful way to analyze and enjoy sports content.

8 Conclusion

In this project, we developed a comprehensive system for automated event recognition and enriched highlight generation in broadcast soccer videos. By leveraging the SoccerNet dataset and integrating the Context-Aware Loss Function (CALF) model, we successfully addressed the inherent challenges of detecting sparse, temporally ambiguous events across long, untrimmed match footage. Our pipeline demonstrated how robust temporal modeling, combined with context-aware loss design, enhances the accuracy and reliability of spotting critical in-game actions such as goals, penalties, fouls, shots on target, cards, and direct free-kicks.

Beyond event detection, we advanced the viewing experience by introducing tactical mini-map visualizations, powered by a fine-tuned YOLOv12 detector within the SoccerNet Game State Reconstruction framework. This innovation provided spectators with both the emotional immediacy of the highlight and the analytical depth of spatial context, bridging the gap between raw footage and tactical insight. The integration of these dual modalities in the final highlight reel created a richer and more immersive way to consume and analyze sports content.

Looking forward, this framework offers substantial opportunities for further exploration. Expanding the set of detectable events, improving the accuracy of player and ball tracking, and enabling real-time deployment are promising directions that could enhance its practical value. Moreover, integrating interactive interfaces and advanced analytics could transform the system into a versatile tool for fans, coaches, and analysts alike. Ultimately, our work demonstrates the potential of artificial intelligence to not only automate highlight generation but also to redefine how sports are experienced, understood, and enjoyed.

References

[1] G. Cioppa, A. Deliège, S. Giancola, et al., *SoccerNet 2023 Challenges Results: Broadcast Camera Calibration, Field Segmentation, and Player Tracking*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2023.

[2] S. Giancola, M. Amrani, and B. Ghanem, *SoccerNet-v2: A Dataset and Benchmarks for Holistic Understanding of Broadcast Soccer Videos*, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021.

[3] A. Glenn Jocher et al., *YOLOv12: Real-Time Object Detection*, arXiv preprint arXiv:2404.00554, 2024.

[4] J. Li, S. Giancola, et al., *SoccerNet Game State Reconstruction: Predicting Camera Calibration and Player Coordinates from Broadcast Video*, arXiv preprint arXiv:2303.09433, 2023.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.