STATS 402 - Interdisciplinary Data Analysis

Football AI Tracker

Final Report

Nizar Talty & Othmane Echchabi
nizar.talty@duke.edu | othmane.echchabi@duke.edu

Abstract

Football tracking has become a cornerstone of modern sports analytics, offering critical insights for match analysis, tactical enhancements, and fan engagement through dynamic visualizations. This project addresses the challenge of achieving accurate and cost-effective tracking of players, referees, and the ball using standard video footage, eliminating the reliance on expensive, specialized hardware such as tracking cameras or GPS-based systems.

We propose a robust, scalable framework that integrates state-of-the-art computer vision techniques to achieve this goal. Our methodology employs YOLOv8 for efficient and precise detection of players and referees, and ByteTrack for reliable multi-object tracking to maintain consistent identities across frames (Jocher et al., 2023; Zhang et al., 2022). Additionally, we incorporate PnLCalib, a points-and-lines calibration method, for accurate keypoint detection and homography transformations, enabling the mapping of participants onto a two-dimensional pitch representation. This integration allows for seamless handling of lower-quality video inputs while preserving spatial accuracy (Gutiérrez-Pérez & Agudo, 2024).

To validate our approach, we applied the framework to video recordings of Duke Kunshan University's Suzhou League football matches. The results demonstrated its ability to accurately track participants and the ball, even in suboptimal video conditions. The system consistently mapped movements onto a virtual pitch with high fidelity, offering insights comparable to those generated by high-end commercial systems. This framework has the potential to transform football analytics by addressing the limitations of current systems, particularly in accessibility and scalability. By eliminating the need for costly infrastructure, our approach democratizes advanced football analytics, making it accessible to a broader range of teams, leagues, and enthusiasts worldwide. This work not only advances the field of sports analytics but also paves the way for further innovations in affordable and scalable computer vision applications in sports.

Introduction

Football tracking has become an integral aspect of modern sports analytics, offering essential insights for match analysis, tactical improvements, and enhanced fan engagement through dynamic visualizations. Traditional systems, such as those using GPS trackers or high-speed cameras, have achieved impressive results but remain prohibitively expensive and inaccessible to amateur teams and smaller leagues. This disparity has fueled a growing interest in developing cost-effective, scalable solutions that leverage advancements in computer vision and deep learning.

Our project aims to address this challenge by proposing a robust framework that uses standard video footage to achieve accurate tracking of players, referees, and the ball. By integrating state-of-the-art techniques such as YOLOv8 for object detection, ByteTrack for multi-object tracking, and PnLCalib for field mapping, we demonstrate a system capable of maintaining high spatial accuracy even in low-quality video conditions. The framework eliminates reliance on expensive hardware, democratizing football analytics for a broader range of stakeholders, from youth academies to amateur clubs.

In this report, we outline our methodology, discuss related work, and present the performance of our framework, validated on real-world football matches. By addressing the limitations of existing systems, our approach offers a scalable, cost-efficient alternative poised to advance sports analytics and accessibility.

Related Work

In recent years, the application of computer vision and deep learning techniques to football video analysis has rapidly advanced, reflecting growing interest from academic researchers, commercial providers, and practitioners seeking more accessible solutions. Early work in this field typically relied on manual annotations or simple feature tracking, which proved costly and error-prone in capturing the complex behaviors of athletes and the dynamics of ball movement. As camera technology and machine learning methods matured, researchers began to develop robust frameworks that fused object detection models and sophisticated multi-object tracking algorithms to isolate and follow players and balls accurately. For example, the emergence of large-scale datasets, such as SoccerNet, facilitated the training and benchmarking of deep learning methods specifically tailored for player and ball detection in challenging, broadcast-quality footage, thus enabling holistic event recognition and action spotting (Giancola et al., 2018).

In parallel, multi-object tracking techniques like DeepSORT and ByteTrack leveraged deep appearance features and robust motion models to link detection boxes across frames, ensuring smooth and continuous identity preservation for multiple players on a congested pitch(Wojke et al., 2017; Zhang et al., 2022). Researchers then sought more complex scene understanding by extending beyond bounding-box tracking to extract body poses, relational cues, and tactical patterns, paving the way for richer analyses of player movements and team formations. Although powerful, these models still relied heavily on calibrated camera setups or high-quality broadcast footage, often leaving amateur teams and lower-tier leagues at a disadvantage due to the prohibitive costs of specialized equipment and lack of standardized camera views.

Moreover, commercial tracking systems, such as those offered by Hawk-Eye Innovations, have demonstrated near-perfect accuracy by deploying multiple high-speed cameras and controlled capture environments (Ltd.). While these systems set the gold standard for professional-level analysis, they remain costly and complex to install, effectively limiting their widespread adoption. The academic community responded to this gap by exploring more accessible frameworks that integrate advanced deep learning pipelines with robust geometric transformations. Approaches like panoramic field construction and the generation of tactical heatmaps have illustrated how high-level insights—such as passing patterns and positional advantages—may be distilled from raw video data (Lucey et al., 2013). However, many existing solutions still implicitly assume certain minimum levels of video quality or stable camera parameters, making them less suitable for resource-constrained contexts.

Our framework aims to bridge this gap by combining state-of-the-art object detection and multi-object tracking methods with an adaptive homography inference pipeline that remains effective even when confronted with low-quality, single-camera recordings. By removing the dependency on specialized, high-resolution broadcasting equipment and precise calibration protocols, our approach ensures that accurate field registration, player tracking, and game reconstruction are accessible to a wide range of stakeholders. In this way, amateur clubs, youth academies, and educational projects can benefit from advanced data-driven insights into player performance and tactical configurations without incurring the costs and logistical demands traditionally associated with professional systems. As interest in football analytics continues to expand, and as machine learning methods for object detection and camera calibration become more robust, such accessible frameworks are poised to reshape how teams and analysts at every level derive value from the abundant data hidden in everyday match footage.

Proposed Method

Overview

The proposed workflow for football tracking integrates multiple state-of-the-art computer vision methods, from data collection and model training to player tracking and field mapping. The complete pipeline is illustrated in Figure 1, which outlines the primary components of our system, including data preprocessing, object detection, team classification, homography computation, and tracking. Each step is detailed in the subsections below.

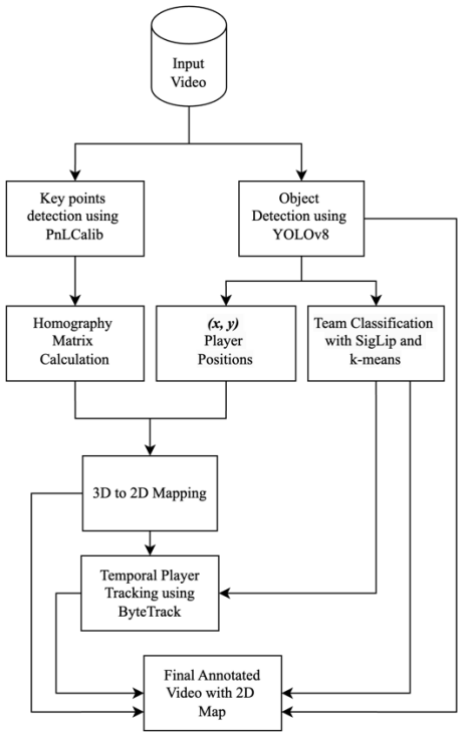

Figure 1. Project Workflow

Data Collection and Model Training

To thoroughly evaluate the proposed football player detection system, we captured custom video footage during the Suzhou League games held at Duke Kunshan University. These recordings, acquired using a Sony Alpha 7 camera, provided authentic test data, enabling a rigorous validation of our approach in real-world, non-professional sports settings. The sample frame from this footage (Figure 2) exemplifies the challenges inherent in dynamic sports environments, such as varying lighting conditions, fast player movements, and overlapping entities.

Figure 2. Sample frame from the source video of a Suzhou League game.

YOLOv8 is one of the most advanced object detection models available, achieving exceptional detection performance (Jocher et al., 2023). Therefore, for training the model, we leveraged the "Football Players Detection" dataset available on Roboflow Universe8. This curated dataset included high-quality annotations for various entities commonly observed in football matches, such as players, referees, and the ball (Figure 3). By fine-tuning YOLOv8 with this dataset, we tailored the model to excel in detecting objects specific to football gameplay (Roboflow, 2024). The dataset's rich annotations and diversity in scenarios ensured that the model could generalize effectively.

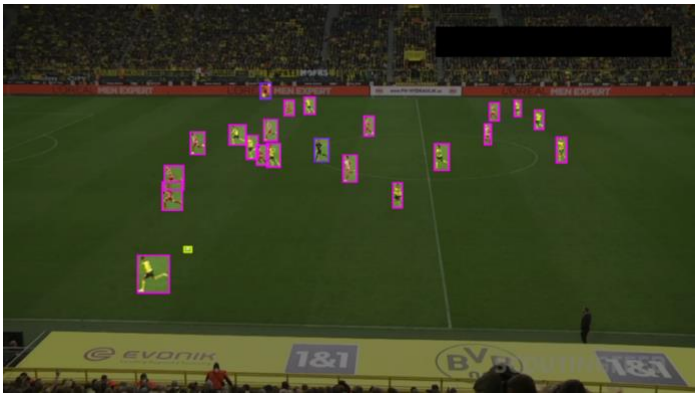

Figure 3. Sample frame from the Roboflow annotated dataset.

The YOLOv8 model was evaluated on our custom test set, and the performance metrics are presented in Table 1.

Keypoint detection

To achieve accurate field mapping, we utilized the PnLCalib zero-shot detection method, which excels in identifying keypoints and calibrating camera views without requiring prior training. PnLCalib solves the Perspective-n-Line (PnL) problem by leveraging geometric constraints to map 2D image points to 3D world coordinates. This method allowed us to project field lines onto video frames (Figure 4), ensuring precise calibration and homography computation. PnLCalib's robust approach provided accurate spatial alignment, enabling seamless mapping of detected players and the ball to the real-world field layout. Its adaptability to varying camera angles and field conditions made it a reliable choice for this project.

Figure 4. Projected Field Lines Using PnLCalib

Team Classification

After detecting players using YOLOv8, we extracted their bounding boxes and processed them for team classification. Initially, we attempted to use simple RGB-based classification by following these steps:

1. **Field Masking:** Masked the green field area by setting a range for green values and calculating the average green color on the field.
2. **Jersey Color Calculation:** Masked the bottom half of each player crop (after masking the green background) to focus on the jersey area, then calculated the mean color of the isolated jersey region.

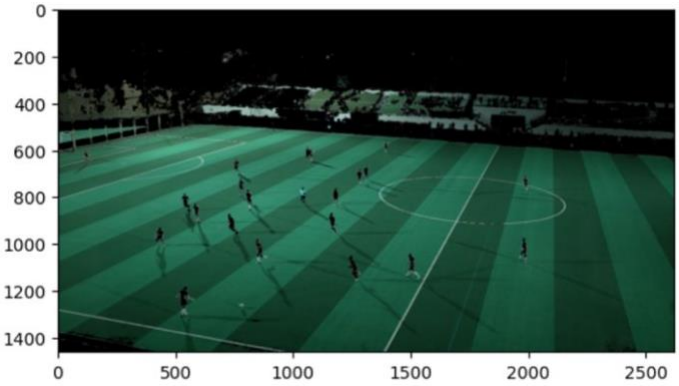

Figure 5. Masked field to isolate grass and compute reference mean color.

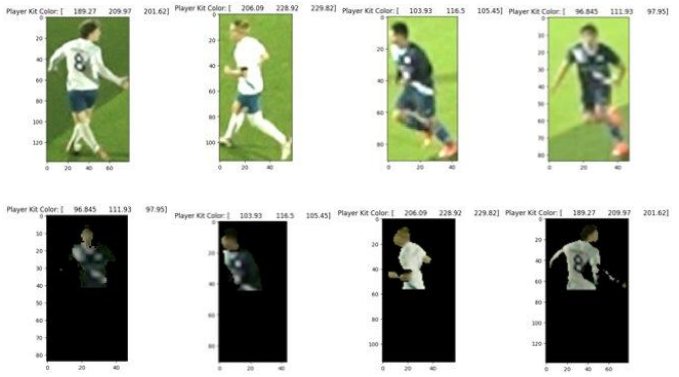

Figure 6. Results of jersey isolation before and after background masking.

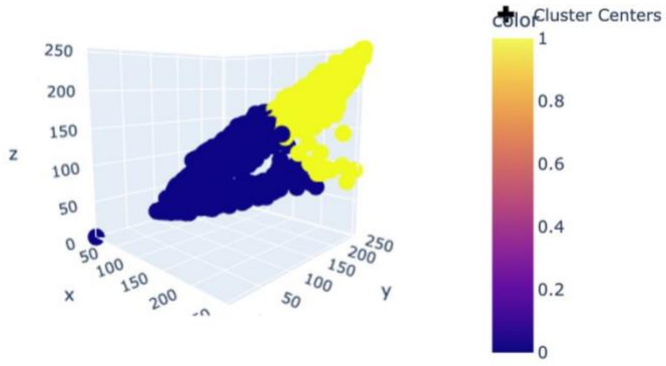

Figure 7. 3D clustering of jersey colors; centroids indicate team averages.

Figure 5 shows the masked field used to isolate the grass and computer a reference mean color, while Figure 6 illustrates the results of jersey isolations before and after background masking. However, this approach was not effective due to inconsistencies in lighting conditions across video frames, leading to unreliable color-based classification. Figure 7 displays a 3D clustering of jersey colors. The results highlight that RGB-based clustering fails to clearly separate the teams

To overcome this limitation, we employed the SigLip model to generate embeddings for each of the crops we got from the object detection model bounding boxes (Zhai et al., 2023). This way, we can extract more robust and meaningful features from the player images, beyond just the raw pixel values. The model helps to capture patterns and characteristics that are less affected by lighting variations, leading to more accurate feature representations. We then project our embeddings from N, 768) to (N, 3) using UMAP and then perform a two-cluster division using k-Means (McInnes et al., 2018).

This method allows us to group players into distinct teams based on the similarity of their extracted features, enabling more reliable team classification even in challenging visual conditions.

Figure 8. Team classification using SigLip and k-means

Figure 8 demonstrates how the generated embeddings provide better separation of players into teams, offering a more robust solution for team classification under varying environmental conditions.

Homography Computation and Mapping

Using the key points detected via PnLCalib, we computed the homography matrix to map player positions onto a two-dimensional pitch representation. This transformation enabled the creation of a top-down view of player movements, which is essential for tactical analysis and visualization. To ensure temporal consistency, we applied a moving average with a window size of 5 to smooth the homography computation.

In an ideal scenario, where a sufficient number of well-distributed key points are detected, the homography matrix provides accurate estimations of 2D positions. This is demonstrated in Figures 9 and10.a, where Figure 9 shows an example frame with bounding box centers and detected key points, and Figure 10.a illustrates the corresponding 2D mapping. However, challenges arise when the homography computation fails, leading to incorrect mappings, as shown in Figure 10.b. These failures prompted the development of an improved framework to enhance the accuracy and reliability of the homography matrix calculation.

Figure 9. Example frame with centers of bounding boxes and key points shown

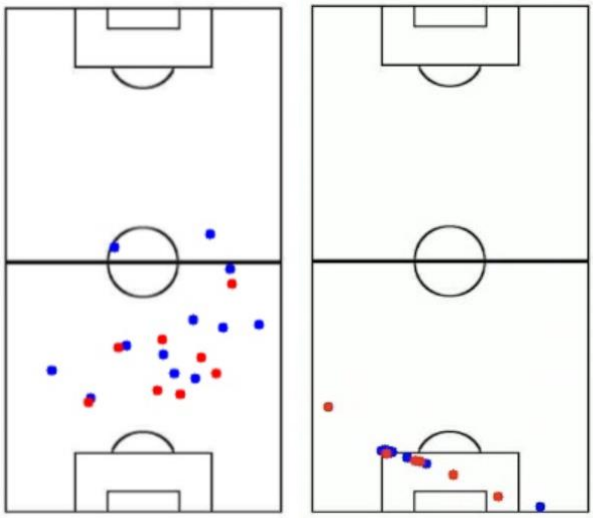

Figure 10. ***a***: Correct 2D Mapping of players, ***b***: Incorrect mapping due to errors in computing H

Several factors were identified as contributors to the inaccuracies in homography computation:

1. **Insufficient Key Points:** The homography matrix requires at least four key points for a reliable transformation. When fewer points are detected, the computation fails.
2. **Misaligned Field Key Points:** Detection errors using PnLCalib can result in inaccurately positioned field key points, significantly degrading the matrix's accuracy.
3. **Frame-to-Frame Discrepancies:** Large variations in the homography matrix between consecutive frames can lead to unstable mappings, especially in dynamic scenarios where smooth transitions are critical.

To resolve these issues, we adopted a refined routine for computing the homography matrix and projections. This method ensures stability by incorporating error detection for significant deviations and a weighted moving average for temporal smoothing. The matrix $\boldsymbol{H}$ is computed from detected keypoints, and if keypoints are insufficient, the previous matrix $\boldsymbol{H_{prev}}$ is used to maintain continuity. Player positions are transformed into homogeneous coordinates $[x \quad y \quad 1]^T$ and projected using $\boldsymbol{H}$. Deviations are calculated as:

$$\Delta = \| p_{mapped} - p_{prev\_mapped} \|$$

where p_{mapped} and p_{prev_mapped} represent points in the current and previous frames, respectively. If Δ exceeds a threshold that we setup, $\boldsymbol{H_{prev}}$ is retained. Valid matrices are stored in a buffer to compute a weighted moving average:

$$H_{\text{avg}} = \frac{\sum_{i=1}^{n} w_i \cdot H_i}{\sum_{i=1}^{n} w_i}$$

with w_i as normalized weights prioritizing recent matrices.

Tracking Players and Ball

The next step in our project involves tracking players across consecutive video frames to ensure the consistent assignment of unique identifiers (IDs). To achieve this, ByteTrack is used as the foundational tracking method (Zhang et al., 2022). Figure 11. Shows how smoothly ByteTrack tracks players across different frames assuming no occlusion. However, challenges arise in scenarios where players become occluded, temporarily leave the frame then reappear in later frames, often leading to mismatches in ID assignments. To address these challenges, we designed our own labeling methodology. This method predefines a fixed set of IDs, assigns them to all detected players in the initial frame, and systematically manages unassigned or missing IDs in subsequent frames.

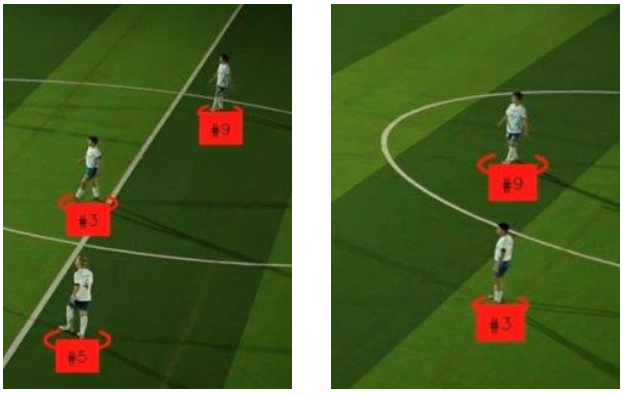

Figure 11. Results using ByteTrack for multi-object tracking.

The proposed approach is structured as follows:

1. **Distance Calculation**: Compute the distances between current player positions and two reference sets: the positions of players from the previous frame and the last known positions of unmatched players. These distances form the basis for prioritizing ID assignment.
2. **Sorting**: Rank all potential matches based on proximity, prioritizing closer matches to minimize ID mismatches and improve assignment accuracy.
3. **ID Assignment**: Assign IDs to players in the current frame by matching their positions to IDs from either the previous frame or the last known positions. This ensures no duplication of IDs and assigns any remaining IDs to unmatched players.
4. **Update Last Known Positions**: For players who remain unmatched, their last known positions are updated. IDs that have been reassigned are removed from this record to avoid conflicts in subsequent frames.

This tracking and labeling method delivers more reliable results, particularly in challenging conditions such as player occlusions or when players temporarily leave the frame. Figures 12 and 13 illustrate a comparison between our method and the baseline labeling approach, specifically showing how our method handles situations where a player exits and re-enters the frame.

Figure 12. Labeling using our method. Left: Player before leaving the frame. Right: Player after rejoining. The label is the same.

Figure 13. Labeling using ByteTrack only. Left: Player before leaving the frame. Right: Player after rejoining. The label is **not** the same.

However, it is important to note that while tracking based on 2D positions generally provides good results, it does not guarantee optimal performance in all scenarios. For instance, when multiple players disappear from the frame while remaining near each other, the tracking algorithm may struggle to distinguish between them. This can lead to ID mismatches, as the system may incorrectly assign different labels to different players or fail to track a player correctly when they reappear.

Figure 14. Example frame crops showing labeled players before their tracks are lost and interchanged

Figure 14 highlight a key issue that arises when the position of a player in the current frame is very close to the position of another player from the previous frame. In such cases, the algorithm may incorrectly assign the label of one player to another due to the proximity of their positions, leading to ID mismatches. This problem is particularly evident when players are near each other, causing the tracking system to erroneously link the wrong player in the current frame to a player in the previous frame.

Results and Performance Evaluation

The objection detection results using YOLOv8 demonstrate strong detection capabilities, with an overall mAP@50 of 88.0% and mAP@50-95 of 67.3%. Entity-specific metrics indicate exceptional precision for goalkeepers, referees, and players, with mAP@50 exceeding 94% in all cases. However, ball detection remains a challenging task, achieving a relatively lower mAP@50 of 59.9% and mAP@50-95 of 32.7%. This discrepancy underscores the complexity of detecting smaller and faster-moving objects within the scene.

Class	*mAP@50 (%)*	*mAP@50-95 (%)*
Overall	88.0	67.3
Ball	59.9	32.7
Goalkeeper	94.8	79.4
Player	99.4	86.2
Referee	97.7	71.0

Table 1. YOLOv8 Model Validation Metrics

The integration of ByteTrack provided reliable multi-object tracking, ensuring consistent identity assignments even in complex scenarios involving occlusions or players re-entering the frame. While challenges such as ID mismatches in crowded areas and proximity errors during player reappearances were noted, our custom labeling methodology mitigated many of these issues effectively. Similarly, the PnLCalib method delivered robust homography transformations, accurately mapping player positions onto a 2D pitch despite occasional failures caused by sparse or misaligned key points.

These shortcomings were addressed with a weighted moving average approach, which improved the temporal stability of projections. The amalgamation of all the modules allows us to achieve satisfactory results. The final output is an annotated video that accurately tracks the players, the ball, and the referees. Figure 15 shows an example frame with all the components applied.

Figure 15. Example frame from the output with all components implemented

Given that the video is at 50 FPS. Using two inferences on each frames makes the process computationally intensive. That is, we managed to reduce the processing time from 6 hours for a 1-minute video to approximately 2 hours. This significant improvement in efficiency was achieved by leveraging two NVIDIA A40 GPUs. One GPU is dedicated to running the YOLOv8 model for object detection, while the second GPU takes care of the PnLCalib inference for key point detection. The parallelization of these tasks across two GPUs has enabled a substantial reduction in overall processing time, allowing for faster video analysis without compromising the accuracy or quality of the results.

Conclusion and future work

In conclusion, this project has made significant progress toward developing an AI-driven football tracking system capable of delivering reliable player, referee, and ball tracking using standard video footage. Through the integration of state-of-the-art computer vision techniques, including YOLOv8 for precise detection, ByteTrack for robust tracking, and PnLCalib for keypoint detection and homography transformations, we have created a scalable and cost-effective framework. This framework has been successfully validated on video footage of Duke Kunshan University's Suzhou League football matches, demonstrating its ability to track participants and the ball accurately even under challenging conditions such as poor lighting and low detection quality.

The challenges faced during this project, particularly the effects of inconsistent lighting and detection quality, have driven us to implement targeted improvements, including fine-tuning the YOLOv8 model and enhancing team classification through more advanced feature extraction techniques. These efforts have led to improvements in player and ball detection, tracking consistency, and clustering accuracy. As we move forward, our next steps will involve developing a more robust method for tracking and labeling, addressing issues such as player occlusions and re-identification through the integration of methods like Kalman filtering. Additionally, we plan to further enhance the classification model by either testing other pretrained models to better handle the variability in player appearances. Finally, we will explore ways to make the framework faster and more efficient.

With these updates and further iterative refinements, we are confident that our system will continue to evolve into a high-performing solution for football tracking, offering valuable insights for match analysis, tactical development, and fan engagement. Ultimately, this work has the potential to democratize advanced football analytics, making it accessible to a broader range of teams, leagues, and enthusiasts, and paving the way for future innovations in affordable and scalable sports analytics.

References

Giancola, S., Amine, M., Dghaily, T., & Ghanem, B. (2018, 18-22 June 2018). SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),

Gutiérrez-Pérez, M., & Agudo, A. (2024). PnLCalib: Sports Field Registration via Points and Lines Optimization. *arXiv preprint*. https://arxiv.org/abs/2404.04244

Jocher, G., Chaurasia, A., & Qiu, J. (2023). Ultralytics YOLOv8. In.

Ltd., H.-E. I. *Hawk-Eye*. https://en.wikipedia.org/wiki/Hawk-Eye

Lucey, P., Bialkowski, A., Carr, P., Morgan, S., Matthews, I., & Sheikh, Y. (2013). Representing and Discovering Adversarial Team Behaviors using Player Roles. *2013 Ieee Conference on Computer Vision and Pattern Recognition (Cvpr)*, 2706-2713. https://doi.org/10.1109/Cvpr.2013.349

McInnes , L., Healy , J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*. https://arxiv.org/abs/1802.03426

Roboflow. (2024). football-players-detection Dataset. In *Roboflow Universe*: Roboflow.

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple Online and Realtime Tracking with a Deep Association Metric. *2017 24th Ieee International Conference on Image Processing (Icip)*, 3645-3649. <Go to ISI>://WOS:000428410703155

Zhai , X., Mustafa , B., Alexander, K., & Beyer , L. (2023). Sigmoid Loss for Language Image Pre-Training. *International Conference on Computer Vision (ICCV)*. https://arxiv.org/pdf/2303.15343

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., & Wang, X. (2022). ByteTrack: Multi-Object Tracking by Associating Every Detection Box. *Proceedings of the European Conference on Computer Vision (ECCV)*. https://arxiv.org/abs/2110.06864