

Article

Automatic Estimation of Football Possession via Improved YOLOv8 Detection and DBSCAN-Based Team Classification

Rong Guo ^{1,2,3}, Yucheng Zeng ^{1,2}, Rong Deng ¹, Yawen Lei ¹, Yonglin Che ^{1,2}, Lin Yu ¹, Jianpeng Zhang ³, Xiaobin Xu ³, Zhaoxiang Ma ^{1,2}, Jiajin Zhang ^{1,2,*} and Jianke Yang ^{2,3,*}

- ¹ College of Big Data, Yunnan Agricultural University, Kunming 650201, China; 2023210516@stu.ynau.edu.cn (R.G.); 2024240731@stu.ynau.edu.cn (Y.Z.); 2003027@ynau.edu.cn (R.D.); 2020314312@stu.ynau.edu.cn (Y.L.); 2024240737@stu.ynau.edu.cn (Y.C.); 2024210530@stu.ynau.edu.cn (L.Y.); 18363797822@163.com (Z.M.)
- ² Center for Sports Intelligence Innovation and Application, Yunnan Agricultural University, Kunming 650201, China
- ³ College of Physical Education, Yunnan Agricultural University, Kunming 650201, China; zhangjianpeng@ynau.edu.cn (J.Z.); xuxiaobin@ynau.edu.cn (X.X.)
- * Correspondence: zjjcl@ynau.edu.cn (J.Z.); yangjianke@ynau.edu.cn (J.Y.)

Abstract

Recent developments in computer vision have significantly enhanced the automation and objectivity of sports analytics. This paper proposes a novel deep learning-based framework for estimating football possession directly from broadcast video, eliminating the reliance on manual annotations or event-based data that are often labor-intensive, subjective, and temporally coarse. The framework incorporates two structurally improved object detection models: YOLOv8-P2S3A for football detection and YOLOv8-HWD3A for player detection. These models demonstrate superior accuracy compared to baseline detectors, achieving 79.4% and 71.1% validation average precision, respectively, while maintaining low computational latency. Team identification is accomplished through unsupervised DBSCAN clustering on jersey color features, enabling robust and label-free team assignment across diverse match scenarios. Object trajectories are maintained via the Norfair multi-object tracking algorithm, and a temporally aware refinement module ensures accurate estimation of ball possession durations. Extensive experiments were conducted on a dataset comprising 20 full-match Video clips. The proposed system achieved a root mean square error (RMSE) of 4.87 in possession estimation, outperforming all evaluated baselines, including YOLOv10n (RMSE: 5.12) and YOLOv11 (RMSE: 5.17), with a substantial improvement over YOLOv6n (RMSE: 12.73). These results substantiate the effectiveness of the proposed framework in enhancing the precision, efficiency, and automation of football analytics, offering practical value for coaches, analysts, and sports scientists in professional settings.

Keywords: computer vision; football possession; YOLOv8; DBSCAN; automation

Academic Editors: Honggang Chen,
Chao Ren

Received: 28 January 2026

Revised: 12 February 2026

Accepted: 13 February 2026

Published: 14 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

With the advent of the internet era, artificial intelligence (AI) technologies have demonstrated vigorous development across various sectors, continuously undergoing rapid evolution and iteration. From early-stage game-theoretic algorithms and expert systems to modern intelligent algorithms centered on machine learning and deep learning, the application boundaries of AI have been significantly expanded. As a critical subfield of AI,

computer vision has seen growing integration into sports sciences in recent years. Driven by advancements in GPU computing power and the accelerated maturation of relevant algorithms, human pose estimation—an essential component of computer vision—has emerged as a prominent research focus in the field of sports science. Alongside the sustained development of international sports technology, the industry is undergoing a paradigm shift from being “experience-driven” to “technology-driven,” with scientific and technological innovation now serving as a key driver of sports science system development. Concurrently, the accelerating pace of digital transformation is refining the theoretical frameworks and technical foundations of AI. Within this context, elite sports—at the core of the broader sports ecosystem—have increasingly incorporated AI-powered solutions to meet the evolving demands of athletes, coaches, and referees. These include intelligent training systems based on human motion capture and recognition, sparring robots, tactical optimization platforms, and automated officiating and decision-support tools. These technologies have been widely deployed across training, competition, and officiating scenarios, enhancing athletic performance, mitigating injury risks, and offering crucial support for the development of a technologically empowered sports nation 1.

In recent years, with the rapid advancement of deep learning architectures, object detection and multi-object tracking technologies have become central components in automated sports video analysis. Comprehensive reviews have summarized the evolution of one-stage detection frameworks such as the YOLO series, highlighting architectural improvements from early anchor-based models to recent anchor-free and feature-enhanced versions including YOLOv10 and YOLOv11, which significantly improve real-time inference efficiency and small-object detection capability (Ali & Zhang, 2024) 2. These developments have further strengthened the applicability of YOLO-based models in dynamic and complex sports scenarios.

This study focuses on the automatic extraction of ball possession information from spatiotemporal tracking data in football matches. In a typical football analytics pipeline, the estimation of ball possession and match status constitutes a foundational step for understanding game events and their interrelationships. In the absence of such information, analytical efforts are often limited to basic physical metrics such as movement distance and velocity. However, with access to accurate possession data, the scope of analysis can be significantly expanded: not only can core match statistics be computed, but individual passing actions can be identified and evaluated, the game segmented into discrete tactical units, and the behavioral patterns of players and teams during offensive and defensive phases further analyzed.

Possession percentage—defined as the proportion of total match time during which a team controls the ball—serves as a key indicator of team dominance and is widely utilized in tactical football analysis. Research has shown that possession metrics are closely linked to several critical performance indicators, and can be used to assess differences in physical and technical execution between teams 3, as well as to support the interpretation of tactical decisions and match dynamics 4.

Moreover, analyses based on data from major international football tournaments suggest that higher possession percentages are often associated with favorable outcomes. For instance, in all matches of the 2014 FIFA World Cup, the winning teams generally recorded possession rates exceeding 50%, with Germany—the eventual champion—achieving the highest average possession 5. Similar trends have been observed in the English Premier League, where successful teams tend to sustain longer periods of ball possession 6.

At present, data acquisition in top-tier football competitions predominantly relies on professional sports analytics companies such as OPTA Sports data. This company employs two teams of experienced operators to manually annotate matches in real time, supplementing the process with computer-generated data to derive key performance metrics such as

possession percentage 7. However, this approach is labor-intensive and time-consuming, and it introduces a degree of subjectivity due to its dependence on human annotation and professional judgment. Notably, OPTA has also revised its definition of possession, proposing that successful passes better reflect a team's ability to retain control. Accordingly, possession is defined as the proportion of a team's completed passes relative to the total passes made by both teams. This pass-based approach neglects other important aspects of possession, such as individual ball retention and tactical time-wasting, thereby limiting its comprehensiveness.

Traditional methods for estimating possession typically rely on manual tracking or offline analysis of event-based data, which are constrained by high subjectivity, limited real-time capabilities, and incomplete data coverage. With the rapid advancement of computer vision technologies in sports analytics, new opportunities have emerged for the automated and fine-grained extraction of possession-related information. Deep learning-based object detection models—particularly the YOLO (You Only Look Once) family—have demonstrated remarkable performance in object recognition tasks and laid a robust technical foundation for real-time sports data analysis.

Recent studies have further demonstrated the effectiveness of integrating advanced YOLO models with multi-object tracking algorithms for football analytics. For example, Shankara et al. (2023) combined YOLOv8 with ByteTrack to perform accurate detection and tracking of players and the ball, enabling automated extraction of football-related statistics and supporting possession analysis based on persistent track identities 8. Similarly, Wang (2025) enhanced YOLOv5 by incorporating DeepSORT and attention mechanisms to improve robustness under occlusion and motion blur conditions in sports videos. These integrated detection–tracking pipelines reflect the current mainstream approach in football Video analytics 9.

Nevertheless, applying object detection models to possession estimation remains challenging due to issues such as suboptimal player identification accuracy, difficulty in inferring possession attribution, and limited computational efficiency in real-time contexts. Furthermore, the absence of reliable possession information significantly hinders subsequent data-driven analyses (Richly et al., 2017) 10. One of the primary motivations for developing automated possession extraction methods is to circumvent the inefficiencies and limitations of manual processes, while also enabling the recovery of missing possession data from historical matches—thus extending the utility of existing spatiotemporal tracking datasets.

In addition to detection challenges, maintaining consistent player identities across frames is critical for reliable possession inference. Recent advancements in sports-oriented multi-object tracking have addressed this issue. Sun et al. (2024) proposed a Global Tracklet Association (GTA) framework to enhance identity consistency in sports tracking scenarios, while Gran-Henriksen et al. (2024) introduced Deep HM-SORT, integrating improved motion modeling and deep feature representations to strengthen tracking robustness 11. Furthermore, the release of large-scale sports multi-object tracking datasets such as Scott (2024) has facilitated benchmarking and evaluation of full-pitch tracking systems in football, promoting methodological improvements in this domain 13.

Despite the critical importance of automated possession estimation in football analytics, research on this topic remains relatively scarce in the scientific literature, particularly approaches that utilize spatiotemporal tracking data. This gap is largely due to the lack of publicly available, high-quality datasets.

Existing methods for possession estimation are often based on varying definitions of “possession,” such as pass-based metrics (Glasser, 2014; Sarkar et al., 2019) 14 or rule-based heuristics manually designed to reflect expert knowledge (Bradley et al., 2013; Khaustov & Mozgovoy, 2020; Morra et al., 2020) 1617**Error! Reference source not found.** However,

these approaches exhibit notable limitations when addressing the complexity of dynamic in-game scenarios or approximating human-level interpretative logic.

To address the aforementioned limitations, this study proposes an automated possession estimation framework that integrates computer vision and deep learning techniques. The proposed method extracts spatiotemporal tracking data of players and the ball directly from the match video, eliminating the need for handcrafted rule sets based on domain expertise, and enabling accurate estimation of possession status.

Specifically, the system first employs the YOLOv8 object detection model to identify players and the ball in each Video frame. Subsequently, a team classification process is conducted using a color-based clustering algorithm. Given the visual uniformity in jersey colors among players on the same team, a density-based clustering algorithm, DBSCAN (Density-Based Spatial Clustering of Applications with Noise), is applied to segment the jersey regions and automatically assign players to their respective teams.

Once object detection and initial classification are complete, the system proceeds to the multi-object tracking stage. To ensure consistent identity assignment across frames, a modular tracking framework based on the SORT (Simple Online and Realtime Tracking) algorithm is implemented via the Norfair library. SORT combines Kalman filtering with the Hungarian algorithm, offering high efficiency and accuracy in real-time tracking tasks. Norfair further enhances SORT's capabilities, making it more suitable for the complex motion patterns characteristic of sports scenarios, thereby enabling stable multi-object tracking throughout the match.

Following tracking, the system enters the possession inference and estimation phase. A temporal decision mechanism is introduced, leveraging frame-wise time-series information. By computing the Euclidean distance between the ball and each player in every frame, the system identifies the player closest to the ball. If this minimum distance falls below a predefined threshold, the player is considered to be in possession. Possession frames are then tallied according to team affiliation, and the final possession percentage is calculated as the ratio of each team's possession frames to the total number of frames. This method accounts for both spatial proximity and temporal continuity between players and the ball, allowing for robust estimation of dynamic possession states across varying tactical styles and match rhythms.

The primary contributions of this study are as follows:

- An integrated player identification and team classification method that combines YOLOv8 object detection with color-based clustering;

- A robust and efficient multi-object tracking module built upon the Norfair framework, tailored for complex sports environments;

- A novel temporal decision mechanism for dynamic possession attribution and percentage estimation, enhancing both accuracy and real-time performance.

Through this structured pipeline, we present a comprehensive possession estimation system that offers a scalable, automated, and accurate solution for football data analytics. Comparative evaluations demonstrate that the proposed method significantly outperforms traditional rule-based approaches.

The remainder of this paper is organized as follows: Section 2 reviews related work in the domain of sports analytics; Section 3 details the proposed methodology; Section 4 presents the experimental setup and evaluation results; and Section 5 concludes the study with a discussion of future research directions.

2. Related Work

In recent years, the analysis of sports events—particularly event detection in football matches—has garnered increasing attention and has been addressed through a variety of methodological approaches. Existing literature typically categorizes these methods based

on the type of input data utilized, which generally falls into three groups: visual data, tracking data, or a fusion of both.

2.1. Visual Data

The application of computer vision in football analytics can be traced back to the early use of traditional computer vision and signal processing algorithms. During this period, researchers achieved relatively high accuracy in detecting and tracking both players and the ball 18. However, these approaches heavily relied on handcrafted feature extraction techniques, which lacked robustness in complex match environments and were particularly sensitive to variations in lighting conditions, occlusion, and motion blur.

The advent of deep learning marked a transformative shift in the field. The introduction of AlexNet 19, for instance, significantly accelerated the development of computer vision. Subsequent studies leveraged artificial neural networks (ANNs) for the detection and tracking of the ball, goalposts, and field markings, particularly in the context of robotic football competitions 20. In parallel, research began to focus on the detection of critical in-game events, such as offside decisions 21, penalty situations, and dead-ball scenarios 22, laying a solid foundation for automated match analysis. Nevertheless, investigations into ball possession and the calculation of possession time remain relatively underexplored. Achieving accurate and real-time estimation of ball possession continues to pose an open challenge.

In recent years, computer vision and deep learning have made substantial progress in the automated recognition of match data, player tracking, and tactical analysis. The YOLO (You Only Look Once) family of object detection models, known for their end-to-end structure and high-speed inference, has been widely adopted in football analytics for detecting players, balls, and match events 23. For instance, the YOLO model introduced by Redmon et al. (2016) 24 has demonstrated outstanding performance in real-time detection tasks involving both players and the ball. Hong et al. (2018) 25 employed convolutional neural networks (CNNs) to detect key events in football matches, while Xu and Tasaka (2020) 26 enhanced event detection accuracy through the integration of multi-view Videodata.

Beyond object detection, time-series analysis has also played a critical role in sports Videoanalytics. For example, Sorano et al. (2021) 27 proposed the use of convolutional recurrent neural networks (CRNNs) and temporal convolutional networks (TCNs) to extract temporal features, thereby enhancing the analysis of tactical patterns and ball possession transitions. However, the high computational demands of these methods often limit their applicability in real-time scenarios. To address this limitation, recent studies have integrated multi-object tracking (MOT) techniques to improve real-time tracking of player and ball trajectories. Combined with temporal analysis, these approaches aim to refine the determination of ball possession and enhance the accuracy and practicality of possession time estimation. In addition to large-scale object detection tasks, few-shot detection has also attracted growing attention in domain-specific applications where labeled data are scarce. Hu et al. (2021) investigated railway automatic switch stationary contact wear detection under few-shot conditions using a YOLO-based framework 28. Their study addressed the challenge of maintaining detection accuracy when only a limited number of annotated samples were available, which is common in industrial inspection scenarios. By adapting the detection architecture and optimizing the training strategy for small-sample environments, the proposed approach demonstrated that YOLO-based models can achieve stable and reliable performance even under constrained data conditions. This work further highlights the adaptability and generalization capability of deep learning-based object detection frameworks across diverse application domains.

2.2. Tracking Data

In the field of sports match analysis, multi-object tracking (MOT) is considered one of the core tasks for enabling dynamic tactical recognition and accurate ball possession evaluation. Early MOT research predominantly relied on classical techniques such as Kalman filtering and the Hungarian algorithm, as demonstrated by Berclaz et al. (2011) 29. In recent years, the integration of deep learning with traditional tracking methodologies has significantly enhanced MOT performance.

SORT combines Kalman filtering with the Hungarian algorithm to achieve efficient multi-target tracking. Its success inspired the development of more robust models, such as DeepSORT (Wojke et al., 2017) 30, which incorporates deep feature embeddings to improve target association under occlusions and appearance variations. Building upon advances in the MOT domain, Tryolabs introduced the modular tracking library Norfair 31. Norfair provides a flexible tracking framework that supports custom matching strategies and performs well in challenging conditions such as target occlusion and nonlinear movement, making it highly suitable for sports analytics applications.

Beyond detection and tracking, trajectory-based data has also become increasingly important in sports analysis. For example, Horton (2020) 32 proposed a trajectory-embedding-based learning approach to characterize player movement patterns in American football. Lucey et al. (2013) 33 introduced a role-based, rather than identity-based, framework for tactical analysis, which better reflects team structure and strategy. In football, identifying ball possession is a key analytic goal. Link and Hoernig (2017) 34 developed a rule-based approach that infers possession changes based on ball acceleration patterns. Similarly, Sanford et al. (2020) 35 conducted a comparative study of I3D and TCN + Transformer models using both visual and trajectory data, showing that incorporating trajectory features significantly improves event detection accuracy.

Building on this body of work, the present study proposes a real-time football match analysis system that integrates state-of-the-art detection, clustering, and tracking techniques to enhance the accuracy and real-time performance of ball possession estimation. Specifically, the system adopts the latest generation of the YOLO model—YOLOv8—for frame-by-frame detection of players and the ball in football match videos. YOLOv8 offers notable improvements in detection precision and inference speed, making it well-suited for real-time applications.

To differentiate between teams, the system applies the DBSCAN clustering algorithm to group players based on the distribution of jersey colors in HSV space. This unsupervised approach eliminates the need for manual labeling or identity-based team classification, thereby improving adaptability in real-world match conditions. For persistent tracking of multiple players and the ball across video frames, an enhanced version of the Norfair tracking library, based on SORT, is employed. This module combines Kalman filtering, the Hungarian matching algorithm, and customizable distance functions, ensuring robust identity consistency even under challenging conditions such as player occlusion or similar uniforms. The output trajectories serve as the foundation for subsequent possession analysis.

A dedicated possession recognition module is then applied to determine possession status based on detected events. This module includes a ball-holder detector, a pass event recognizer, and a time controller, working collaboratively to define each possession episode. When the ball remains in proximity to a player from a particular team for a predefined duration, possession is considered established. A successful pass or a change in ball control marks the end of the possession cycle. By summing each team's possession duration and comparing it to total match time, the system calculates time-based possession rates.

Unlike common approaches that define possession based on the number of completed passes, this system directly relies on temporal ratios. This strategy reduces reliance on

accurate pass detection, which may be error-prone in high-intensity or turnover-prone matches, and thereby enhances the stability and practical applicability of possession analysis.

3. Proposed Method

In this study, we explicitly distinguish between player-level possession determination and team-level ball possession estimation. Player-level possession determination refers to identifying whether the ball is controlled by a specific player at a given moment based on spatial and temporal constraints. Team-level ball possession estimation refers to calculating the percentage of total match time during which a team controls the ball. The former serves solely as an intermediate step to support the latter, which constitutes the primary objective of this study.

This section outlines the implementation framework of the proposed football match analysis system, as illustrated in Figure 1. The system takes a video sequence as input and employs a YOLO-based object detection model to identify players and the ball on a frame-by-frame basis. To automatically distinguish between teams, the detected player targets are further clustered using the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm [36]. This unsupervised method groups players based on jersey color distributions in the RGB space, eliminating the need for manual annotations or identity-specific information.

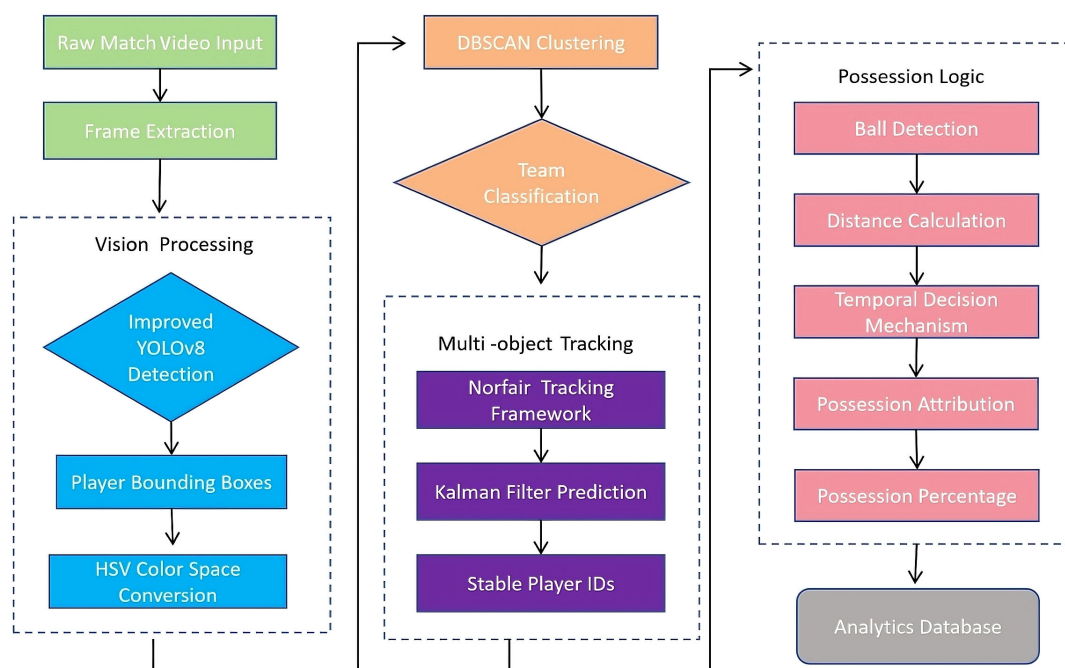


Figure 1. Experimental flow chart.

After classification, multi-object tracking is performed to maintain the temporal consistency of both player and ball identities across frames. This is achieved using the Norfair tracking library, an extension of the Simple Online and Realtime Tracking (SORT) algorithm proposed by Bewley et al. (2016) [37]. Norfair incorporates Kalman filtering, Hungarian matching, and customizable distance functions, allowing for robust performance under challenging conditions such as player occlusion or similar jerseys. Finally, the system uses the algorithm described in Algorithm 1 to infer ball possession events based on player proximity and possession duration. By aggregating the time each team maintains possession, the

system computes the overall ball possession rate, providing a more stable and context-aware alternative to common pass-based metrics.

Algorithm 1: Calculating for ball possession rate

Input: The bounding boxes of the ball $\mathbf{BB} = (X, Y)$, the bounding boxes of the players $\mathbf{bb}_i = [(x_1, y_1), (x_2, y_2), \dots, (x_i, y_i)]$, the image array of the players $\mathbf{Arr}_i$, enter/exit hysteresis (T_{in}, T_{out}) , minimum switch duration K , hold length H for missing detections, FPS for Videoframe rate.

Output: Ball possession rate

```

1. n_frame_0 = 0; n_frame_1 = 0; //Total frame count of possession for Team 0 and 1.
2. state = UNKNOWN; cand = UNKNOWN; k = 0; h = 0;
3. while (Current_Frame < Max_Frame) do
  //Step 1: Assign players to two teams.
4.  $Half\_Arr_i = Arr_i[0 : Arr_i.shape[0]/2, :]$ ; //only shirt is selected for clustering.
5.  $TP\_dict[i] = DBSCAN(\epsilon = 0.5, min\_sample = 6, Half\_Arr_i)$ ; //A dictionary of player
  and team relationships is obtained through DBSCAN clustering.
  //Step 2: Judging ball possession
6. if ball not detected then
7.    $h = h + 1$ ;
8.   if  $h = H * FPS$  then
9.     state = UNKNOWN;  $h = 0$ ;
10.   $D[i] = \sqrt{(X - x_i)^2 + (Y - y_i)^2}$ ; //Calculate the Euclidean distance between the foot-
    ball and the players.
11.  if  $\min(D)/heights[\min\_Index] < T_{in}$  then //where heights[min_Index] denotes the height
    of the bounding box of the nearest player to the ball, which is used to normalize the dis-
    tance.
12.    cand =  $TP\_dict[\min\_Index]$ ;
13.  else if  $\min(D)/heights[\min\_Index] > T_{out}$  then
14.    cand = UNKNOWN;
  //Step 3: Transition stabilization
15.  if cand == state then
16.     $k = 0$ ;
17.  else if cand != UNKNOWN then
18.     $k = k + 1$ ;
19.    if  $k \geq K * FPS$  then
20.      state = cand;  $k = 0$ ;
21.  if state == 0 then
22.    n_frame_0 = n_frame_0 + 1;
23.  else if state == 1 then
24.    n_frame_1 = n_frame_1 + 1;
  //Step 3: Calculating current ball possession rate
25.  bp_rate_0 =  $n\_frame\_0 / (n\_frame\_0 + n\_frame\_1)$ ;
26.  bp_rate_1 =  $n\_frame\_1 / (n\_frame\_0 + n\_frame\_1)$ ;
27. end while

```

3.1. Player and Ball Tracking

The SORT (Simple Online and Realtime Tracking) algorithm employs a fast-object detection framework, typically based on region-based convolutional neural networks (RCNN), to achieve efficient and real-time object tracking. It models object motion using a linear constant velocity model and applies a Kalman filter to predict and update the state of each object. For data association, SORT calculates a cost matrix based on the Intersection over Union (IoU) between detected objects and existing tracks, and uses the Hungarian

algorithm to optimally assign detections to trackers. When a new object enters the camera view, SORT initializes a new tracker to record its position and size. To prevent false positives, newly initialized trackers undergo a probationary period during which multiple consecutive detections are required before confirming the object's existence. Conversely, if an object is not detected for a predefined number of frames, the corresponding tracker is terminated to prevent resource waste and reduce tracking errors. Despite its simplicity and lightweight design—which make it highly suitable for real-time applications—SORT may underperform in challenging scenarios involving long-term occlusions, complex backgrounds, or overlapping targets.

In this system, YOLOv8 provides spatial observations of targets in each frame, such as bounding boxes or centroids, while Norfair performs cross-frame identity association using Kalman filtering and distance-based matching mechanisms. In scenarios such as football matches with frequent occlusions, tracking bounding box coordinates yields better performance. YOLO detection outputs are first converted into Norfair-compatible detection objects by encoding each target's bounding box or centroid as observation points (points), along with corresponding confidence scores (scores) and class labels. Norfair then predicts the state of each track using a Kalman filter and performs data association between detections and existing tracks using distance metrics such as IoU or Euclidean distance combined with the Hungarian algorithm. Matched detections are used to update track states, while unmatched detections initialize new tracks, and unmatched tracks are maintained for several frames through prediction, enabling stable multi-object tracking.

In real-world applications, the system is affected by environmental conditions such as poor lighting, player occlusions, and abrupt viewpoint changes. In such cases, Kalman filtering alone is insufficient, leading to missed or false detections and consequently broken trajectories that degrade prediction performance. To address this, appearance embeddings are introduced for target re-identification (ReID). Originally, when detections are missing for more than a certain threshold, the corresponding track is deleted and a reappearing target is assigned a new ID. With ReID, instead of being immediately removed, disappeared tracks are placed into a cache. When a target reappears, it is compared against both active tracks and cached tracks using a distance function; if it is determined to be the same target, the original track is recovered.

Camera motion or zoom can cause YOLO's detection boxes to drift globally in the image coordinate system, which may mislead the data association module and result in track loss. To address this, the system estimates global background motion using optical flow and maps detections into a virtual stationary coordinate system before tracking. Norfair provides a MotionEstimator module whose parameters include the maximum number of sampled corner points (max_points), minimum corner spacing (min_distance), optical flow window size, camera motion model (e.g., translation or homography), and corner quality threshold.

For each frame, the system first detects corner points with strong texture features in the previous frame and tracks them to the next frame using optical flow, retaining only successfully matched point pairs. It then applies Random Sample Consensus (RANSAC) to enforce geometric consistency by randomly sampling minimal point sets to fit a motion model and selecting the model with the largest number of inliers over multiple iterations, thereby removing outliers caused by foreground object motion or mismatches. During camera motion estimation, target regions are deliberately excluded from optical flow and corner computation so that only background points are used, preventing foreground motion from contaminating background flow estimation. Finally, the estimated transformation is passed to the Norfair Tracker to correct target state prediction, improving trajectory stability under camera motion.

3.2. Player Classification Module

Given the wide variety of jersey styles in real-world scenarios, using clustering algorithms to distinguish between different jerseys can significantly reduce manual labeling efforts compared to traditional classification methods. To minimize noise and eliminate irrelevant information during the experiment, the target images were cropped prior to player classification based on jersey colors. The cropping was guided by the coordinates and object labels returned by the model, with both width and height limited to the central 15–85% region of the image. The cropped images were then processed using median filtering to reduce noise and smooth the images.

Following preprocessing, HSV color histogram features were extracted from each player's region. The HSV color space was chosen over RGB because it is more robust under variable lighting conditions, which are common in football matches. The extracted features were converted into NumPy arrays and clustered using the DBSCAN algorithm.

Two key parameters in DBSCAN, ϵ and min_samples , greatly influence clustering performance. To achieve optimal clustering, a heuristic grid search was employed. Specifically, candidate values for ϵ ranged from 0.3 to 1.0 in increments of 0.1, and min_samples ranged from 3 to 10 (integers only). A nested loop iterated over all possible parameter combinations. For each combination, the silhouette score was computed to evaluate clustering quality. The silhouette score ranges from -1 to 1 , with higher values indicating better-defined clusters. The algorithm retained the parameter set yielding the highest silhouette score as the optimal configuration.

After clustering, players were categorized into two groups—Team1 and Team2—based on jersey styles. These new labels were stored in a list for use in subsequent modules.

3.3. Model Construction

To provide a clear understanding of the overall system design, the framework for calculating ball possession rate is illustrated in Figure 2. In this framework, Video frames are processed through two improved YOLOv8-based networks (YOLOv8-P2S3A for football detection and YOLOv8-HWD3A for player detection) to extract bounding box information. This information is subsequently fed into a post-processing module to determine ball possession rates over time.

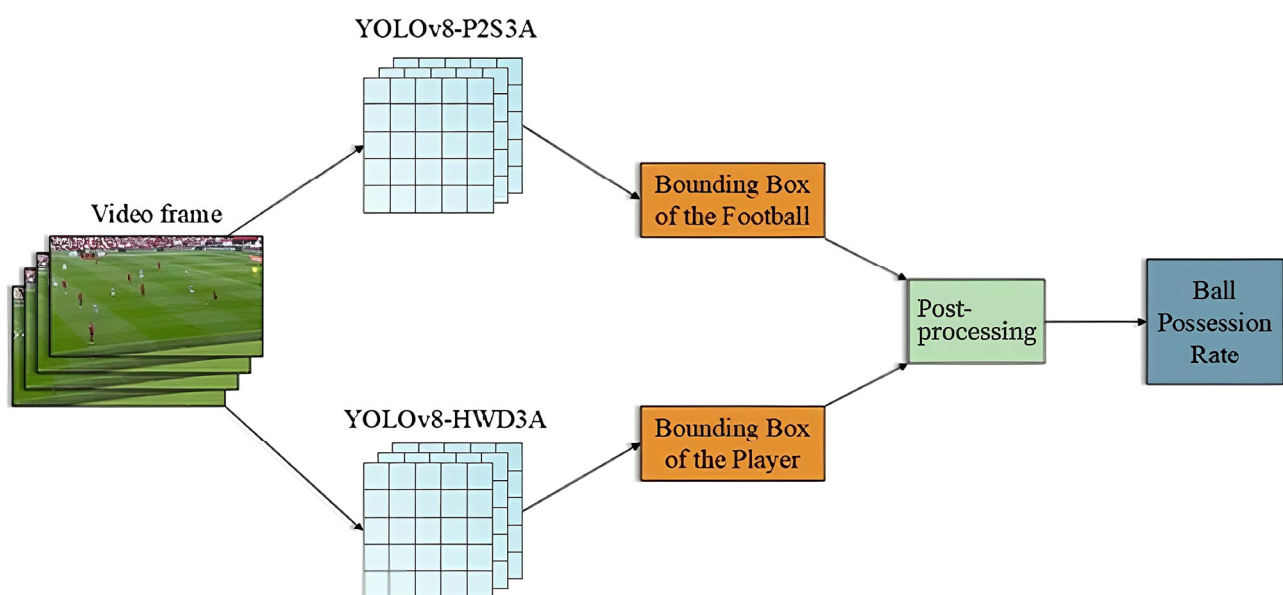


Figure 2. The illustration of the framework for calculating ball possession rate. The Video frames are processed through two improved YOLOv8 networks (YOLOv8-P2S3A and YOLOv8-HWD3A) to

obtain the bounding box information for the football and the players, and finally, this information is used to calculate the ball possession rate.

To further clarify the rationale behind our model enhancements, we focused primarily on the latter stages of the processing pipeline. As illustrated in Algorithm 1, this stage relies heavily on the accuracy of bounding box outputs to calculate ball possession. Consequently, the precision of possession estimation is inherently dependent on the detection performance of the YOLO models.

Recognizing this critical relationship, we reframed the core challenge of this study as improving the detection accuracy of YOLOv8 models when applied to football match footage, thereby ensuring more reliable ball possession calculation.

3.4. Data Processing

As illustrated in Figure 3, the video footage is first decomposed into frame-by-frame image sequences, each timestamped to ensure accurate temporal alignment for downstream analysis. These frames are then processed using the improved YOLOv8 model to detect player targets. For each detected bounding box, only the jersey region is cropped and uniformly resized. A square region centered on the cropped image is extracted and converted into the HSV color space, which helps minimize background interference and better represent the jersey color while mitigating the effects of lighting variations.

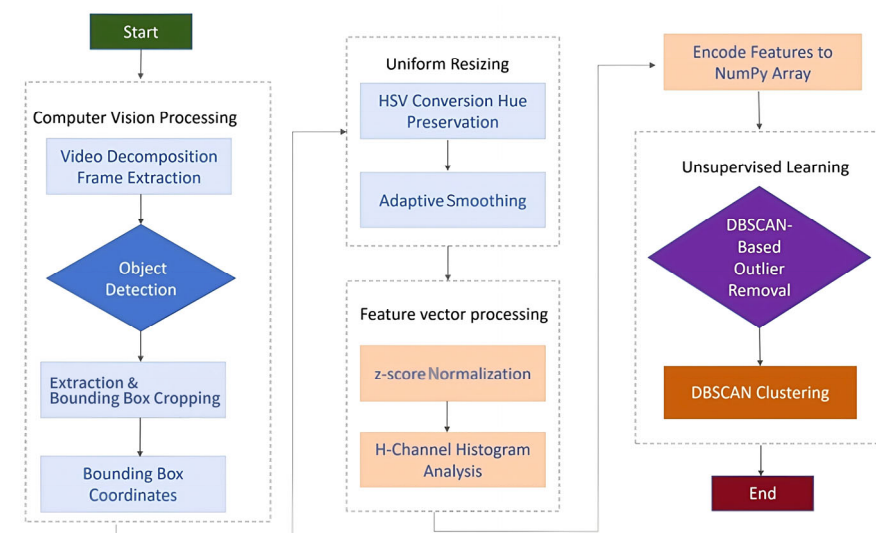


Figure 3. Flowchart for player image preprocessing and team clustering via jersey color analysis.

In cases where players overlap or are partially occluded, a confidence-based filtering mechanism selects the most prominent bounding box per region. Additionally, background pixels with low saturation or brightness (in HSV space) are filtered out before feature extraction to further suppress non-jersey interference.

Each frame undergoes a series of preprocessing steps to enhance robustness. The cropped jersey region is smoothed and transformed into a feature vector describing its visual characteristics. Prior to clustering, these feature vectors are standardized using z-score normalization to ensure consistent scaling and minimize bias caused by lighting changes or camera exposure.

The system calculates the color histogram of the H (hue) channel for each jersey region, divides it into 16 bins, and normalizes the values to form a compact color feature vector. These vectors are then converted into NumPy arrays for efficient computation and are clustered using the DBSCAN algorithm, which groups players into teams based on jersey color.

To maintain clustering accuracy throughout the game, this process is repeated dynamically every N frames, enabling adaptive classification under varying lighting or camera conditions. Furthermore, DBSCAN's built-in ability to identify outliers ensures that only relevant player data are included in possession calculations by discarding noise points with minimal cluster association. This completes the image preprocessing and feature clustering stage, providing reliable input for real-time possession estimation and team identification.

4. Experiments

4.1. Dataset Construction

Due to significant differences in both the number and size of players and the ball in football match footage, this system requires dedicated object detection models and corresponding training datasets for players and the ball respectively.

Existing public football datasets on the internet predominantly consist of static, single large images of footballs, which are not suitable for detecting footballs in dynamic match scenarios. Therefore, we constructed a new dataset specifically for this task. We downloaded match footage from several major football leagues and the UEFA Champions League from sports media websites. The videos were sampled at an average rate of 16 frames per second (fps), and after mixing frames from various games, 1000 screenshots were randomly selected. These images were annotated using the YOLO format in LabelImg, and the dataset was split into training, validation, and test sets in a 7:2:1 ratio.

For player detection, we utilized a player dataset available on Roboflow, which contains 1462 annotated images from football matches. Among these, 1023 images were used for training, 293 for validation, and 146 for testing.

4.2. YOLOv8 Architecture Enhancement

4.2.1. YOLOv8-P2S3A

YOLOv8-P2S3A is a modified version of the YOLOv8 model specifically optimized for football detection tasks. One of the key challenges in detecting footballs lies in their inherently small size. As the depth of a neural network increases, successive convolutional downsampling can lead to the loss of critical feature information for such small objects, which significantly compromises detection performance.

To address this issue, one effective strategy is to preserve and transmit more detailed information to the detection head. In this work, the P5 layer of the original YOLOv8 architecture was removed, and a new detection head was added at the P2 layer to better retain the features of small targets. Additionally, the StarBlock module 38 was integrated into the Neck component to enhance feature representation. A novel detection head, referred to as 3A-Detect, incorporating attention mechanisms, was designed to further improve detection accuracy (as illustrated in Figure 4).

In StarBlock, there is an element-wise multiplication of one-step tensors, which can be expressed as $(W_1^T X + B_1) * (W_2^T X + B_2)$. If we assume $W = \begin{bmatrix} W \\ B \end{bmatrix}$, $X = \begin{bmatrix} X \\ 1 \end{bmatrix}$, then it can be expressed as $(W_1^T X) * (W_2^T X)$ and then we expand it:

$$\begin{aligned} W_1^T X * W_2^T X &= \left(\sum_{i=1}^{d+1} w_1^i x^i \right) * \left(\sum_{j=1}^{d+1} w_2^j x^j \right) \\ &= \sum_{i=1}^{d+1} \sum_{j=1}^{d+1} w_1^i w_2^j x^i x^j \\ &= a_{(1,1)} x^1 x^1 + \dots + a_{(4,5)} x^4 x^5 + \dots + a_{(d+1,d+1)} x^{d+1} x^{d+1} \end{aligned} \quad (1)$$

The symbol “*” denotes the multiplication operation. We can see that this operation generates $\frac{(d+2)(d+1)}{2} \approx \left(\frac{d}{\sqrt{2}}\right)^2$ different terms, each of which is a nonlinear combination of the input features, indicating that they are independent latent dimensions. This operation can increase the dimension of the feature space and enrich the expression of features 39.

In 3A-Detect, we incorporate the Triplet Attention module into the original detection head of YOLOv8, as depicted in Figure 5. This enhancement allows the detection head to effectively learn nonlinear local dependencies among the channel, height, and width dimensions, thereby improving the model's performance with negligible additional computational cost.

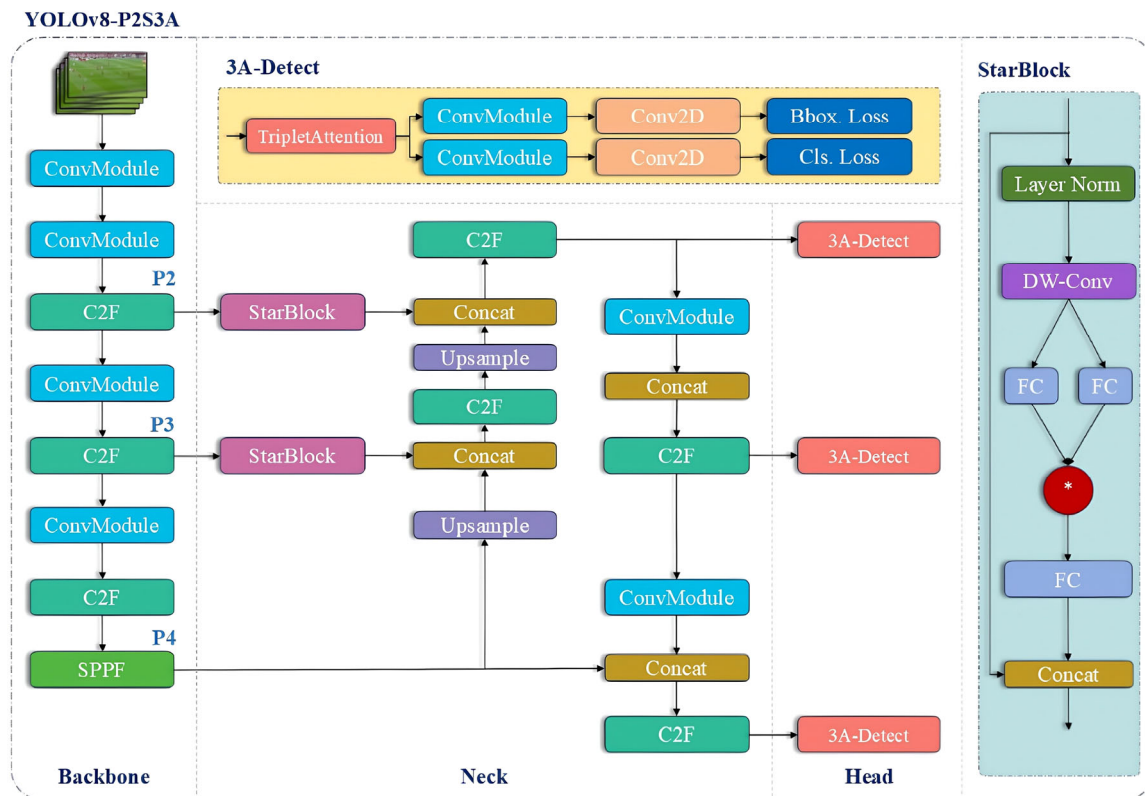


Figure 4. The illustration of YOLOv8-P2S3A.

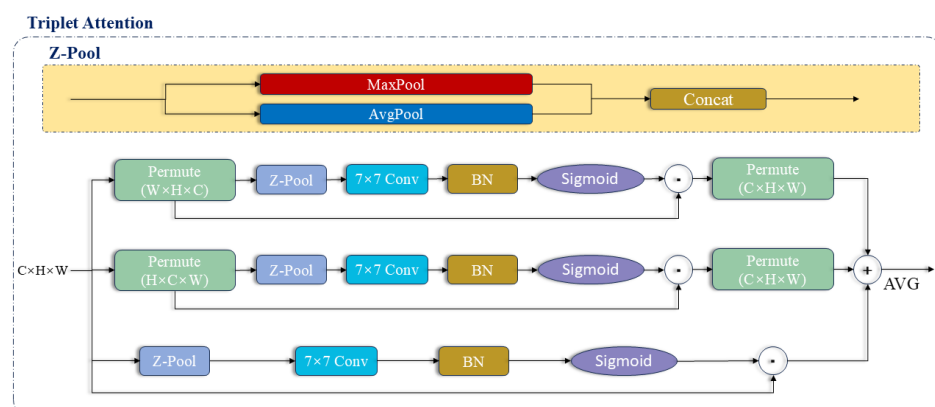


Figure 5. Illustration of the Triplet Attention module.

To better illustrate the effectiveness of our architectural improvements, we visualize and compare the attention heatmaps generated by the baseline YOLOv8 model and the proposed YOLOv8-P2S3A model, as shown in Figure 6.

The heatmaps clearly demonstrate that YOLOv8-P2S3A exhibits a much stronger and more focused attention on the football compared to the baseline YOLOv8 model. This indicates that the enhancements, particularly the introduction of the P2 detection head and

attention mechanisms, significantly improve the model's ability to capture small and critical objects such as footballs.



Figure 6. Visualization of attention heatmaps for football detection: YOLOv8 vs. YOLOv8-P2S3A.

4.2.2. YOLOv8-HWD3A

In Section 3.4 we discussed strategies for enhancing detection accuracy by transmitting relatively uncompressed information to the detection head. Initially, we attempted to apply a similar approach to player detection. However, our preliminary experiments revealed that this method yielded limited results in this task. Unlike footballs, players are not small-scale targets; thus, adding a P2-level detection head had limited benefits, while removing the P5-level detection head adversely affected performance.

In response, we retained the P5 detection head and introduced Haar wavelet downsampling 40 to mitigate information loss during downsampling. Based on this strategy, we redesigned the Backbone and Neck structures of the model, as illustrated in Figure 7.

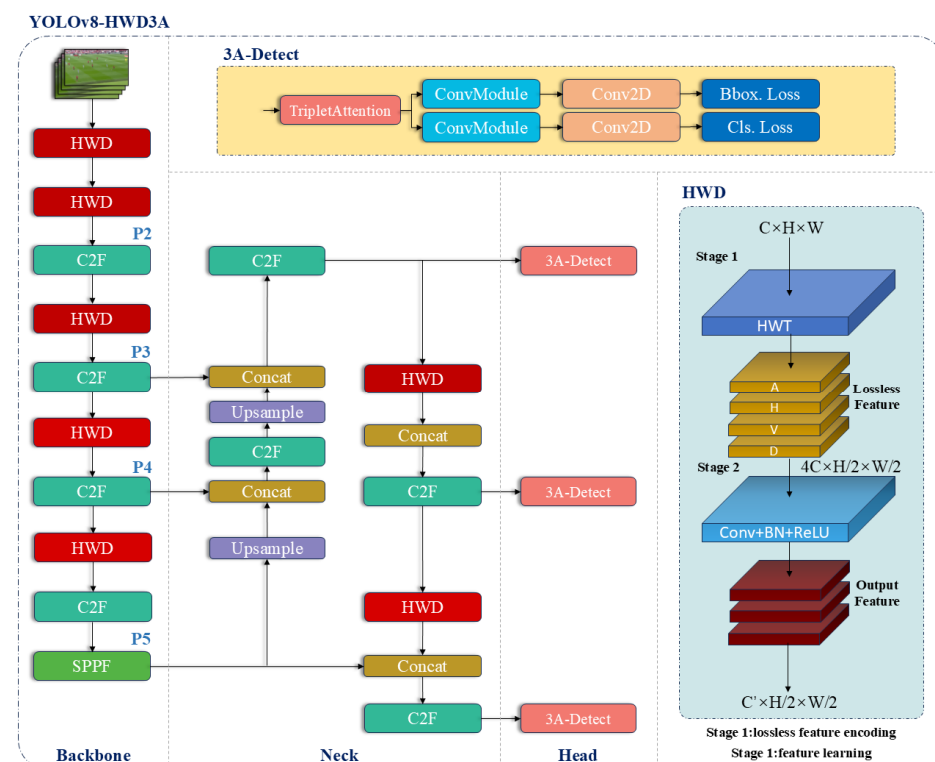


Figure 7. Illustration of YOLOv8-HWD3A. A: the approximate (low-frequency) component. H: the detail components (high-frequency) in the horizontal directions. VW: the detail components (high-

frequency) in the vertical directions. D: the detail components (high-frequency) in the diagonal directions.

The Haar wavelet transform decomposes input signals into four components: one low-frequency (approximate) and three high-frequency (detail) components along the horizontal, vertical, and diagonal directions. This decomposition helps preserve critical information related to object boundaries, scales, and textures during downsampling 41, which is particularly valuable in football scenarios where frequent player occlusions occur.

To visualize the effect of these modifications, Figure 8 presents the attention heatmaps generated by the baseline YOLOv8 model and the proposed YOLOv8-HWD3A model during inference. While the improvement in focus is not as visually pronounced as in Figure 6, Section 3.4 will provide a more detailed quantitative comparison demonstrating the superior performance of YOLOv8-HWD3A. In the attention heatmaps, warmer colors (red and yellow) indicate higher attention weights and stronger model focus, while cooler colors (blue and green) represent lower attention responses.



Figure 8. Visualization of attention heatmaps for player detection: YOLOv8 vs. YOLOv8-HWD3A.

4.2.3. Quantitative Evaluation of the Improved YOLOv8 Models

After introducing the structural enhancements and visual detection improvements of the proposed models, we conducted a series of experiments to comprehensively evaluate their performance across multiple tasks. The evaluation focused on football detection, player detection, and ball possession estimation, comparing the proposed models with several state-of-the-art methods.

The overall experimental results are summarized in Tables 1 and 2. In the football detection task, the proposed YOLOv8-P2S3A model achieves the highest accuracy with only 1.9 M parameters, reaching 79.4% Validation AP (%) and an inference latency of just 9.5 ms. It outperforms both YOLOv10n (67.7%) and YOLOv8n (66.3%), demonstrating superior precision and computational efficiency.

Table 1. Comparisons with state-of-the-art models for football detection.

Model	Parameters (M)	Validation AP (%)	Latency (ms)
YOLOv8n 42	3.0	66.3	7.4
YOLO11 43	2.5	64.3	2.6
YOLOv10n 44	2.7	67.7	10.1
YOLOv8-P2S3A (Ours)	1.9	79.4	9.5
YOLOv5n 45	2.5	62.5	3.2
YOLOv6n 46	4.2	14.4	10.7

Note: “Parameters (M)” denotes the number of model parameters (in millions). “Validation AP (%)” represents the average precision of the validation set for ball and player detection, respectively. “Latency” measures the average inference time per image (in milliseconds). The root mean square

error (RMSE) is computed as $RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y - \hat{y})^2}$ and MAE as $\frac{1}{n} \sum_{i=1}^n |y - \hat{y}|$, where y denotes the ground-truth possession percentage reported in the official match statistics, $\hat{y}$ denotes the predicted possession percentage, and n is the total number of evaluated matches.

Table 2. Comparisons with state-of-the-art models in player detection.

Model	Parameters (M)	Validation AP (%)	Latency (ms)
YOLOv8	3.0	70.6	3.7
YOLO11	2.5	69.2	2.6
YOLOv10n	2.7	69.7	8.2
YOLOv8-HWD3A(Ours)	2.6	71.1	3.3
YOLOv5n	2.5	70.1	2.6
YOLOv6n	4.2	66.2	9.6

Note: “Parameters (M)” denotes the number of model parameters (in millions). “Validation AP (%)” represents the average precision of the validation set for ball and player detection, respectively. “Latency” measures the average inference time per image (in milliseconds).

In the player detection task, the YOLOv8-HWD3A model achieves 71.1% Validation AP (%) with only 2.6 M parameters, and an inference latency of 3.3 ms, surpassing YOLOv11 (69.2%) and YOLOv5n (70.1%) in both accuracy and speed.

Moreover, as shown in Table 3, in the ball possession estimation task, the combination of YOLOv8-P2S3A + YOLOv8-HWD3A yields a root mean square error (RMSE) of 4.85, the lowest among all tested models, and significantly lower than the YOLOv6n-based combination (12.75).

Table 3. Comparisons of possession rate estimation performance.

Ball Detection Model	Player Detection Model	RMSE	MAE
YOLOv8n	YOLOv8 42	5.07	4.23
YOLO11	YOLO11 43	5.16	4.39
YOLOv10n	YOLOv10n 44	5.09	4.35
YOLOv8-P2S3A (Ours)	YOLOv8-HWD3A (Ours)	4.85	3.98
YOLOv5n	YOLOv5n 45	5.11	4.32
YOLOv6n	YOLOv6n 46	12.75	9.03

Furthermore, to verify the robustness and reproducibility of the proposed ball possession estimation module, we performed a one-factor sensitivity analysis on the key parameters (T_{in} , T_{out} , K , H). As shown in Table 4, the default configuration ($T_{in} = 0.6$, $T_{out} = 1.0$, $K = 0.30$, $H = 0.20$) achieves the lowest error with $RMSE = 4.87$ and $MAE = 3.98$. When scaling the hysteresis thresholds by α from 0.8 to 1.2, the performance varies only mildly ($RMSE$: 4.87–5.01; MAE : 3.98–4.10), indicating that the distance-threshold decision is relatively stable around the default region. In contrast, the temporal parameters exhibit a clearer preferable range: setting K too small ($K = 0.10$) increases the error ($RMSE = 5.07$, $MAE = 4.13$), likely due to frequent state oscillations under short-term noise, while overly large K ($K = 0.50$) also degrades performance ($RMSE = 5.10$, $MAE = 4.15$), suggesting delayed response to genuine possession transitions. A similar trend is observed for the hold length H : removing the hold mechanism ($H = 0.00$) yields higher errors ($RMSE = 5.01$, $MAE = 4.12$), whereas an excessively long hold ($H = 0.40$) further increases MAE to 4.25, implying that over-holding can smear rapid exchanges.

In summary, the YOLOv8-P2S3A + YOLOv8-HWD3A pipeline achieves the best overall performance across football detection, player detection, and ball possession estimation tasks, making it the optimal solution under current experimental conditions.

Table 4. Sensitivity analysis of the proposed method.

Group	Setting ID	T _{in}	T _{out}	K	H	RMSE	MAE
Threshold scale	$\alpha = 0.8$	0.48	0.8	0.3	0.2	4.91 (↑0.06)	4.03 (↑0.05)
	$\alpha = 0.9$	0.54	0.9	0.3	0.2	4.89 (↑0.04)	3.99 (↑0.01)
	$\alpha = 1.0$ (default)	0.6	1	0.3	0.2	4.85	3.98
	$\alpha = 1.1$	0.66	1.1	0.3	0.2	4.93 (↑0.08)	4.01 (↑0.03)
	$\alpha = 1.2$	0.72	1.2	0.3	0.2	5.01 (↑0.16)	4.10 (↑0.12)
Switch duration	K = 0.10	0.6	1	0.1	0.2	5.07 (↑0.22)	4.13 (↑0.15)
	K = 0.20	0.6	1	0.2	0.2	4.94 (↑0.09)	4.04 (↑0.06)
	K = 0.30 (default)	0.6	1	0.3	0.2	4.85	3.98
	K = 0.40	0.6	1	0.4	0.2	4.96 (↑0.11)	4.05 (↑0.07)
	K = 0.50	0.6	1	0.5	0.2	5.10 (↑0.25)	4.15 (↑0.17)
Hold length	H = 0.00	0.6	1	0.3	0	5.01 (↑0.16)	4.12 (↑0.14)
	H = 0.10	0.6	1	0.3	0.1	4.92 (↑0.07)	4.05 (↑0.07)
	H = 0.20 (default)	0.6	1	0.3	0.2	4.85	3.98
	H = 0.30	0.6	1	0.3	0.3	4.95 (↑0.10)	4.10 (↑0.12)
	H = 0.40	0.6	1	0.3	0.4	5.09 (↑0.24)	4.25 (↑0.27)

Note: “↑” denotes that the corresponding metric shows performance improvement compared with the baseline model.

4.3. Team Classification via DBSCAN Clustering

To ensure accurate identification of teams in football match analysis, the task of classifying detected players based on their jersey colors becomes essential for ball possession analysis. Given that only players from the two teams and the ball are relevant, and that the referee has already been filtered out in the previous stage, this step focuses on distinguishing between the two teams based on the high contrast often observed in their jersey colors during official football matches. This allows for an unsupervised learning approach, which reduces the reliance on extensive manual annotations typical of traditional supervised methods.

In this study, we propose using the DBSCAN (Density-Based Spatial Clustering of Applications with Noise) algorithm, a powerful density-based clustering technique capable of handling noise and non-convex cluster shapes without requiring prior assumptions about the number or shape of clusters. DBSCAN’s ability to automatically determine the number of clusters makes it an ideal choice for this task.

Two key parameters influence the performance of DBSCAN:

ϵ (epsilon): the radius of the neighborhood, which defines the similarity threshold between player jersey color features (i.e., how close two players must be to be considered on the same team).

min_samples: the minimum number of neighbors required for a point to be considered a core point.

To optimize clustering performance, we generated 100 candidate values for ϵ and min_samples using a fixed step size within a specified range. Since the feature vectors (based on normalized color histograms) typically range from 0 to 1, the ϵ values were chosen within (0, 0.3). For min_samples, the range was set between 10% and 50% of the total number of player targets per image, which aligns with the typical player count in 11-a-side football matches.

Each pair of (ϵ , min_samples) candidates was applied to the DBSCAN algorithm to obtain clustering results. In practice, although each team consists of players wearing different jersey colors (e.g., outfield players and goalkeepers), goalkeepers are involved in far fewer actions and have limited ball contact. Thus, for simplicity, goalkeepers were excluded

from clustering, and we selected parameter combinations that resulted in exactly two clusters.

To further improve accuracy, we introduced a cost function to identify the parameter set that best balances the number of players assigned to each team. Let N_1 and N_2 represent the number of data points classified as Team 1 and Team 2, respectively. The cost function is defined as follows:

$$\text{cost}(\epsilon, \min_{\text{samples}}) = |N_1 - N_2| \quad (2)$$

The combination of ϵ and $\min_{\text{samples}}$ that minimizes this cost is selected as the final clustering parameters.

To evaluate the effectiveness of this method, we randomly extracted 20 frames from five different football match videos. For each frame group, we applied the clustering approach described above and recorded the number of correctly and incorrectly classified player targets for Team 1 and Team 2. The classification accuracy results are summarized in Table 5.

The results demonstrate that the method performs well and remains stable across different matches. As shown in Figure 9, the proposed algorithm can accurately distinguish between players, even in scenes where player occlusion is severe.

Table 5. Team classification accuracy.

Video	Team 1	Team 2
1	97.8%	92%
2	95.3%	92.3%
3	90.9%	95.2%
4	93.7%	89.4%
5	92.1%	91.7%

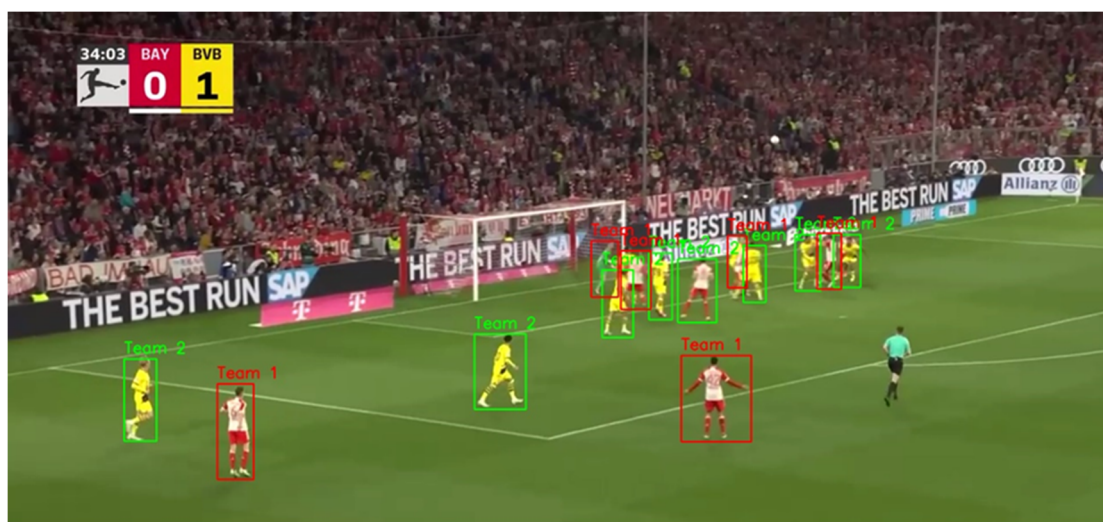


Figure 9. Team classification performance.

4.4. Performance of the Ball Possession Estimation System

The ball possession visualization shown in the image is implemented using the Python Imaging Library (PIL version 9.5.0) for real-time frame-by-frame image processing and rendering. The system analyzes each video frame to track cumulative ball possession time, total match duration, and calculate the possession ratio between the two teams. Whenever the accumulated game time reaches a full second (e.g., 1 s, 2 s, 3 s, etc.), the system updates the

visualization, overlaying a possession bar chart and timestamp that dynamically display the percentage of ball possession held by each side.

In the visualization, Team 1 and Team 2 correspond to the home and away teams, respectively, with numeric values indicating the current percentage of ball possession for each team up to that point in the match. As shown in the figure, the system also overlays contextual match information—including player positions, speeds, distances, and ball location—to enhance tactical understanding. All processed frames are compiled into a new output video, enabling continuous real-time tracking of possession throughout the game.

To validate the effectiveness of the approach, the system was tested on four high-profile football matches, including Al-Ahli SFC (white kit) vs. Al Nassr FC, Manchester United (red kit) vs. Manchester City, Liverpool FC vs. Real Madrid (white kit), and Shanghai Port vs. Shanghai Shenhua (blue kit). The results, visualized in Figures 10–13, clearly demonstrate the differing ball possession percentages between the teams in each match.

The red triangle indicates the player currently in possession of the ball, green triangles show potential next ball receivers, numbers above players denote jersey numbers, text below players indicates running speed and distance covered, and colored circles represent different teams. The ball control bar (bottom-right) shows each team's possession percentage, with colors corresponding to the team on the field.

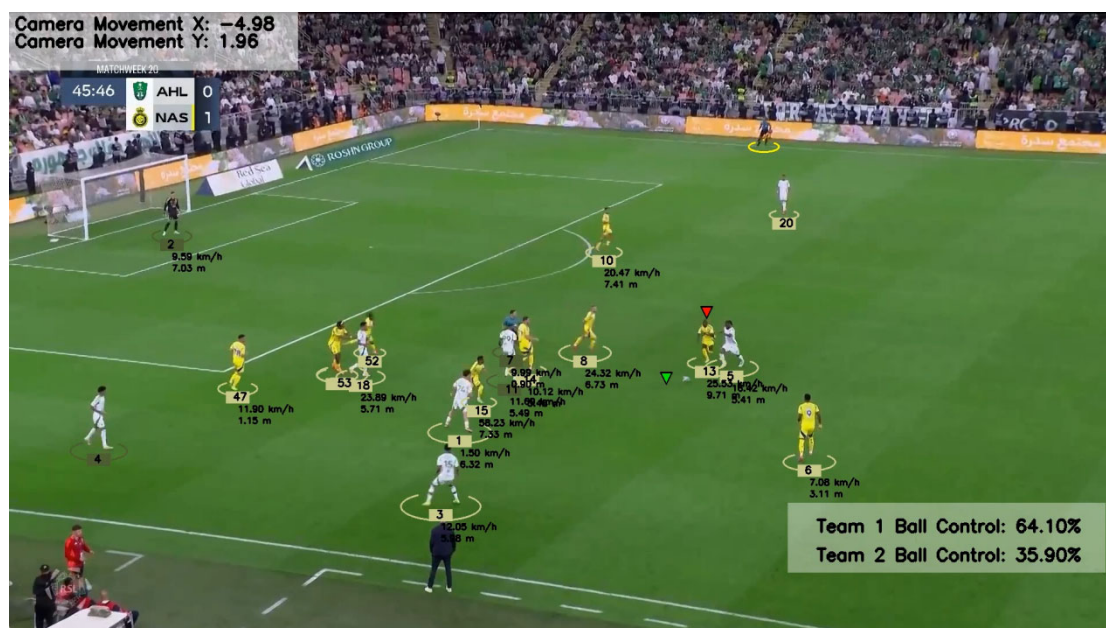


Figure 10. Al-Ittihad Club vs. Al Nassr FC. Figure 10 is from <https://www.bilibili.com/video/BV155KPerEeo>, accessed on 6 March 2025.

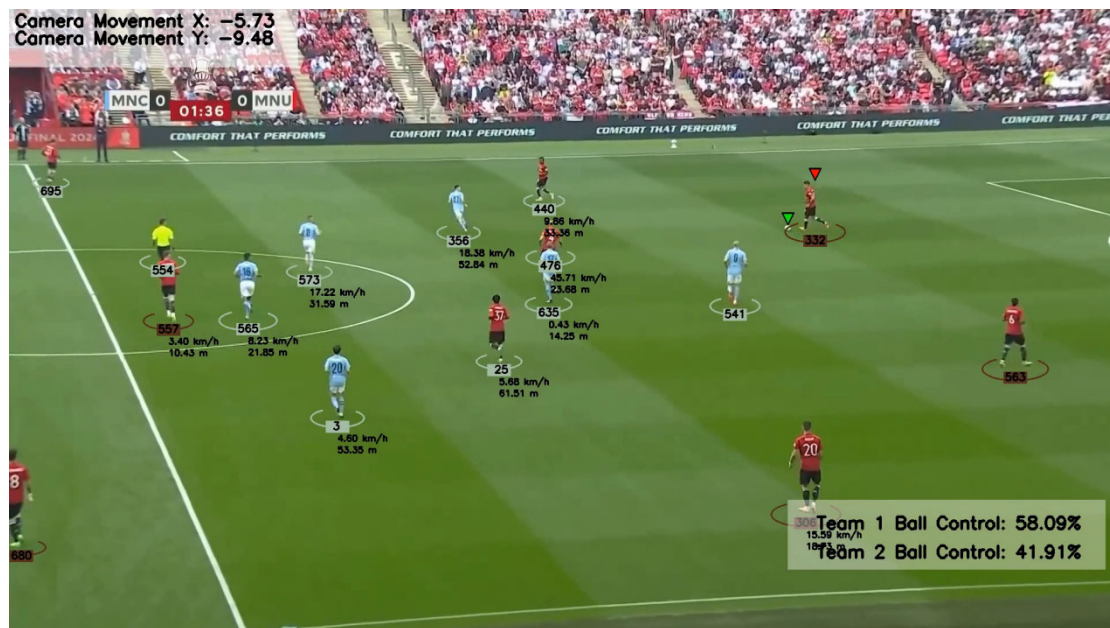


Figure 11. Manchester City vs. Manchester United. Figure 11 is from <https://www.bilibili.com/video/BV1tz421a79G/?p=2>, accessed on 6 March 2025.



Figure 12. Liverpool FC vs. Real Madrid. Figure 12 is from <https://www.youtube.com/watch?v=oXn0KPPHzuY>, accessed on 6 March 2025.

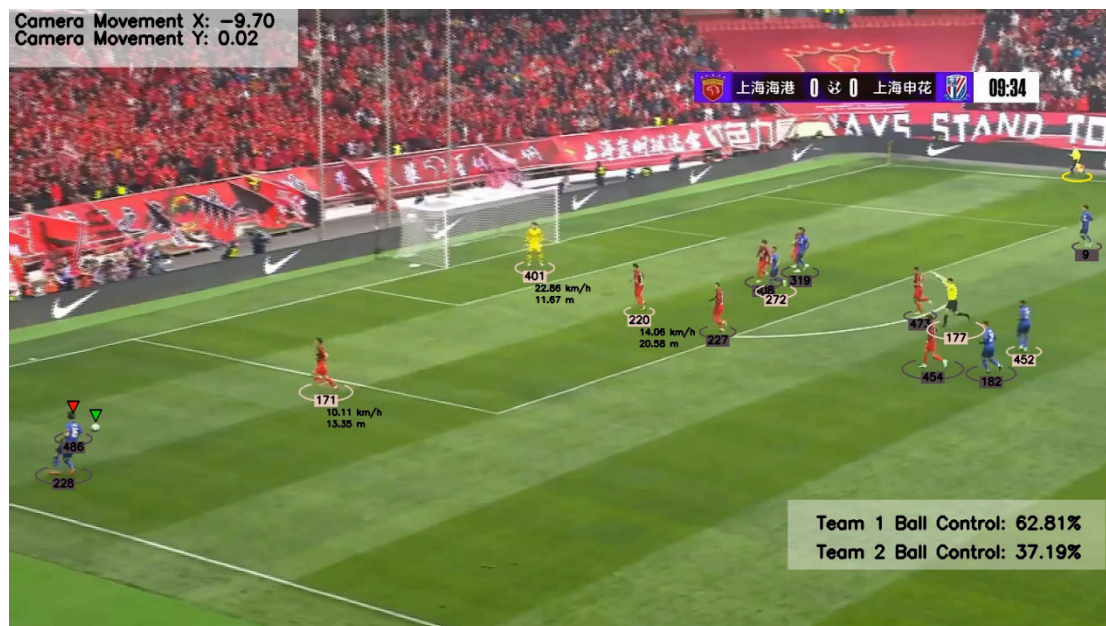


Figure 13. Shanghai Port Football Club vs. Shanghai Shenhua Football Club” Figure 13 is from <https://b23.tv/pNOhG53>, accessed on 6 March 2025.

5. Conclusions

In this study, we proposed a footballVideoanalysis system based on deep learning techniques, capable of automatically estimating ball possession in real time. This approach aims to alleviate the heavy reliance on manual statistics in traditional possession analysis and enhance the efficiency of football match interpretation.

To improve detection performance in complex match scenarios, we designed and integrated two customized YOLOv8 variants: YOLOv8-P2S3A for accurate detection of the football and YOLOv8-HWD3A for robust identification of players. These tailored models enhanced the detection accuracy, especially for small or occluded targets, addressing common challenges in sportsVideoanalysis. The system achieved an overall detection precision of 0.837 across all classes, providing a solid foundation for downstream tasks such as team classification and possession computation.

For team classification, we adopted an unsupervised clustering strategy to address the variability of jerseys across different teams and matches. Compared to traditional supervised methods requiring extensive manual labeling, this approach significantly reduces workload while maintaining high clustering accuracy and robustness.

Through the integration of improved detection modules and a reliable post-processing framework, our system can efficiently determine the bounding boxes of players and the football, calculate possession events, and derive meaningful ball possession rates in real time. This lightweight yet effective solution delivers valuable tactical insights to coaches, analysts, and players through automated visual analysis of game footage.

6. Future Work

While the current system has demonstrated promising results, real-world football matches present numerous challenges that require further investigation. These include the frequent changes in player motion, complex interactions and occlusions between multiple objects, and special scenarios such as corner kicks, free kicks, and penalty situations, which may temporarily alter game dynamics and possession patterns.

To enhance the accuracy and adaptability of the system, future work will focus on the following directions:

Improving illumination robustness for team clustering: although HSV is effective for color-based clustering, abrupt lighting changes may still affect hue stability. Future work will explore more illumination-invariant representations (e.g., Lab space, color constancy, or learned appearance embeddings). Improving possession inference beyond image-plane Euclidean distance: the distance cue can be unreliable under occlusions, overlaps, fast passes, and strong perspective distortion. Future work will explore homography-based field normalization and richer ball–player interaction cues (e.g., 3D trajectory features). Incorporating temporal information using sequence modeling (e.g., LSTM or transformer-based approaches) to better understand player movement and ball transitions over time. Enhancing robustness in densely populated scenes and occluded scenarios through multi-frame fusion or 3D pose estimation. Expanding the dataset to include a broader variety of match types, jersey styles, and lighting conditions to improve generalization. Integrating multi-modal data, such as commentary text, player statistics, or sensor-based tracking, for a more comprehensive match analysis framework. In addition, future research will further investigate scalability and computational efficiency, particularly for high-resolution broadcast inputs and potential real-time deployment scenarios. Model compression, lightweight architectures, and inference acceleration techniques may be explored to improve practical applicability.

With continuous refinement, the system holds the potential to become a valuable tool for intelligent sports analytics and decision-making in both professional and amateur football settings.

Author Contributions: R.G.: Writing—review and editing, Writing—original draft, Methodology. Y.Z.: Writing—review and editing, Validation. R.D.: Supervision. Y.L.: Investigation, Supervision. Y.C.: Software. L.Y.: Data curation. X.X.: Data curation. J.Z. (Jianpeng Zhang): Supervision. Z.M.: Data curation. J.Y.: Writing—review and editing, Funding acquisition, Conceptualization. J.Z. (Jiajin Zhang): Writing—review and editing, Funding acquisition, Conceptualization. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Yunnan Province Basic Research Joint Project (Project No.: 202301BD070001-114) and the Undergraduate Education and Teaching Reform Research Projects of Yunnan Agricultural University (Project Nos.: 2024-55 and 2021YLKC126).

Institutional Review Board Statement: This study did not involve any direct interaction with human participants or the collection of personally identifiable data. All video materials used for the analysis were obtained from publicly available football match recordings. Therefore, ethical approval and informed consent were not required, in accordance with the institutional and journal guidelines.

Data Availability Statement: The datasets generated and analyzed during the current study are publicly available at <https://www.soccer-net.org/data> (accessed on 27 January 2026) and https://github.com/AkagiRitsuko-r/soccer_data (accessed on 27 January 2026). The code used for data preprocessing, model training, and analysis in this study is openly accessible at <https://github.com/jiecao233/YOLOv8-P2S3A-HWD3A> (accessed on 27 January 2026).

Conflicts of Interest: No potential conflicts of interest were reported by the authors.

Reference

1. Bradley, P.S.; Lago-Peñas, C.; Rey, E.; Gomez Diaz, A. The effect of high and low percentage ball possession on physical and technical profiles in English FA Premier League soccer matches. *J. Sports Sci.* **2013**, *31*, 1261–1270.
2. Ali, M.L.; Zhang, Z. The YOLO framework: A comprehensive review of evolution, applications, and benchmarks in object detection. *Computers* **2024**, *13*, 336.
3. Bradley, P.S.; Lago-Peñas, C.; Rey, E.; Sampaio, J. The influence of situational variables on ball possession in the English Premier League. *J. Sports Sci.* **2014**, *32*, 1867–1873.

4. Göral, K. Passing success percentages and ball possession rates of successful teams in 2014 FIFA World Cup. *Int. J. Sport Cult. Sci.* **2015**, *3*, 86–95.
5. Jones, P.D.; James, N.; Mellalieu, S.D. Possession as a performance indicator in soccer. *Int. J. Perform. Anal. Sport* **2004**, *4*, 98–102.
6. Liu, H.; Hopkins, W.; Gómez, A.M.; Molinuevo, S.J. Inter-operator reliability of live football match statistics from OPTA Sportsdata. *Int. J. Perform. Anal. Sport* **2013**, *13*, 803–821.
7. Barros, R.M.; Misuta, M.S.; Menezes, R.P.; Figueroa, P.J.; Moura, F.A.; Cunha, S.A.; Anido, R.; Leite, N.J. Analysis of the distances covered by first division Brazilian soccer players obtained with an automatic tracking method. *J. Sports Sci. Med.* **2007**, *6*, 233.
8. Shankara, V.; Ahmed, S.; Sneha, M.; Jayabalasamy, G. Object Detection and Tracking for Football Data Analytics. In Proceedings of the 1st International Conference on Artificial Intelligence, Communication, IoT, Data Engineering and Security, IACIDS 2023, Lavasa, India, 23–25 November 2023.
9. Wang, B. Football sportsVideotracking and detection technology based on YOLOv5 and DeepSORT. *Discov. Appl. Sci.* **2025**, *7*, 563.
10. Richly, K.; Moritz, F.; Schwarz, C. Utilizing artificial neural networks to detect compound events in spatio-temporal soccer data. In Proceedings of the 2017 SIGKDD Workshop MiLeTS, Halifax, NS, Canada, 14 August 2017; pp. 13–17.
11. Sun, J.; Huang, H.; Yang, C.; Jiang, Z.; Hwang, J. Gta: Global tracklet association for multi-object tracking in sports. In Proceedings of the Asian Conference on Computer Vision (ACCV) Workshops, Hanoi, Vietnam, 8–12 December 2024.
12. Gran-Henriksen, M.; Lindgaard, H.A.; Kiss, G.; Lindseth, F. Deep HM-SORT: Enhancing Multi-Object Tracking in Sports with Deep Features, Harmonic Mean, and Expansion IOU. *arXiv* **2024**, arXiv:2406.12081.
13. Scott, A.; Uchida, I.; Ding, N.; Umamoto, R.; Bunker, R.; Kobayashi, R.; Koyama, T.; Onishi, M.; Kameda, Y.; Fujii, K. Teamtrack: A dataset for multi-sport multi-object tracking in full-pitch videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, Seattle, WA, USA, 17–18 June 2024.
14. Glasser, H. The problem with possession: The inside story of soccer's most controversial stat. *Slate*, 27 June 2014.
15. Sarkar, S.; Chakrabarti, A.; Prasad Mukherjee, D. Generation of ball possession statistics in soccer using minimum-cost flow network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–17 June 2019.
16. Khaustov, V.; Mozgovoy, M. Recognizing events in spatiotemporal soccer data. *Appl. Sci.* **2020**, *10*, 8046.
17. Morra, L.; Manigrasso, F.; Canto, G.; Gianfrate, C.; Guarino, E.; Lamberti, F. Slicing and dicing soccer: Automatic detection of complex events from spatio-temporal data. In Proceedings of the International Conference on Image Analysis and Recognition, Póvoa de Varzim, Portugal, 24–26 June 2020; Springer: Cham, Switzerland, 2020; pp. 107–121.
18. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25, Proceedings of the 26th Annual Conference on Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–8 December 2012*; Morgan Kaufmann Publishers, Inc.: Burlington, MA, USA, 2012.
19. Awaludin, I.; Hidayatullah, P.; Hutahae, J.; Parta, D.G. Detection and object position measurement using computer vision on humanoid soccer. In Proceedings of the 2013 International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia, 7–8 October 2013; IEEE: New York, NY, USA, 2013; pp. 88–92.
20. Panse, N.; Mahabaleshwarkar, A. A dataset & methodology for computer vision based offside detection in soccer. In Proceedings of the 3rd International Workshop on Multimedia Content Analysis in Sports, Seattle, WA, USA, 16 October 2020; pp. 19–26.
21. Borghesi, M.; Costa, L.D.; Morra, L.; Lamberti, F. Using Temporal Convolutional Networks to estimate ball possession in soccer games. *Expert Syst. Appl.* **2023**, *223*, 119780.
22. Ester, M.; Kriegl, H.P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), Portland, OR, USA, 2–4 August 1996; Volume 96, No. 34, pp. 226–231.
23. Jiang, H.; Lu, Y.; Xue, J. Automatic soccerVideoevent detection based on a deep neural network combined CNN and RNN. In 2016 IEEE 28th International Conference on Tools with Artificial Intelligence (ICTAI), San Jose, CA, USA, 6–8 November 2016; IEEE: New York, NY, USA, 2016; pp. 490–494.
24. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
25. Hong, Y.; Ling, C.; Ye, Z. End-to-end soccerVideoscene and event classification with deep transfer learning. In Proceedings of the 2018 International Conference on Intelligent Systems and Computer Vision (ISCV), Fez, Morocco, 2–4 April 2018; IEEE: New York, NY, USA, 2018; pp. 1–4.
26. Xu, J.; Tasaka, K. Keep your eye on the ball: Detection of kicking motions in multi-view 4K soccer videos. *ITE Trans. Media Technol. Appl.* **2020**, *8*, 81–88.

27. Sorano, D.; Carrara, F.; Cintia, P.; Falchi, F.; Pappalardo, L. Automatic pass annotation from soccerVideostreams based on object detection and LSTM. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science and Demo Track: European Conference, ECML PKDD 2020, Ghent, Belgium, 14–18 September 2020, Proceedings, Part V*; Springer International Publishing: Cham, Switzerland, 2021; pp. 475–490.
28. Hu, X.; Cao, Y.; Sun, Y.; Tang, T. Railway automatic switch stationary contacts wear detection under few-shot occasions. *IEEE Trans. Intell. Transp. Syst.* **2021**, *23*, 14893–14907.
29. Berclaz, J.; Fleuret, F.; Turetken, E.; Fua, P. Multiple object tracking using k-shortest paths optimization. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1806–1819.
30. Wojke, N.; Bewley, A.; Paulus, D. Simple online and realtime tracking with a deep association metric. In Proceedings of the 2017 IEEE international conference on image processing (ICIP), Beijing, China, 17–20 September 2017; IEEE: New York, NY, USA, 2017; pp. 3645–3649.
31. Kandimalla, V.; Richard, M.; Smith, F.; Quirion, J.; Torgo, L.; Whidden, C. Automated detection, classification and counting of fish in fish passages with deep learning. *Front. Mar. Sci.* **2022**, *8*, 823173.
32. Horton, M. Learning feature representations from football tracking. In Proceedings of the MIT Sloan Sports Analytics Conference, Boston, MA, USA, 6–7 March 2020.
33. Lucey, P.; Bialkowski, A.; Carr, P.; Morgan, S.; Matthews, I.; Sheikh, Y. Representing and discovering adversarial team behaviors using player roles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 2706–2713.
34. Link, D.; Hoernig, M. Individual ball possession in soccer. *PLoS ONE* **2017**, *12*, e0179953.
35. Sanford, R.; Gorji, S.; Hafemann, L.G.; Pourbabaee, B.; Javan, M. Group activity detection from trajectory and Videodata in soccer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 898–899.
36. Deng, D. DBSCAN clustering algorithm based on density. In Proceedings of the 2020 7th International Forum on Electrical Engineering and Automation (IFEEA), Hefei, China, 25–27 September 2020; IEEE: New York, NY, USA, 2020; pp. 949–953.
37. Bewley, A.; Ge, Z.; Ott, L.; Ramos, F.; Upcroft, B. Simple online and realtime tracking. In Proceedings of the 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, 25–28 September 2016; IEEE: New York, NY, USA, 2016; pp. 3464–3468. Available online: <https://tryolabs.com/blog/2022/09/20/announcing-norfair-2.0-open-source-real-time-multi-object-tracking-library> (accessed on 27 January 2026).
38. Wang, Z.; Zheng, L.; Liu, Y.; Li, Y.; Wang, S. Towards real-time multi-object tracking. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer International Publishing: Cham, Switzerland, 2020; pp. 107–122.
39. Ma, X.; Dai, X.; Bai, Y.; Wang, Y.; Fu, Y. Rewrite the Stars. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024.
40. Misra, D.; Nalamada, T.; Arasanipalai, A.U.; Hou, Q. Rotate to attend: Convolutional triplet attention module. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Virtual, 5–9 January 2021.
41. Xu, G.; Liao, W.; Zhang, X.; Li, C.; He, X.; Wu, X. Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern Recognit.* **2023**, *143*, 109819.
42. Jocher, G. YOLOv8. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 18 September 2025).
43. Jocher, G. YOLO11. 2024. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 18 September 2025).
44. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J. YOLOv10: Real-time end-to-end object detection. *arXiv* **2024**, arXiv:2405.14458.
45. Jocher, G. YOLOv5. 2024. Available online: <https://github.com/ultralytics/yolov5> (accessed on 18 September 2025).
46. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.