INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM ENGENHARIA DE INTELIGÊNCIA ARTIFICIAL




Sistema Inteligente de Apoio à Arbitragem: uma abordagem de Visão por Computador e Aprendizagem Profunda para a deteção de faltas no futebol amador

NUNO RAFAEL DE SOUSA ALVES
outubro de 2025

P.PORTO

Intelligent Refereeing Support System: A Computer Vision and Deep Learning Approach for Foul Detection in Amateur Football

Nuno Rafael Sousa Alves
Student No.: 1200753

Dissertation for the Master's Degree in Artificial Intelligence Engineering

Supervisor: Constantino Martins
Co-Supervisor: Paulo Matos

Evaluation Committee:
President:
Diogo Emanuel Pereira Martinho, Assistant Professor, Institute of Engineering, Polytechnic of Porto

Members:
Rui Pedro Ferreira Pinto, Assistant Researcher, Faculty of Engineering, University of Porto
António Constantino Lopes Martins, Associate Professor, Institute of Engineering, Polytechnic of Porto

Porto, October 18, 2025

Abstract

Recent advancements in computer vision and artificial intelligence have revolutionized sports technology, particularly in professional football through systems like Video Assistant Referee (VAR). However, a significant technological gap exists between professional and amateur levels of the sport, with professional systems costing thousands of euros remaining inaccessible to grassroots football. This research addresses this disparity by developing an automated referee assistant system using consumer-grade smartphones.

The proposed system integrates three specialized YOLO-based models: YOLOv12 for player and ball tracking, a custom-trained YOLO11 Pose model for field keypoint detection, and another YOLO11 Pose model for player pose estimation. A key innovation is our proximity-based processing strategy that triggers pose analysis only when players are near the ball, reducing computational overhead by approximately 65% while maintaining detection accuracy. The system employs dual-camera panoramic stitching to achieve 180-degree field coverage, overcoming parallax challenges through optimized camera positioning guidelines.

Our implementation achieves 86.8% mean average precision for player detection and 99.5% for field keypoint detection, though ball detection remains challenging at 51.7% due to object size limitations. The system successfully detects handball violations outside the penalty area and ball out-of-bounds situations in real-time at 15-20 frames per second. We created a custom dataset of 500 annotated images with 27 field keypoints, addressing a critical gap in publicly available football field detection resources.

While the system faces limitations in 3D spatial analysis for airborne balls, it demonstrates that meaningful referee assistance is achievable with consumer hardware. This research contributes to democratizing sports technology by providing an accessible, cost-effective solution that brings automated officiating capabilities to amateur football, where the vast majority of matches worldwide currently lack any technological support.

Keywords: Computer Vision, Deep Learning, Football Analytics, Automated Refereeing, Privacy-Preserving AI, Real-time Object Detection

Resumo

Os recentes avanços em visão computacional e inteligência artificial revolucionaram a tecnologia desportiva, particularmente no futebol profissional através de sistemas como o Video-árbitro (VAR). No entanto, existe uma lacuna tecnológica significativa entre os níveis profissional e amador do desporto, com sistemas profissionais a custar milhares de euros a permanecerem inacessíveis ao futebol de base. Esta investigação aborda esta disparidade desenvolvendo um sistema automatizado de assistente de arbitragem através do uso de smartphones.

O sistema proposto integra três modelos especializados baseados em YOLO: YOLOv12 para deteção de jogadores e bola, um modelo YOLO11 Pose personalizado para deteção de pontos-chave do campo, e outro modelo YOLO11 Pose para estimação de pose de jogadores. Uma inovação chave é a nossa estratégia de processamento baseada em proximidade que ativa a análise de pose apenas quando os jogadores estão perto da bola, reduzindo a sobrecarga computacional em aproximadamente 65% mantendo a precisão de deteção. O sistema combina imagens de duas câmaras para criar uma vista panorâmica com cobertura de campo de 180 graus, superando desafios de paralaxe através de diretrizes otimizadas de posicionamento de câmaras.

A nossa implementação alcança 86,8% de precisão média para deteção de jogadores e 99,5% para deteção de pontos-chave do campo, embora a deteção da bola permaneça desafiante com 51,7% devido a limitações do tamanho do objeto. O sistema deteta com sucesso violações de mão na bola fora da área de penálti e situações de bola fora de campo em tempo real a 15-20 frames por segundo. Criámos um conjunto de dados personalizado de 500 imagens anotadas com 27 pontos-chave de campo, abordando uma lacuna crítica nos recursos públicos disponíveis para deteção de campos de futebol.

Embora o sistema enfrente limitações na análise espacial 3D para bolas no ar, prova que telemóveis convencionais são suficientes para assistir eficazmente a arbitragem. Esta investigação contribui para democratizar a tecnologia desportiva ao fornecer uma solução acessível e económica que traz capacidades de arbitragem automatizada ao futebol amador, onde a vasta maioria dos jogos a nível mundial atualmente carece de qualquer suporte tecnológico.

Palavras-chave: Computer Vision, Deep Learning, Football Analytics, Automated Refereeing, Privacy-Preserving AI, Real-time Object Detection

Acknowledgement

I wish to express my sincere appreciation to my supervisor, Professor Constantino Martins, and my co-supervisor, Professor Paulo Matos, for their essential guidance, encouragement, and insightful feedback throughout the course of this dissertation. Their knowledge and support were crucial for the successful completion of this work. I am equally thankful to my family and friends for their patience, understanding, and constant encouragement, which gave me the determination to overcome the most demanding moments. I would also like to give special thanks to my close friends, Bruno Pereira and Diogo Machado, whose friendship and support have been a powerful source of motivation along this path. Lastly, I am deeply grateful to my twin brother, Rafael Alves, for embarking on this ambitious project with me. Sharing this experience with him not only made the process lighter but also turned it into a journey full of joy, collaboration, and memorable moments

Contents

List of Figures

List of Tables

List of Acronyms

A2	Area Attention.
AI	Artificial Intelligence.
ANN	Artificial Neural Network.
CNN	Convolutional Neural Network.
CV	Computer Vision.
DSR	Design Science Research.
FIFA	Fédération Internationale de Football Association.
GDPR	General Data Protection Regulation.
mAP	mean Average Precision.
MPJPE	Mean Per Joint Position Error.
P	Precision.
PBDM	Player and Ball Detection Model.
PR	Precision-Recall.
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
R	Recall.
R-ELAN	Residual Efficient Layer Aggregation Networks.
R-CNN	Region-based Convolutional Neural Network.
VAR	Video Assistant Referee.
YOLO	You Only Look Once.

Chapter 1

Introduction

Modern football has undergone a significant transformation with the introduction of refereeing assistance technologies, as demonstrated by the success of Video Assistant Referee (VAR) in reducing decision errors in professional competitions (Held, Itani, et al. 2024; Holder, Ehrmann, and König 2022). However, while professional football benefits from these technological advances, the amateur level, which represents the vast majority of games played worldwide, remains solely dependent on traditional human refereeing (Gurau et al. 2023; Promvijittrakarn and Charoenpong 2023).

Recent advances in computer vision and artificial intelligence have shown significant potential to revolutionize sports analysis through automated player detection and event recognition systems (Chopra, Mundody, and Reddy Guddeti 2023; V and G 2024). These technologies, when properly adapted, can offer accessible solutions to improve the quality of refereeing at amateur levels of the sport.

This research proposes the development of an automated foul detection system based on computer vision, with the aim of democratizing access to refereeing support tools. The proposed system uses advanced deep learning techniques for real-time analysis of plays and player interactions, with a specific focus on creating a solution that is both technically robust and economically viable for implementation in amateur contexts.

1.1 Context

Football continues to evolve through technological innovation, particularly in professional leagues. The introduction of VAR marked a significant milestone in football officiating technology, fundamentally changing how crucial decisions are made in professional matches (Carlos, Ezequiel, and Anton 2019; D'Orazio and Leo 2010). However, a significant technological gap exists between professional and amateur football, where the latter represents the vast majority of matches played worldwide yet lacks access to these advancements (Held, Cioppa, et al. 2024).

Recent developments in player tracking and detection systems have shown promising results in professional settings. These systems can now effectively monitor player movements and interactions throughout matches, providing valuable data for both real-time decision support and post-match analysis (Held, Cioppa, et al. 2024). The success of these technologies in professional environments has demonstrated their potential for broader applications across different levels of the sport(Held, Cioppa, et al. 2024).

This research addresses the technological disparity between professional and amateur football by exploring how modern computer vision and artificial intelligence technologies can be adapted and implemented in amateur settings. The focus is on developing accessible solutions that can provide similar benefits to those currently available only in professional football, while considering the unique constraints and requirements of amateur environments.

1.2 Motivation

The development of computer vision systems for amateur football presents unique challenges that differ significantly from controlled laboratory environments. Real-world deployment must address varying lighting conditions, dynamic camera movements, and complex player interactions, making this an important frontier for advancing computer vision capabilities (Chopra, Mundody, and Reddy Guddeti 2023; Promvijittrakarn and Charoenpong 2023; Shao and He 2025; V and G 2024). These technical challenges provide an opportunity to contribute to both theoretical understanding and practical applications of computer vision in unpredictable environments.

The need for accessible technological solutions in amateur football is becoming increasingly apparent. While professional football benefits from sophisticated officiating tools, the amateur level faces significant barriers to accessing similar technologies. The development of cost-effective solutions could democratize access to fair play tools, addressing a crucial gap in sports technology (Diop et al. 2022; Öberg 2021). This economic motivation drives the search for innovative solutions that balance functionality with affordability.

Beyond the technical aspects, improving amateur football officiating has broader societal implications. Enhanced fairness and reduced conflicts in amateur matches can lead to improved player experiences and stronger community engagement in sports (Joshan Athanesious and Kiruthika 2024). The implementation of objective decision-support systems can help reduce tensions arising from disputed calls and enhance the overall quality of amateur competitions.

1.3 Objectives

The primary aim of this research is to develop a robust computer vision system capable of functioning as an automated referee assistant in football matches. While VAR systems have proven valuable in professional settings, their high cost and complexity make them impractical for an amateur environment. Our goal is to create a more accessible alternative that leverages recent advances in computer vision and artificial intelligence to provide reliable referee assistance in amateur football contexts.

The central hypothesis of this research:

> **Can a computer vision-based system effectively detect and analyze football fouls in real-time with sufficient accuracy and reliability to serve as an automated referee assistant in amateur matches, providing a cost-effective alternative to professional VAR systems while maintaining high standards of officiating?**

This builds upon recent success in player tracking and event detection systems, extending these capabilities to the specific challenge of foul detection and analysis.

To address this hypothesis, we establish three specific objectives. First, we aim to develop a computer vision system capable of accurate player tracking and motion analysis in real-time. This foundation is essential for understanding player interactions and potential foul situations, requiring robust detection algorithms that can handle the dynamic nature of football matches.

Our second objective focuses on developing and implementing foul detection algorithms. This involves creating a system that can analyze player interactions, detect potential fouls based on defined rules and patterns, and provide instant feedback. The system must process this information quickly enough to be useful during live matches, while maintaining high accuracy to ensure reliable assistance.

Finally, we aim to create practical implementation guidelines that address the specific needs of amateur football environments. This includes determining optimal camera positioning, establishing calibration procedures, and designing an intuitive interface that allows players to effectively utilize the system's capabilities during games.

1.4 Contributions

This research aims to make significant contributions to the field of sports technology, particularly focusing on making advanced refereeing assistance accessible to non-professional football. Building upon recent advances in object detection technologies, such as You Only Look Once (YOLO) and Faster Region-based Convolutional Neural Network (R-CNN) (Aleza and Vetrithangam 2023; Benakesh and Rajeev 2024), and recent developments in player tracking systems (Grotenhuis 2022; Song 2022), our work extends beyond current technical boundaries to create practical solutions for diverse football environments.

The primary technical contribution of this research lies in developing a comprehensive computer vision system specifically designed for non-professional football settings. By addressing the unique challenges of real-time foul detection in uncontrolled environments, this work advances the current state of computer vision applications in sports. Our approach focuses on creating robust, reliable solutions that can operate effectively with minimal infrastructure, making advanced refereeing technology accessible to matches at various levels. From a practical perspective, this research will provide non-professional football with tools previously available only in professional settings. The developed system will assist players in making more accurate decisions in self-officiated matches. This support system aims not to replace human officials but to provide additional tools for matches without referees, thereby improving the quality and fairness of officiating.

The broader impact of this research extends beyond technical innovation. By democratizing access to advanced refereeing technology, this work has the potential to transform how football is officiated across different organizational levels. Various types of matches will benefit from improved analysis capabilities, while the technology can assist in resolving contentious decisions. This technology can also serve as a training tool, providing concrete examples and immediate feedback during matches.

Furthermore, this research establishes a framework for implementing sophisticated computer vision systems in resource-constrained environments. The methodologies and guidelines developed through this work will serve as a foundation for future developments in sports technology, particularly in making advanced technical solutions accessible to football at all organizational levels. This contribution is particularly significant as it addresses the growing technological divide between professional and non-professional football.

In summary, while the technical innovations in computer vision and automated foul detection form the core of this research, its true value lies in its potential to democratize access to advanced sports technology. By providing non-professional football with tools that enhance the quality and fairness of matches, this work contributes to the broader goal of making sports technology accessible to all levels of organization, ultimately benefiting players and the sport as a whole.

1.5 Research Methodology

The process to find a suitable research methodology involved an initial search, which led to choosing Design Science Research (DSR) methodology (Dresch, Lacerda, and Antunes 2015). This methodology was selected due to its focus on creating innovative technological solutions to solve real-world problems, which aligns perfectly with the goal of developing a recommendation system to enhance youth football athletes' development.

DSR is a research methodology that focuses on the creation of innovative artifacts to solve complex problems. This process was split into six activities in (Brocke, Hevner, and Maedche 2020), which are problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and lastly, communication. The process is displayed in 1.1.

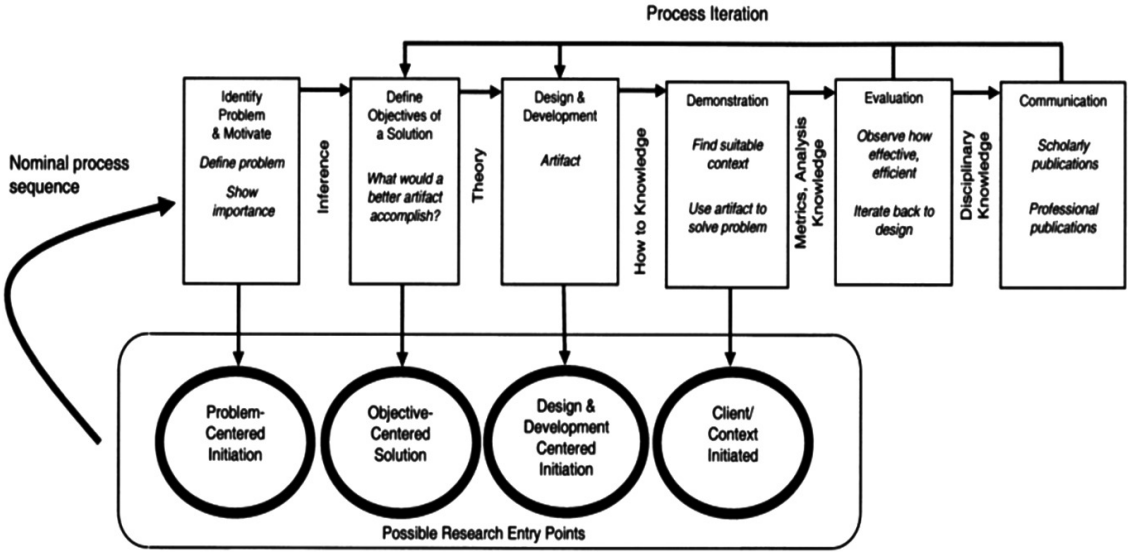

Figure 1.1: Design Science Research (Brocke, Hevner, and Maedche 2020)

For this thesis, only one iteration of the DSR methodology took place, where the solution was implemented and evaluated, leading to conclusions on its effectiveness and the discussion of necessary changes for future iterations of the system. The process involved the following stages:

1. **Problem identification and motivation:** This stage involved defining the problem of lack of accessible automated refereeing assistance for non-professional football matches, and establishing the motivation for developing a computer vision-based rule violation detection system. Through analysis of amateur football environments, it became clear that most matches lack proper officiating support due to cost and availability constraints.

2. **Objective definition:** This stage involved specifying the goals of the research, focusing on developing an automated referee assistant system capable of detecting handball violations outside the penalty area and ball out-of-bounds situations in real-time. The main objective was to create a system that operates with consumer-grade hardware and minimal infrastructure requirements.

3. **Design and development:** During this stage, the automated referee assistant system was designed and developed, incorporating three specialized YOLO models: YOLO12 extra large version for player and ball detection, custom-trained YOLO11 Pose extra large version for field keypoint detection, and YOLO11 Pose large version for player pose estimation. The system architecture includes video processing, multi-camera stitching, proximity-based filtering, and rule violation assessment algorithms.

4. **Demonstration:** The system was tested in real amateur football environments through experimentation with multiple camera configurations and diverse match conditions.

Testing encompassed single camera, dual camera stitched setups across different lighting conditions and field types to validate system performance and reliability.

5. **Evaluation:** The system's performance was evaluated through standard computer vision metrics including mean Average Precision (mAP), precision, recall, and F1-scores for each detection component. Real-world validation assessed system accuracy, processing efficiency, false positive rates, and practical deployment considerations across extended testing periods.
6. **Communication:** The final stage involved documenting and sharing the research findings through this thesis, focusing on transparency of both technical achievements and limitations discovered during development and real-world testing phases.

Additionally, during the development phase of DSR, the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology was used to conduct a systematic literature review on recommendation systems in football. This review provided an insightful overview on the existing systems, the user modeling techniques and the most common algorithms used.

The methodology chosen proved adequate for this research, as it provided a systematic framework for developing and evaluating the recommendation system. The iterative nature of DSR allowed for continuous refinement of the solution.

1.6 Ethics and Data Protection

Since the system involves video recording of amateur football matches, which may include minors and recreational players, ensuring data protection and privacy is crucial (Pandey and Mishra 2024). The research strictly adheres to the General Data Protection Regulation (GDPR)and relevant local data protection laws.

The core GDPR principles followed in this research include the right to erasure, right of access to personal data, right to restriction of processing, right to rectification, right to object, and right to data portability. Additionally, we maintain transparency through notification obligations regarding any rectification or erasure of personal data.

All video data collected during matches undergoes automatic anonymization through computer vision techniques (Sepehri et al. 2023). This includes blurring of faces and any other identifying features that are not essential for foul detection analysis. The system is designed to focus solely on detecting potential fouls while minimizing the capture of personal identifying information.

1.7 Document Structure

This dissertation is organized into five chapters that present the research, development, and evaluation of an automated referee assistant system for sports rule violation detection. The structure is as follows:

Chapter 1 introduces the research context, motivation, and objectives. It presents the research methodology used and addresses important ethical and data protection considerations for implementing Artificial Intelligence (AI) systems in sports environments.

Chapter 2 provides a comprehensive review of the state of the art in computer vision applications for sports analysis. It examines the evolution of these technologies, current approaches to player detection and tracking, and presents a systematic literature review that identifies key research gaps and opportunities.

Chapter 3 details the system design and implementation of the automated referee assistant. It describes the technical architecture, including the computer vision models for player detection, field analysis, and pose estimation, as well as the video processing pipeline and rule violation assessment algorithms.

Chapter 4 presents the experimental results and performance evaluation. It provides comprehensive analysis of each system component's accuracy, processing efficiency, and real-world validation results, demonstrating the system's capabilities and identifying areas for improvement.

Chapter 5 concludes the dissertation by summarizing the key findings, discussing limitations, and suggesting directions for future research. It reflects on the broader implications of this work for sports technology and the potential for extending the system to additional sports and rule types.

Chapter 2

State of the Art

The field of Computer Vision (CV) in football analysis has undergone significant transformations over the past decade, driven by advancements in artificial intelligence and the increasing demand for automated decision support systems in sports (Bi et al. 2023). This evolution reflects the sport's growing complexity and the need for more sophisticated tools to assist in various aspects of the game, from tactical analysis to refereeing decisions.

The integration of CV systems in football has become particularly relevant as the sport faces increasing pressure for accuracy and fairness in decision-making. The VAR system, introduced by Fédération Internationale de Football Association (FIFA), represents a significant milestone in this evolution, demonstrating both the potential and limitations of technology in football (Albzeirat et al. 2020). However, while VAR has improved decision accuracy to 99.3% in professional competitions, the system still relies heavily on human interpretation and can lead to significant game interruptions (Xu et al. 2021).

Recent developments in deep learning and CV have opened new possibilities for more sophisticated and automated approaches to football analysis. Modern systems can now handle complex tasks such as player tracking, ball detection, and event recognition with increasing accuracy (Rashid and Liew 2022). These advances are particularly significant in addressing traditional challenges such as occlusion, rapid movement tracking, and real-time processing requirements (Nguyen and Tran 2023).

The application of artificial intelligence in football spans multiple domains, including automated referee assistance, tactical analysis, and performance evaluation. The development of these systems requires addressing various technical challenges, from the processing of multi-camera inputs to the real-time analysis of complex game situations (Theagarajan and Bhanu 2021). The integration of various technologies, including Convolutional Neural Network (CNN), YOLO architectures, and transformer-based models, has enabled more robust and accurate systems (Aleza and Vetrithangam 2023).

In professional environments, these technologies are increasingly being used to support decision-making processes and provide detailed analysis of game events. However, there remains a significant gap between professional and amateur levels in terms of access to and implementation of these technologies (Theagarajan and Bhanu 2021). This disparity has motivated research into more accessible and efficient solutions that can be deployed across different levels of the sport.

This chapter provides a comprehensive review of the current state of CV technologies in football, examining both the theoretical foundations and practical implementations. The review begins with an analysis of the evolution of CV in football, followed by a detailed

examination of specific technologies and approaches, and concludes with a systematic review of current research in the field.

2.1 Evolution of Computer Vision in Football

The evolution of computer vision in football represents a significant technological advancement in sports analysis and decision-making support. This section explores the historical development, traditional approaches, and current technological landscape of computer vision applications in football.

2.1.1 Historical Development of Game Analysis Technologies

The early stages of football analysis relied heavily on manual observation and basic video recording techniques. Traditional methods involved human observers manually coding match events and player movements, which was both time-consuming and prone to human error (Theagarajan and Bhanu 2021). The introduction of basic computer vision systems in the late 1990s and early 2000s marked the beginning of automated analysis in football.

Early computer vision systems focused primarily on basic player tracking and ball detection using traditional image processing techniques (Di Salvo Valter and Marco 2006). These systems typically relied on fixed camera positions and controlled environments, which limited their practical application in real match scenarios. The initial implementations faced significant challenges in handling dynamic environments, varying lighting conditions, and player occlusions.

The advent of digital broadcasting and multiple camera setups in professional football provided new opportunities for more sophisticated analysis systems. This period saw the development of semi-automated tracking systems that could provide basic positional data and simple event detection capabilities (Joshan Athanesious and Kiruthika 2024).

2.1.2 Traditional Approaches vs Deep Learning Solutions

Traditional computer vision approaches in football analysis relied primarily on classical image processing techniques such as background subtraction, color-based segmentation, and motion tracking. These methods, while foundational, had significant limitations in handling complex scenarios such as player occlusions, rapid movements, and varying environmental conditions (Rashid and Liew 2022).

The introduction of deep learning techniques marked a paradigm shift in football video analysis. CNNs and their variants have demonstrated superior performance in player detection and tracking, even under challenging conditions such as occlusions and varying weather conditions (Xu et al. 2021). This transition from traditional computer vision to deep learning-based approaches has significantly improved the accuracy and reliability of automated analysis systems.

Key advancements include:

- Enhanced player detection and tracking capabilities
- Improved ball tracking in complex scenarios
- More accurate event detection and classification

- Better handling of occlusions and environmental variations

2.1.3 Current Technological Landscape

The current landscape of computer vision in football incorporates multiple technologies and approaches, creating a rich ecosystem of analysis tools. Modern systems typically integrate several key components (Aleza and Vetrithangam 2023):

- VAR systems
- Goal-line technology
- Player tracking systems
- Ball tracking systems
- Event detection and classification systems

These systems increasingly rely on artificial intelligence and machine learning techniques, particularly deep learning architectures. The integration of YOLO architectures and other real-time object detection systems has enabled faster and more accurate analysis (Nguyen and Tran 2023). Additionally, transformer-based architectures are emerging as powerful tools for understanding complex game scenarios and temporal relationships in match events.

Recent developments have also seen the introduction of multi-modal analysis systems that combine video data with other sources such as sensor data and statistical information. This integration provides a more comprehensive understanding of game events and player performance (Theagarajan and Bhanu 2021). However, while these technologies have become standard in professional football, their implementation at lower levels remains challenging due to cost and infrastructure requirements.

2.2 YOLO

YOLO is a family of one-stage object detectors that reframes detection as a single regression problem, predicting bounding boxes and class probabilities in one forward pass. Compared with two-stage pipelines (e.g., R-CNN/Faster R-CNN), this design drastically reduces latency while preserving competitive accuracy, which is crucial for live sports scenarios such as player/ball tracking and referee assistance (Jegham et al. 2025; Sapkota, Flores-Calero, Qureshi, et al. 2025).

Over the last decade, the YOLO series has evolved through successive architectural and training innovations, consistently improving the speed–accuracy–efficiency trade-off. Early versions (v1–v4) established the paradigm; mid-generation models (v5–v8) broadened usability (anchor-free heads, decoupled classification/regression) and task coverage; the recent v9–v12 line emphasizes efficiency, scalability, and robust feature aggregation for real-time deployment (Jegham et al. 2025; Sapkota, Flores-Calero, Qureshi, et al. 2025).

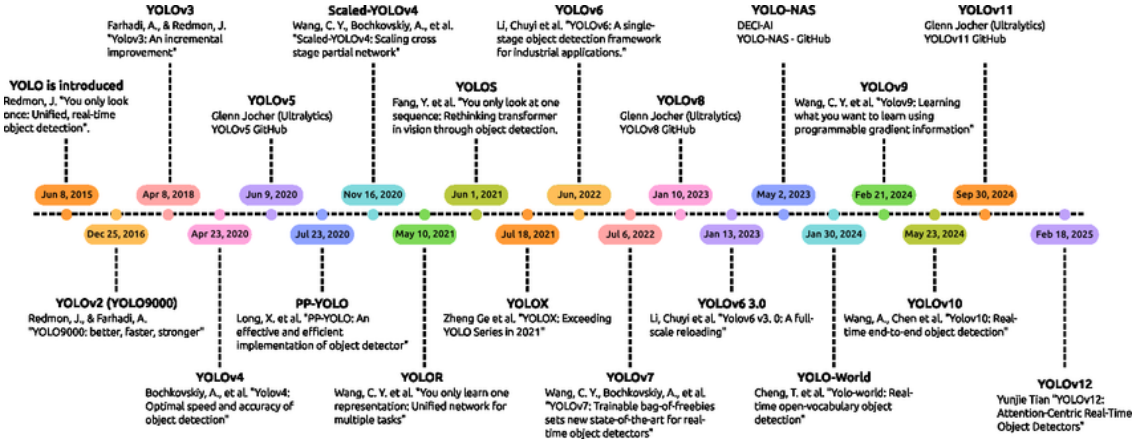

Figure 2.1: Evolution of YOLO Algorithms throughout the years Jegham et al. 2025.

Ultralytics' YOLOv11 refines the backbone and attention design to balance accuracy and throughput. Two notable elements are the *C3k2* backbone block (replacing earlier C2f/C3 variants) and the *C2PSA* (Parallel Spatial Attention) module in the neck, which together improve spatial awareness, particularly for small or partially occluded instances. The detection head uses a lightweight hybrid-convolution strategy to reduce parameters and FLOPs with minimal impact on precision, enabling deployment from edge devices to cloud environments Jegham et al. 2025; Ultralytics 2024.

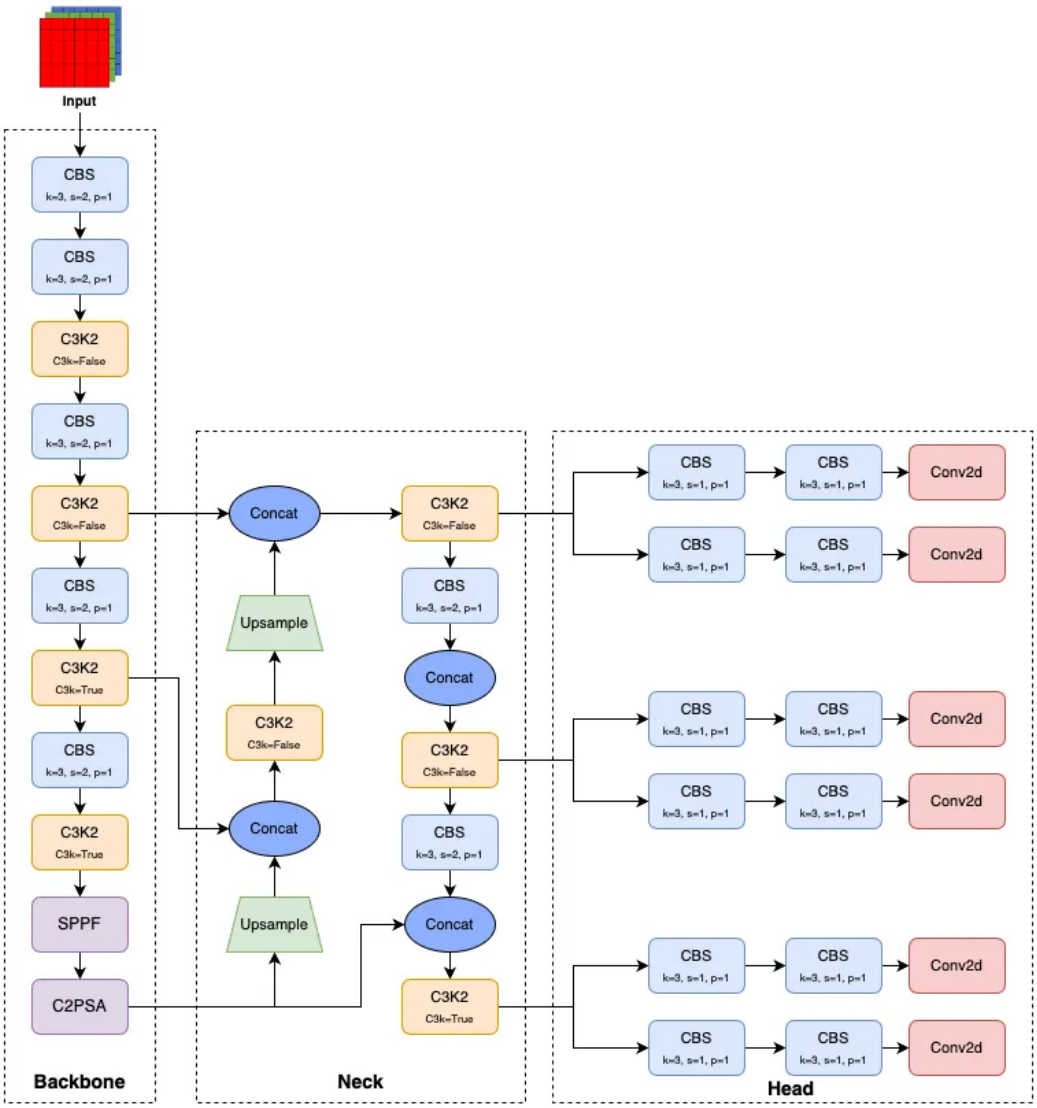

Figure 2.2: YOLO11 architecture showcasing the new C3k2 blocks and the C2PSA module Jegham et al. 2025.

The latest iteration shifts further toward attention-centric design. YOLOv12 integrates *Area Attention (A2)* for efficient large receptive fields and *Residual Efficient Layer Aggregation Net-works (R-ELAN)* to stabilize optimization in deeper models. Together with streamlined heads and training refinements, these changes improve the latency, mAP curve and object contour quality while keeping real-time performance, strengthening suitability for time-critical sports analytics Jegham et al. 2025; Sapkota, Flores-Calero, Qureshi, et al. 2025; YOLO 2025.

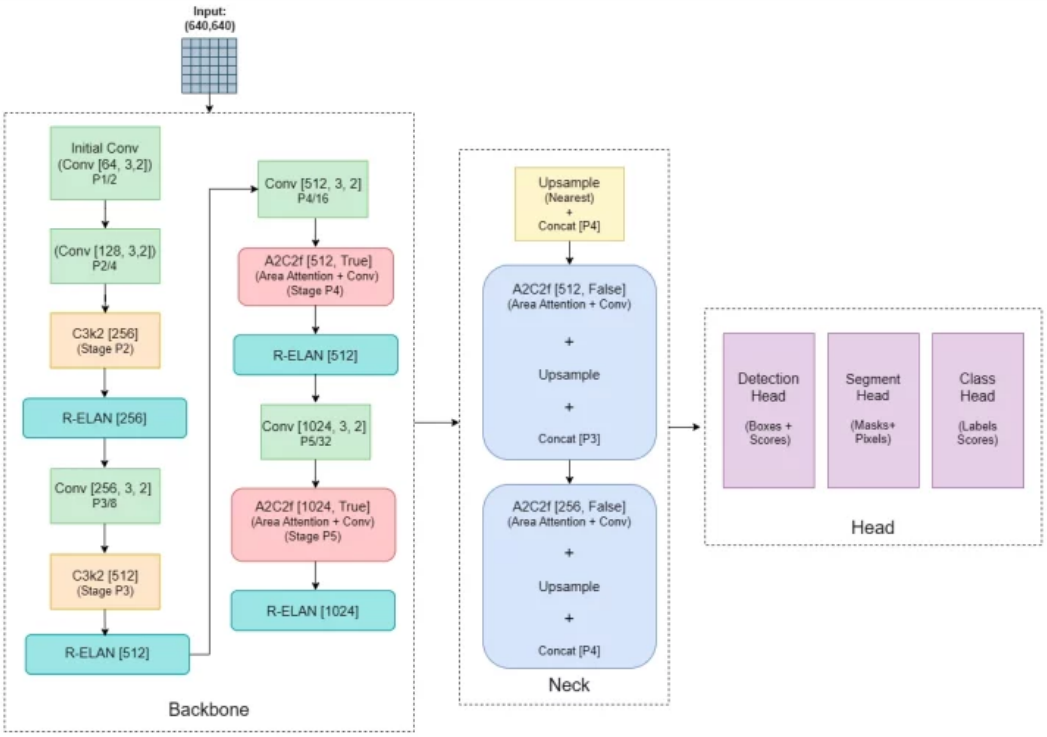

Figure 2.3: Yolov12 architecture with the new R-ELAN and A2 module Jegham et al. 2025.

Beyond bounding boxes, *YOLO-Pose* augments the detector with a keypoint branch that directly regresses human skeletal joints. This unifies detection and pose estimation in a single forward pass, preserving YOLO's real-time behavior. Recent studies report robust performance under crowded scenes, occlusions, and complex backgrounds conditions typical in football while remaining lightweight enough for mobile or drone-class capture Ding et al. 2024; Maji et al. 2022. In our thesis context, YOLO-Pose enables per-player kinematic cues (e.g., limbs, posture) for downstream logic such as potential foul/handball indicators and player state analytics.

Figure 2.4: An example of pose estimation on a construction site using YOLO11 Ultralytics 2024.

2.3 Convolutional Neural Networks

Convolutional Neural Networks (CNN) are neural networks tailored for images and other grid-like data. In addition to fully connected layers, CNNs introduce *convolutional* and *pooling* layers that share weights across spatial locations and exploit local connectivity, which drastically reduces parameters and computation compared to a plain Artificial Neural Network (ANN) that connects every pixel to every hidden neuron O'Shea and Nash 2015.

Operationally, small, learnable filters (kernels) slide over the input to produce feature maps: early layers capture edges and simple textures, while deeper layers aggregate them into higher-level patterns. Pooling (or strided convolutions) downsample feature maps to control dimensionality and increase the effective receptive field. A final fully connected (or lightweight detection) head maps features to task outputs (e.g., class scores for classification; boxes and classes for detection) O'Shea and Nash 2015.The R-CNN family are *two-stage* detectors that first generate region proposals and then classify/refine them with CNN features: R-CNN Girshick et al. 2014, Fast R-CNN Girshick 2015, and Faster R-CNN with a Region Proposal Network Ren et al. 2015. They are often strong in accuracy (e.g., for small or crowded objects) but typically incur higher latency than *one-stage* methods, making them less suitable for strict real-time constraints.

This brief context clarifies why CNN underpin modern vision models: they are parameter-efficient, translation-equivariant feature extractors that can be trained end-to-end on image data. In our work, we build on one-stage detectors (YOLO family) that inherit these CNN principles while optimizing for real-time performance.

2.4 Systematic Literature Review

A systematic literature review provides a comprehensive and structured approach to analyzing and synthesizing existing research. To ensure rigor and transparency in our review process, we followed the PRISMA methodology (Page, McKenzie, et al. 2021). This methodology offers a standardized framework for conducting systematic reviews, helping to minimize bias and ensure reproducibility in research synthesis.

The PRISMA framework guided our review through several key phases. First, we conducted a comprehensive search across multiple academic databases including IEEE Xplore, ArXiv, and Springer (Vassar et al. 2017). This was followed by a systematic screening process, where papers were evaluated against predefined inclusion and exclusion criteria. The screening process occurred in two stages: initial screening based on titles and abstracts, followed by full-text review of potentially eligible papers.

To ensure comprehensive coverage while maintaining focus on relevant research, we developed a structured search strategy combining three key domains: computer vision technologies, football-specific applications, and event detection approaches (Nasser et al. 2012). This strategy allowed us to capture the intersection of these fields while excluding irrelevant studies. The search strategy was developed following established guidelines for systematic reviews in computer science and sports technology domains.

The review process was documented using the PRISMA flow diagram, which provides a transparent representation of the study selection process (Page, Moher, et al. 2021). This visual documentation includes the number of studies identified, screened, assessed for eligibility, and ultimately included in the review. The flow diagram ensures transparency and reproducibility by clearly showing how the final set of included studies was determined, allowing other researchers to understand and potentially replicate our selection process.

Our systematic review approach emphasizes methodological rigor and transparency throughout all stages of the review process (Okoli and Schabram 2015). This structured approach helps ensure that our findings are comprehensive, unbiased, and valuable for both researchers and practitioners in the field of computer vision applications in football.

2.5 Research Questions

To conduct this systematic literature review, we defined clear research questions to answer our main question

> "Can a computer vision-based system effectively detect and analyze football fouls in real-time with sufficient accuracy and reliability to serve as an automated referee assistant in amateur matches, providing a cost-effective alternative to professional VAR systems while maintaining high standards of officiating?"

databases to be searched, search terms, and inclusion/exclusion criteria. Table 2.1 presents the core research questions that guide our review:

These research questions were carefully formulated to address the key aspects of CV applications in football refereeing, focusing on implementation techniques, performance constraints, and validation methodologies.

Search Strategy

Table 2.1: Research Questions

Identifier	Research Question
RQ1	What computer vision approaches are currently used for event detection in football environments?
RQ2	How are existing systems handling real-time processing and resource constraints in football applications?
RQ3	What methods are being used for system validation in real-world football environments?

Our search strategy combines three key domains: Computer Vision technologies, football-specific applications, and event detection approaches. Table 2.2 outlines the search strings used for each domain.

Table 2.2: Search Strings by Domain

Domain	Search Terms
CV	"computer vision" OR "machine learning" OR "deep learning" OR "artificial intelligence"
Event Detection	"event detection" OR "action recognition" OR "foul detection" OR "player tracking" OR "motion analysis" OR "tracking" OR "player" OR "game analysis" OR "sport analysis" OR "referee" OR "official" OR "VAR" OR "Video Assistant Referee" OR "refereeing"
Football	"football" OR "soccer"

Selection Criteria

To ensure the quality and relevance of selected studies, we defined specific inclusion and exclusion criteria as shown in Tables 2.3 and 2.4.

Table 2.3: Inclusion criteria

Identifier	Inclusion Criteria
IC1	Studies focused on computer vision in football/soccer environments
IC2	Articles presenting practical implementations with clear methodology
IC3	Research involving real-time processing capabilities
IC4	Studies with experimental validation and results

Study Selection Process

Following the PRISMA methodology (Page, Moher, et al. 2021), we conducted a comprehensive literature search and screening process across multiple academic databases. The initial search was performed across three major databases: IEEE Xplore, ArXiv, and Springer. Our search strategy utilized carefully constructed search strings combining terms related to computer vision, artificial intelligence, and football refereeing.

Table 2.4: Exclusion Criteria

Identifier	Exclusion Criteria
EC1	Studies published more than 5 years ago
EC2	Papers without experimental validation
EC3	Sources not in English
EC4	Theoretical proposals without implementation

The initial database search yielded a total of 217 records: 97 from IEEE Xplore, 66 from ArXiv, and 54 from Springer. After collecting these records, we first removed 25 duplicate entries, resulting in 192 unique papers for initial screening. This deduplication process was essential to ensure each study was evaluated only once and to maintain the integrity of our review process.

The screening process was conducted in two phases. In the first phase, we reviewed titles and abstracts of all 192 papers, excluding 80 papers that did not align with our research focus. This initial screening was based on our predefined inclusion and exclusion criteria (Tables 2.3 and 2.4), particularly focusing on relevance to computer vision applications in football and referee decision support systems.

The remaining 112 papers underwent a full-text review in the second phase of screening. During this detailed examination, we excluded an additional 90 papers based on our strict inclusion and exclusion criteria. Common reasons for exclusion at this stage included:

- Lack of experimental validation
- Insufficient technical detail
- Focus on aspects of sports analytics not directly relevant to our research objectives

This selection process ultimately yielded 18 papers that fully met our criteria and form the core of our systematic review. These final selections represent the most relevant and high-quality research in computer vision applications for football refereeing, encompassing various aspects such as player tracking, event detection, and referee decision support systems.

The selected papers were then categorized into three main themes for detailed analysis:

- Player tracking and analysis (7 papers)
- Field keypoints detection (2 papers)
- Game foul detection (6 papers)
- Complementary contributions (3 papers)

This categorization, presented in Tables 2.5, 2.6, and 2.7, enables a structured approach to understanding the current state of research in each key area while maintaining focus on our primary research objectives.

The distribution of papers across these categories reflects the current research landscape in computer vision applications for football, with a particular emphasis on game analysis and forecasting techniques. This distribution aligns with our research objectives of developing comprehensive solutions for automated referee assistance systems, as it provides insights

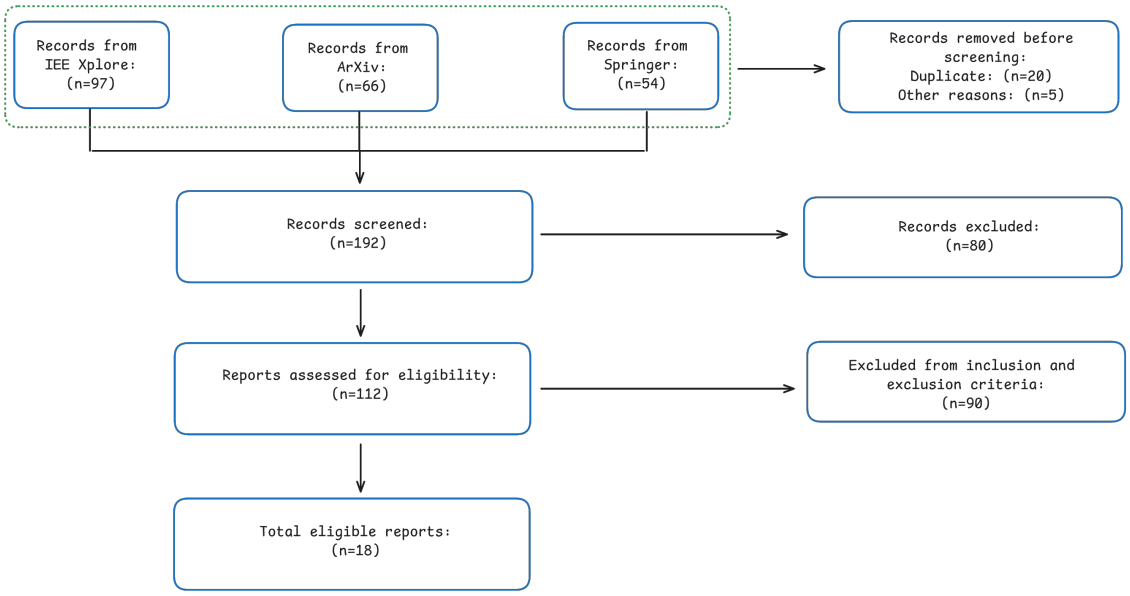

Figure 2.5: PRISMA Flow Diagram of Study Selection Process

into both the technical foundations and practical applications of computer vision in football environments.

Selected Studies Analysis

In this section, we review the selected studies and highlight how each of them contributes to the understanding of football analysis. We also point out which approaches may or may not be suitable for our target scenario of amateur and recreational football, where recording conditions are limited to single-camera setups and mobile devices.

Player and Ball Tracking: One of the essential components for any computer vision system in football is the ability to reliably locate both the ball and the players on the field. Accurate tracking provides the foundation for higher-level tasks such as event detection, tactical analysis, or even referee assistance. However, in amateur contexts the challenge becomes harder, since videos are often recorded with low-cost mobile cameras, limited viewpoints, motion blur, and frequent occlusions. For this reason, it is important to evaluate which of the existing tracking approaches can realistically be applied outside professional broadcast environments.

A common finding across the selected studies is that while they provide strong results in terms of accuracy, they often overlook the constraints of real-time processing and the recording conditions that are typical of amateur football. For instance, (Mavrogiannis and Maglogiannis 2022) and (Joshan Athanesious and Kiruthika 2024) rely on YOLO-based detection and post-processing methods such as ensemble fusion to improve tracking robustness. These approaches achieve high performance on broadcast-quality footage, but they are computationally heavy and depend on visible field markings, making them unsuitable for single-camera recordings on mobile devices where the field is not always well defined.

In contrast, (Vidal-Codina et al. 2022) bypasses the vision problem by consuming pre-processed tracking data sampled at 25Hz from professional environments. Although this allows the authors to propose effective possession and event detection rules, the method cannot be transferred directly to our context because such high-quality tracking feeds are not

Table 2.5: Studies on Player and Ball Tracking

Title	Year	Technique
The Physical Fitness Video Tracking System of Football Players Based on Artificial Intelligence Algorithm	2022	Computer vision tracking system for players using AI-based models
Soccer Player Recognition using Artificial Intelligence and Computer Vision	2022	Player recognition via CNNs, facial recognition and jersey number detection
Footballer Detection on Position-Based Classification Recognition Using Deep Learning Approach	2021	Deep learning-based detection and position classification of players
Analyzing the Feature Extraction of Football Player's Offense Action using Machine Vision	2023	Motion and ball tracking combined with feature extraction using machine vision
Amateur Football Analytics using Computer Vision	2022	YOLO / SSD-based player and ball detection, tracking with optical flow and keypoints
Perspective Transform Based YOLO with Weighted Intersect Fusion for Forecasting the Possession Sequence of the Live Football Game	2024	YOLO-based detection with centroid/GM tracking and homography projection
Automatic Event Detection in Football using Tracking Data	2022	Player and ball trajectory tracking (25Hz positional data) for event spotting

available in amateur scenarios. A similar issue arises in (V and G 2024) and (Benakesh and Rajeev 2024), which outline pipelines that integrate deep learning and multimodal modules but implicitly assume access to GPU resources and clean video input.

Other studies, such as (Wang and Liu 2023), (Rashid and Liew 2022), (Zhang 2022), and (Diop et al. 2022), place emphasis on feature extraction, positional classification, or fitness-oriented tracking. Although these works demonstrate the importance of linking player trajectories to higher-level analysis, they do not directly address the challenges of ball tracking or the need for lightweight solutions. Consequently, they provide useful conceptual insights, but fall short of offering practical methods that can run on resource-constrained devices in real time.

In summary, most existing approaches either assume broadcast-quality footage with stable field conditions or rely on computationally intensive models that cannot be applied in real-time on mobile devices. For our case, where we target amateur football recorded with low-cost cameras, the feasible direction is to adopt lightweight detectors combined with simple tracking strategies.

Field keypoints: Another crucial component for football video analysis is the estimation of field keypoints, which enables camera calibration and pitch registration. By detecting field lines and markings, these methods attempt to align the video frames with a standard pitch model, providing spatial context for higher-level tasks such as event detection and tactical analysis. In professional broadcasts, this process benefits from consistent top-view angles, high-quality footage, and clearly visible white lines. However, in amateur and recreational scenarios recorded with mobile devices, field conditions are highly variable: grass is not always present, lines are often faint or missing, and the camera angle is limited to a single

lateral view. As a result, techniques that strongly depend on well-defined field geometry or broadcast-style footage may not transfer directly to our case.

Table 2.6: Studies on Field Keypoints and Pitch Calibration

Title	Year	Technique
Amateur Football Analytics using Computer Vision	2022	Camera pose estimation and field registration using synthetic data and line detection
Perspective Transform Based YOLO With Weighted Intersect Fusion for Forecasting the Possession Sequence of the Live Football Game	2024	YOLO-based detection combined with homography matrix estimation from pitch markings

The selected studies that focus on pitch calibration demonstrate the importance of mapping the video frames to a standard football field model, but they also highlight limitations when transferred to amateur contexts. In (Mavrogiannis and Maglogiannis 2022), the authors employ camera pose estimation and field registration techniques, supported by synthetic data and line detection, to accurately align the video with the pitch. While effective in controlled broadcast settings, this method assumes that field lines are always well defined and that the surface has a uniform grass appearance. These assumptions are not guaranteed in amateur environments, where fields may be poorly marked or even non-grass surfaces.

Similarly, (Joshan Athanesious and Kiruthika 2024) use a YOLO-based detector combined with homography matrix estimation derived from pitch markings to forecast ball possession sequences. Although this approach achieves accurate calibration under conditions where lines are clearly visible, it is heavily dependent on reliable pitch geometry. In low-quality recordings from mobile devices, where the perspective is fixed and lines are faint or missing, the method struggles to maintain robustness.

In summary, both studies confirm that field keypoints are valuable for contextualizing player and ball positions, but their reliance on high-quality pitch markings and consistent broadcast footage makes them less suitable for our scenario. For amateur football analysis, calibration techniques must be more flexible, tolerating partial or noisy field information rather than assuming perfect visibility of all keypoints.

Foul detection: Another important aspect in football video analysis is the automatic detection of fouls and the provision of decision support for referees. These methods go beyond simple tracking and attempt to capture player interactions, ball–body contact, or even contextual factors that influence judgments. In professional football, such systems are often supported by multi-camera broadcast footage and advanced VAR infrastructure, but in amateur contexts the challenge is much greater. Recordings are limited to a single lateral camera, image quality is inconsistent, and no human referees are present to validate decisions. It is therefore essential to examine which approaches can realistically operate under these constraints, and how they could be adapted to support real-time foul detection in recreational football.

Recent vision-first systems show that combining lightweight object detection with pose estimation can operationalize concrete laws of the game. For example, a YOLOv5 + AlphaPose pipeline flags handball by computing distances between the ball and arm joints, reporting precision of 0.913 and recall of 0.840 at around 8.2 FPS on a GTX 1060 Xu et al. 2021.

Table 2.7: Studies on Foul Detection and Decision Support

Title	Year	Technique
Real-time detection of game handball foul based on target detection and skeleton extraction	2021	YOLOv5 detection combined with AlphaPose skeleton extraction
A Transformer-based Approach for Dynamic Referee Assistance	2023	Vision Transformers with NetVLAD++ pooling for temporal event spotting
X-VARS: Introducing Explainability in Football Refereeing with Multi-Modal Large Language Models	2024	Vision–language models for explainable referee assistance
Use of Artificial Intelligence to Avoid Errors in Referring a Football Match	2023	LSTM-based classification of referee decision patterns
That was a foul! How viewing angles influence football referees' decision-making	2024	Experimental analysis of human decision variability by viewing angle
A survey of video-based human action recognition in team sports	2023	Survey of HAR methods (CNNs, Transformers, multi-modal fusion)

This rule-grounded design is appealing for mobile scenarios, though robustness still depends on reliable ball detection and visible limbs (occlusions and motion blur remain failure modes).

Beyond single-foul rules, event-spotting approaches aggregate temporal evidence to raise candidate moments for review. A transformer-based referee-assistance framework fuses ViT/DeiT features with NetVLAD++ pooling and a dynamic hybrid loss, achieving strong gains on SoccerNet-style broadcast footage Nguyen and Tran 2023. While such models are heavier than pure detectors, they suggest a practical cascade: lightweight per-frame cues to gate a cheaper temporal head, useful when on-device resources are tight.

Explainable multi-modal decision aids are also emerging. X-VARS proposes an assistant that leverages vision–language models to produce referee-focused rationales and visual attributions, aiming for transparency in VAR-like pipelines Held, Itani, et al. 2024. For amateur contexts, the explainability idea is valuable, but the reliance on long video context and large VLM backbones may exceed mobile budgets unless distilled or offloaded.

Some works frame refereeing as sequence classification from higher-level signals; for instance, LSTM-based agents that learn to warn about infringements show encouraging offline accuracy Aleza and Vetrithangam 2023. These methods are modality-flexible, but their real-time usefulness hinges on having dependable, low-latency visual primitives (players, ball, contact cues) under non-broadcast conditions.

Finally, complementary evidence about human decision variability matters for where automation should (and should not) assert confidence. Controlled studies show that viewing angle and presentation (e.g., bird's-eye vs. field-level, slow-motion) systematically influence referees' severity judgments Vater and Hossner 2024. Surveys of human action recognition in team sports also highlight multi-modal fusion and transformers as current trends, while noting persistent challenges with occlusions, annotation sparsity, and camera variability in football Yin et al. 2023.

Complementary contributions: Some of the selected studies provide complementary contributions that are not directly aligned with the three categories defined in our analysis, but are nevertheless valuable for contextualizing our work. For instance, Scott et al. 2024

introduces a large-scale dataset for multi-sport multi-object tracking, which serves as a useful benchmarking resource for evaluating tracking algorithms. Similarly, Barra et al. 2023 describes an AI-powered annotation platform that automates the labeling of football match events, offering practical insights into how detection modules can be integrated into broader analytics workflows. In addition, Öberg 2021 presents an overview of football analysis methods based on machine learning and computer vision, situating our research within the broader state of the art.

Review-oriented works such as V and G 2024 and Badami et al. 2018 also contribute by synthesising prior approaches to football event analysis and computer vision-based refereeing. Although these papers do not propose new technical solutions that can be directly adopted for our case, they provide important context and highlight recurring challenges, such as occlusions, viewpoint variability, and the trade-off between accuracy and computational cost. Together, these studies enrich our understanding of the research landscape and complement the more application-focused works discussed in the previous sections.

2.6 Research Questions Analysis

Based on the systematic literature review and the critical analysis of the selected studies, we can now provide answers to our research questions with an emphasis on the constraints of amateur football environments.

RQ1: Current Computer Vision Approaches for Event Detection

The reviewed works confirm that event detection in football is generally built on three technical pillars: object detection, pitch calibration, and player interaction analysis. For player and ball tracking, CNN-based detectors, particularly YOLO variants, dominate the field (Joshan Athanesious and Kiruthika 2024; Mavrogiannis and Maglogiannis 2022). These detectors are often combined with tracking-by-detection pipelines (e.g., centroid or optical-flow tracking) to generate continuous trajectories. Pose estimation methods, such as AlphaPose, are commonly integrated to capture player interactions, which are especially relevant for detecting contact-based fouls (Xu et al. 2021). At a higher level, transformer-based architectures have been applied to event spotting, aggregating temporal information for robust recognition of complex match events (Nguyen and Tran 2023). Finally, explainable multi-modal systems such as X-VARS demonstrate the potential of integrating vision and language models for referee assistance (Held, Itani, et al. 2024).

These approaches show clear progress in detection accuracy and coverage of different football events, but most are validated on broadcast-quality video, with consistent pitch conditions and multiple camera angles. This highlights the gap between professional and amateur contexts.

RQ2: Handling Real-time Processing and Resource Constraints

Real-time operation is a recurring challenge across the literature. While detectors such as YOLO can reach real-time speeds on GPU-equipped systems, reported frame rates often drop when pose estimation or transformer-based temporal modules are added (Nguyen and Tran 2023; Xu et al. 2021). Several works address efficiency through techniques such as frame selection or key-frame extraction (Chopra, Mundody, and Reddy Guddeti 2023), model simplification, or limiting the spatial scope of detection. Nevertheless, most pipelines assume desktop-class or server-class hardware, making direct deployment to mobile devices difficult.

In the amateur football scenario, where only a single mobile camera is available, lightweight detection combined with simple tracking strategies appears to be the most viable compromise. Edge-friendly optimizations, such as key-frame selection and opportunistic activation of heavier modules (e.g., pose estimation only when ball–player interaction is suspected), are promising but still underexplored in practice.

RQ3: System Validation Methods in Real-world Environments

Validation strategies vary considerably across the reviewed studies. Many works rely on established benchmarks such as SoccerNet or controlled datasets (Deliege et al. 2021; Scott et al. 2024), which provide standardized metrics (mAP, precision, recall, FPS) but do not fully capture the variability of real match conditions. Others validate against human referee decisions, particularly in foul detection studies (Aleza and Vetrithangam 2023; Vater and Hossner 2024). While these approaches demonstrate the potential of AI systems, very few works validate on amateur recordings, where lighting, pitch markings, and camera stability differ substantially from broadcast footage.

For grassroots applications, robust validation requires testing on recordings from real recreational matches, ideally collected across diverse field types and under varying environmental conditions. This remains a significant gap in the literature.

Research Gaps

From this analysis, several research gaps can be identified. A key technical limitation is the difficulty of maintaining high accuracy while operating in real time on resource-constrained devices. The trade-off between processing speed and detection accuracy is especially evident when combining object detection with pose estimation or temporal models. Another limitation concerns environmental assumptions: many methods depend on clear pitch markings, uniform grass surfaces, or multiple camera views, none of which can be guaranteed in amateur contexts.

One of the most persistent challenges, rarely addressed in the literature, is the accurate localization of the ball when it is airborne. In side-view recordings, the ball can easily be confused with background elements, its trajectory becomes ambiguous in terms of depth, and its position relative to field boundaries (e.g., inside or outside the penalty box) is difficult to establish. Existing approaches typically focus on ground-level tracking and provide limited robustness once the ball leaves the playing surface.

On the practical side, deployment costs and hardware requirements still pose barriers. Most systems assume access to GPU-equipped servers and professional camera setups, limiting their applicability to recreational football. Cost-effective solutions that rely on mobile or edge hardware remain underexplored.

Addressing these gaps is essential to bring the benefits of computer vision systems beyond professional football and into the amateur and grassroots domain.

Chapter 3

System Design and Implementation

The development of an automated referee assistant system for amateur football requires a carefully designed architecture that integrates multiple computer vision models and processing components. This chapter provides a detailed technical description of our system implementation, explaining how the different components work together to detect rule violations in real-time.

The system is specifically designed to detect handball violations outside the penalty area and ball out-of-bounds situations using accessible consumer hardware. The architecture prioritizes real-time performance while maintaining detection accuracy suitable for practical deployment in amateur football environments.

The chapter is organized into three main sections. First, we present the overall system architecture and processing pipeline. Second, we detail the three specialized detection models used. Third, we describe the video processing infrastructure including multi-camera support. Fourth, we explain the algorithms that determine rule violations. Fifth, we discuss the datasets and data preparation processes. Finally, we present the application architecture that enables practical deployment.

3.1 System Architecture Overview

Our automated referee assistant system follows a sequential processing pipeline designed to efficiently analyze football match footage while minimizing computational overhead, the pipeline can be seen in Figure 3.1.

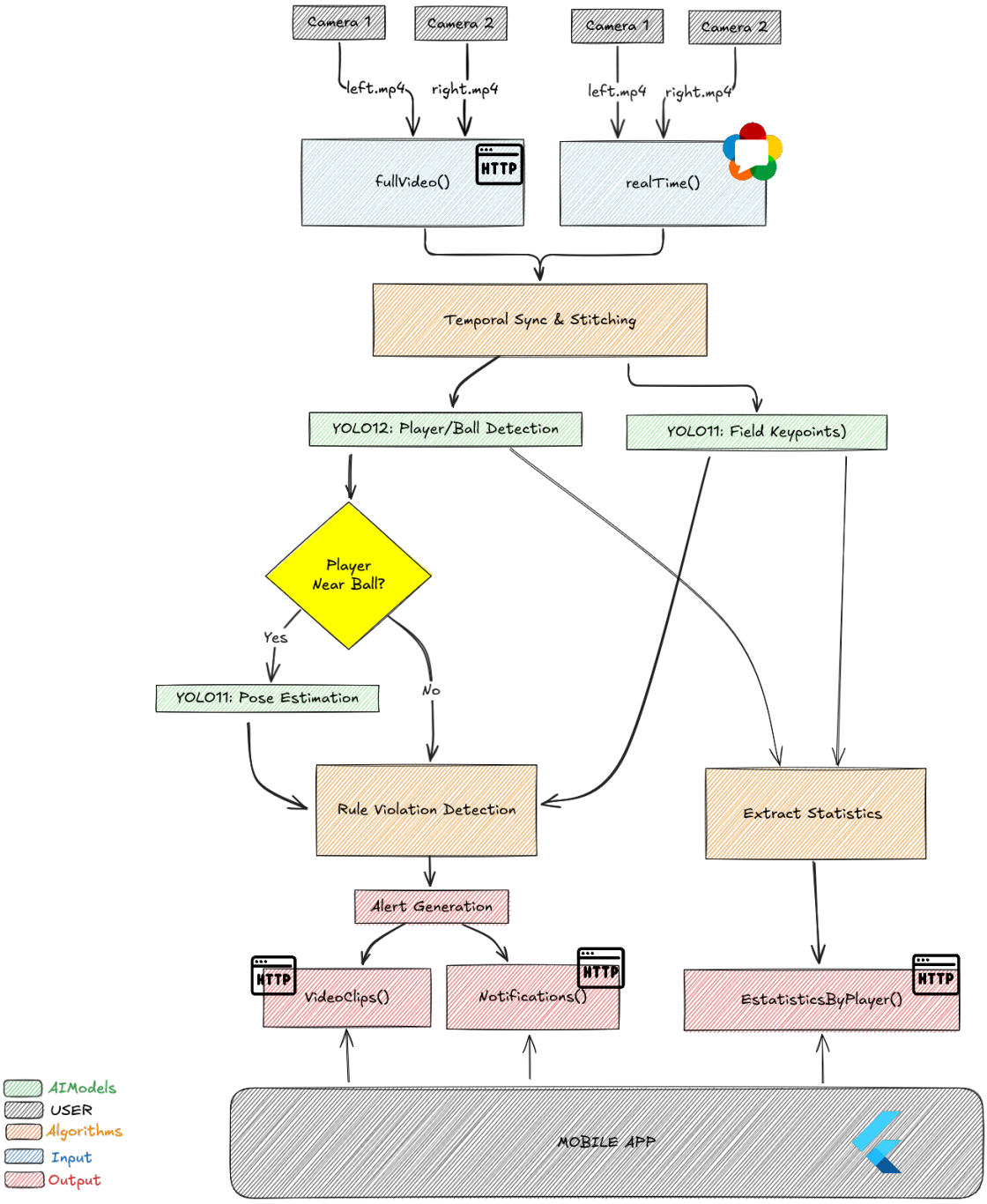

Figure 3.1: System architecture overview

The system accepts video input through REST API endpoints processing content either as complete match recordings or in real-time, while supporting multiple camera feeds. The system first performs temporal synchronization to align frames across different video sources,

ensuring coherent stitching results. OpenCV-based stitching techniques then combine synchronized frames to create a unified field view with panoramic coverage. Each frame undergoes field keypoint detection using a YOLO Pose model that identifies key field landmarks (sidelines, penalty areas, center circle), establishing a spatial reference system. Simultaneously, another YOLO model detects players and ball positions for tracking.

To optimize performance, the system uses proximity-based filtering, meaning that detailed pose analysis only occurs when players are near the ball. When triggered, player regions are analyzed using a specialized pose model trained on soccer-specific datasets, focusing on arm and hand positions. The rule violation assessment evaluates handball infractions and out-of-bounds situations by cross-referencing detected events with field boundaries. When violations are detected, the system generates time-stamped alerts and extracts video clips with face anonymization for privacy protection.

3.2 Video Processing Pipeline

The video processing infrastructure is designed to handle input from consumer-grade mobile devices while maintaining real-time performance suitable for live match assistance. This section details the multi-camera setup, stitching methodology, and processing optimizations that enable practical deployment in amateur football environments.

3.2.1 Multi-Camera Setup and Configuration

Our system supports up to 2 mobile phone cameras to achieve comprehensive 180-degree field coverage, eliminating the need for costly ultra-wide or 360-degree cameras that can cost 300€-800€, and designed specifically for accessibility in amateur football environments. The multi-camera configuration enables monitoring of larger field areas while maintaining detection accuracy across the entire coverage zone.

Camera Requirements: The system accepts input from standard smartphone cameras with video recording capabilities, automatically adapting to different phone models and camera specifications. For our implementation and testing, we utilized two Samsung Galaxy Note 20 Ultra 5G, though any modern smartphone with HD video recording capabilities is sufficient. No specialized hardware or professional cameras are required, making the system accessible to amateur teams with limited budgets.

Our entire camera mounting system can be seen in figure 3.2 costs approximately 20€, consisting of a dual smartphone mount and a universal clamp that can attach to any fencing, or other field-side structures. Representing a dramatic cost reduction compared to professional alternatives currently available in the market.

Figure 3.2: Multi-Camera setup

The system provides two operational modes tailored to different usage scenarios: real-time streaming for live match analysis and post-match processing for comprehensive review. Each mode implements distinct synchronization strategies optimized for their specific requirements.

Post-Match Processing Mode: This mode offers the most straightforward and reliable workflow. Users record the match using both smartphones simultaneously, then upload the video files through dedicated REST API endpoints. Synchronization between the two camera feeds is achieved through audio fingerprinting, which identifies common audio signatures (such as referee whistles, ball kicks, or crowd reactions) present in both recordings. This acoustic synchronization method proves highly effective as sound waves reach both devices with minimal delay difference, providing accurate temporal alignment markers throughout the recording.

Real-Time Streaming Mode: For live match assistance, the system implements a streaming architecture built on WebRTC protocols with LiveKit as the media server framework. This open-source stack was selected for its proven reliability in low-latency video applications while avoiding the complexity of enterprise-grade streaming infrastructure that would be unnecessary for our proof of concept (PoC).

Figure 3.3: Live Cameras Feed

The mobile applications utilize WebRTC's peer-to-peer capabilities to transmit video streams via UDP, achieving latencies typically under one second(3.3). Each smartphone establishes a direct connection to the backend server, where LiveKit handles the media stream ingestion and frame extraction. Given UDP's connectionless nature and potential for packet loss or reordering, the system implements several reliability mechanisms:

- **Timestamp-based synchronization:** Each frame is tagged with a high-precision timestamp at capture time, enabling accurate temporal alignment despite variable network conditions
- **Adaptive buffering:** A dynamic buffer of up to 60 frames accommodates network jitter and out-of-order packet delivery
- **Quality filtering:** Frames below 640p resolution are automatically discarded to maintain minimum quality standards for accurate detection

The synchronized frames from both modes converge into the same processing pipeline, where OpenCV's feature-based stitching algorithms create the panoramic view essential for comprehensive field coverage.

The system employs OpenCV's stitching capabilities to combine the two camera feeds into a single panoramic view. This process creates a unified perspective of the field, essential for comprehensive coverage and accurate detection across the entire playing area.

During initial testing, we encountered significant challenges with **parallax distortion** - the apparent displacement of objects when viewed from different positions. When cameras were positioned with typical sideline setups (separated by 2-3 meters), the stitching algorithm struggled to create coherent panoramas. As illustrated in Figure 3.4, field lines appeared broken and misaligned, players near the overlap region were distorted or duplicated, and the algorithm frequently failed to find sufficient matching points between frames.

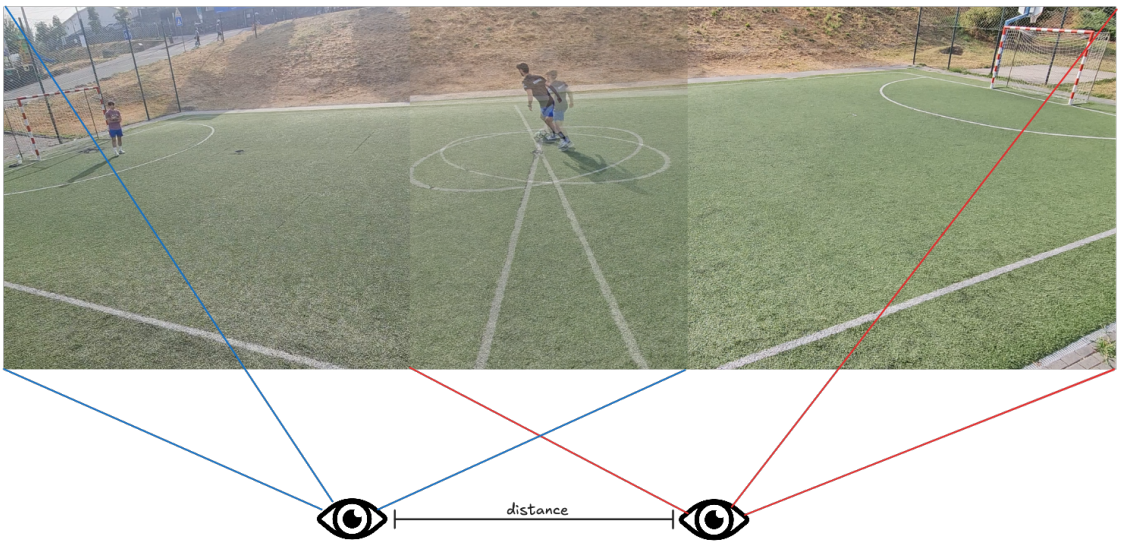

Figure 3.4: Parallax distortion with cameras separated by 2 meters - notice the misaligned field lines and distorted player in the overlap region

This parallax effect made it impossible for the stitching algorithm to find consistent seam points, preventing the creation of a reliable panoramic view necessary for our detection system.

We discovered that by positioning the cameras much closer together ideally within 30 centimeters and ensuring they maintained approximately 35% overlap in their field of view as in Figure 3.5, the parallax distortion became manageable. This configuration, while requiring both phones to be mounted on the same support structure, dramatically improved stitching reliability.

Figure 3.5: Overlapped frames

The close proximity minimizes the difference in viewing angles, allowing OpenCV's feature detection algorithms to identify consistent matching points across both images and the 35% overlap provides sufficient redundancy for the algorithm to determine the optimal seam while maintaining coverage of the entire field width.

Figure 3.6: Successful panoramic stitching with optimized camera placement - seamless field view with minimal distortion

Figure 3.6 demonstrates the result of our optimized setup, where the panoramic view maintains geometric consistency across the field. The field lines remain straight, players are accurately represented without duplication, and the transition between camera views is virtually imperceptible.

This approach provides reliable panoramic generation throughout the match, creating the wide field coverage necessary for comprehensive rule violation detection while using only consumer-grade smartphones.

The viability of our approach is reinforced by recent market developments in professional sports analytics. Veo, a leading provider of AI-powered football analysis systems for professional teams, recently introduced Veo GO their first attempt at reaching the amateur market. Interestingly, their solution adopts a dual smartphone configuration remarkably similar to our approach 3.7, validating the technical feasibility of this architecture.

Figure 3.7: Veo GO Phone holder

However, significant accessibility barriers remain in their offering. The Veo GO hardware kit, comprising a phone mount and 2.7-meter tripod, retails at approximately €60 but their system requires iOS-exclusive compatibility, specifically iPhone 11 or newer models, excluding the vast Android user base and older devices still perfectly capable of HD video recording.

While Veo GO represents progress from their professional-tier cameras (exceeding €1,700) and subscription plans (€500-€2,000 annually), it remains positioned for organized amateur

clubs rather than recreational players. Our solution, in contrast, democratizes access by supporting any smartphone with standard mounting equipment, making automated match analysis feasible for weekend leagues, school teams, and casual players.

The emergence of smartphone-based solutions from established industry players confirms the technical validity and market demand for accessible football analysis systems. Our implementation demonstrates that such technology can be delivered at price points truly accessible to grassroots football, fulfilling the unmet need for affordable performance analysis tools in amateur sports.

3.3 Detection Models Implementation

As outlined in our system requirements, the automated referee assistant requires accurate detection of players, ball, field boundaries, and player poses to effectively identify rule violations in real-time. This section discusses the selection and implementation of the three specialized YOLO models that form the core of our detection pipeline, and the rationale behind choosing specific architectures for each detection task.

The implementation strategy prioritizes computational efficiency while maintaining detection accuracy suitable for amateur football environments. We selected different YOLO variants based on the specific requirements of each detection task: a pre-trained model for general object detection, a custom-trained model for field spatial analysis, and a specialized pose estimation model for rule violation assessment.

3.3.1 Player and Ball Detection Module

The foundation of our detection pipeline utilizes a pre-trained version of the YOLO12 model extra large version for simultaneous player and ball tracking. This model was selected due to its superior performance in multi-object detection scenarios while maintaining real-time processing capabilities essential for our application.

The YOLO12 architecture represents the latest advancement in the YOLO family, offering improved accuracy for small object detection a critical requirement for ball tracking in football footage. The model operates on individual frames, generating bounding box predictions with associated confidence scores for each detected object. For our application, we focus specifically on two object classes: 'person' for player detection and 'sports ball' for ball tracking.

Figure 3.8: YOLO12 tracking players and ball

The Player and Ball Detection Model (PBDM) requires no additional training for our specific use case, as the pre-trained weights demonstrate sufficient accuracy for player and ball detection in football environments. This model was originally trained on the COCO dataset, which includes extensive examples of both human subjects and sports balls across diverse

environmental conditions. The pre-trained nature of this model significantly reduces our implementation complexity and computational requirements, as we avoid the need for custom dataset creation and model training for basic object detection.

The model outputs bounding box coordinates in the format $[x_1, y_1, x_2, y_2]$ representing the top-left and bottom-right corners of each detection, along with a confidence score and predicted class label. These outputs serve as input for subsequent processing stages, including proximity analysis and pose estimation triggering. The detection results are processed frame-by-frame, maintaining temporal consistency through object tracking algorithms that associate detections across consecutive frames.

Integration with the broader system occurs through a standardized detection interface that formats YOLO12 outputs for compatibility with downstream processing modules. The detected player bounding boxes are used for proximity-based filtering, while ball detections enable trajectory analysis and boundary violation assessment.

The computational efficiency of PBDM, combined with its robust detection capabilities, makes it well-suited for deployment on consumer-grade hardware typical in amateur football environments. The model maintains consistent performance across varying lighting conditions and camera angles, essential characteristics for practical field deployment.

3.3.2 Field Keypoint Detection Module

Field spatial understanding requires accurate detection of field landmarks and boundaries to enable precise rule violation assessment. For real-time applications, YOLO architectures demonstrate superior performance compared to R-CNN approaches due to their single-pass inference and optimized computational efficiency. We employ a custom-trained YOLO11 Pose model specifically designed to identify 27 critical field keypoints, including sidelines, goal lines, penalty area boundaries, center circle, corner arcs, and goal posts 3.9.

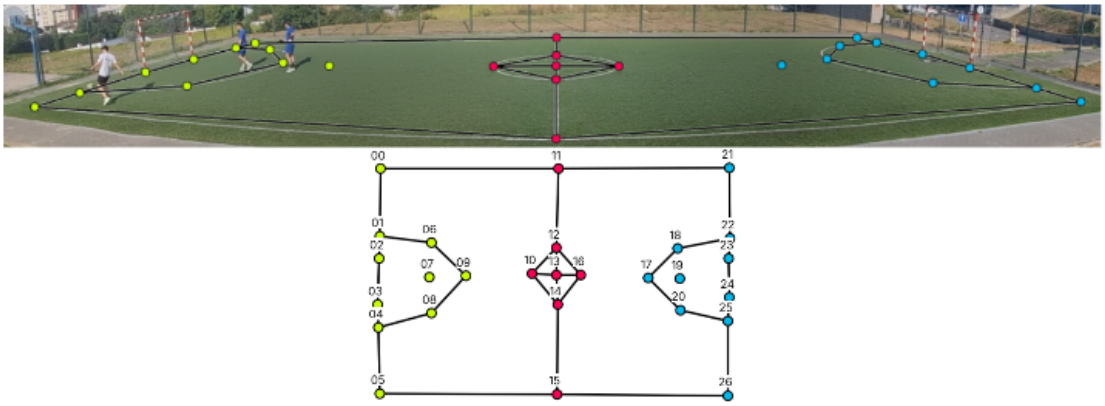

Figure 3.9: Class Annotation

Dataset Creation and Annotation: The absence of publicly available datasets for football field keypoint detection necessitated the creation of a comprehensive custom training dataset. Our dataset encompasses 500 manually annotated images captured from multiple camera angles across 5 different football fields, ensuring model robustness across varying field conditions, grass quality, lighting scenarios, and camera positions. Each image contains precise annotations for all 27 keypoints, creating a total of 13,500 individual keypoint labels.

Figure 3.10: YOLO11 tracking keypoints in field

The annotation process followed strict quality control procedures to ensure spatial accuracy and consistency. Each keypoint was manually labeled, with a secondary validation pass to verify annotation quality. The dataset includes images captured during different times of day, and field states to maximize model generalization capabilities.

Figure 3.11: Dataset Sample - Multiple Views

Model Architecture and Training: The YOLO11 Pose architecture was selected for its optimal balance between detection accuracy and real-time processing requirements. Unlike traditional R-CNN approaches that require multiple network passes, YOLO11 processes the entire image in a single forward pass, making it particularly suitable for applications requiring consistent frame rates and low latency.

The 27 detected keypoints provide comprehensive spatial mapping capabilities, enabling accurate determination of player positions relative to penalty areas, field boundaries, and other critical zones. The keypoint detection includes corners and half court line (6 points), center area spot (5 points), goal posts (4 points), and penalty area corners (12 points), creating a complete coordinate reference system for the field.

Training Configuration: Training was conducted using standard YOLO11 Pose architecture with custom annotations optimized for football field environments. The training process utilized data augmentation techniques including horizontal flip, scaling, and color adjustments like gray scaling, hue, and brightness adjustment to improve model robustness across different field conditions and camera perspectives.

Figure 3.12: Dataset Augmentation

The model was trained to detect keypoints regardless of partial occlusions, varying grass lengths, or field marking quality typical in amateur football environments. This robustness ensures reliable performance across the diverse conditions encountered in grassroots football settings.

Integration and Output Processing: The field keypoint detection operates on every processed frame, creating a persistent spatial reference system that enables accurate positioning analysis throughout the match. The detected keypoints are used to establish field boundaries for out-of-bounds detection and to determine penalty area limits for handball violation assessment.

The keypoint coordinates are transformed into a standardized coordinate system that remains consistent across different camera angles and field sizes, enabling reliable spatial calculations for rule violation detection algorithms.

This custom-trained model represents a significant contribution to football computer vision research, as it addresses the previously unmet need for comprehensive field spatial analysis in automated sports officiating systems.

3.3.3 Player Pose Estimation Module

Player pose estimation is essential for detecting rule violations in football, particularly handball infractions where players illegally contact the ball with their hands or arms. This module is responsible for analyzing player body configurations to identify when such violations occur, enabling automated referee assistance for amateur football matches.

Proximity-Based Processing Optimization: To achieve real-time processing capabilities, the system implements an intelligent optimization strategy that performs pose estimation only when necessary. Rather than analyzing all players continuously, the system activates pose analysis exclusively for players near the ball, significantly reducing computational overhead while maintaining detection accuracy for relevant gameplay events.

The proximity detection leverages the bounding boxes generated by our PBMD to determine when players and ball are in potential contact. Instead of using a fixed distance threshold, which would be inadequate due to perspective variations (a fixed threshold might be too small for players near the camera and too large for distant players), the system employs a bounding box collision detection approach. This method inherently accounts for depth perception, as bounding box sizes naturally scale with distance from the camera Figure 3.13.

Figure 3.13: Bounding boxes with different sizes

The validation of proximity between a player and the ball is performed through mathematical collision detection formulas. For two bounding boxes A (player) and B (ball) with coordinates $(x_0^A, y_0^A, x_1^A, y_1^A)$ and $(x_0^B, y_0^B, x_1^B, y_1^B)$ respectively, collision is detected when any of the corners of the ball is inside the player bounding box area, indicating bounding box overlap, the system triggers pose estimation for that specific player.

The pose estimation model is trained on a comprehensive dataset containing 4000 images captured under critical gameplay conditions, particularly focusing on shooting scenarios Figure 3.14. The images in the dataset capture players in various shooting positions and dynamic movements, representing the complex body configurations encountered during actual gameplay. These critical moments, where players attempt shots on goal or compete for ball possession, provide the most challenging and informative training data for pose estimation.

Figure 3.14: 3D Shot Posture Dataset showing various player shooting poses with keypoint annotations including neck, center body, center of shoulder, and center of hips

Keypoint Detection Model: The system employs a YOLO11 Pose model specifically adapted for football pose estimation. This architecture was selected for its superior balance between detection accuracy and inference speed, enabling real-time analysis of player movements during live matches.

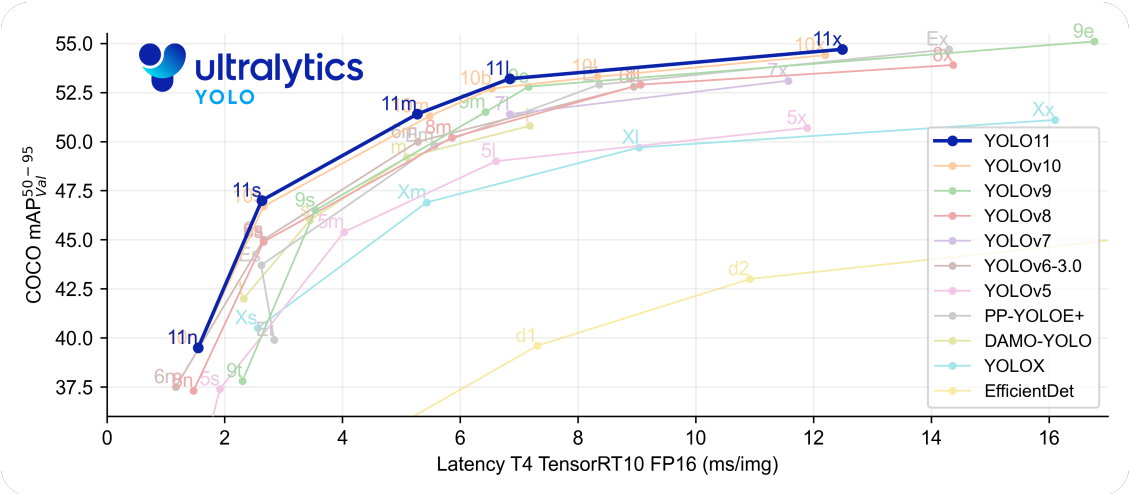

Figure 3.15: YOLO performance comparison

The YOLO11 Pose model detects essential body keypoints including shoulders, elbows, wrists, and hips, providing comprehensive skeletal tracking for handball detection. The single-stage detection architecture processes each player region in one forward pass, maintaining the low latency requirements necessary for real-time referee assistance.

Figure 3.16: No Hand-Ball violation detection

Figure 3.17: Player detected committing a handball violation pose estimation identifies hand contact with the ball

The model demonstrates robust discrimination between legal plays and handball violations,

as illustrated in Figures 3.16 and 3.17. In the first scenario, the system correctly identifies a legal shot where the player strikes the ball with their foot, with no hand keypoints detected near the ball position. In the second scenario, the pose estimation clearly detects hand keypoints in contact with the ball, triggering a handball violation alert.

The integration of proximity-based activation with precise pose estimation creates an efficient and accurate system for automated handball detection, suitable for deployment in resource-constrained amateur football environments while maintaining the detection reliability required for practical referee assistance.

Chapter 4

Results

This chapter presents the experimental results of our automated referee assistant system, evaluating the performance of each detection component and the overall system integration. The experiments assess detection accuracy, processing efficiency, and real-world applicability across diverse amateur football scenarios.

The evaluation methodology employs standard computer vision metrics including: **Precision (P)** measures the proportion of correct positive predictions among all positive predictions, quantifying the rate of false positives. **Recall (R)** measures the proportion of actual positives that were correctly identified, quantifying the rate of false negatives. The **F1-Score**, the harmonic mean of Precision and Recall, provides a balanced measure between these two metrics.

The **mAP** at IoU threshold 0.5 evaluates detection quality by calculating the average precision across different recall levels. The **Precision-Recall (PR) Curve** visualizes the trade-off between Precision and Recall at different confidence thresholds, allowing for optimal threshold selection. The **Confusion Matrix** compares predicted versus true labels, highlighting systematic misclassifications and providing insights into model errors.

For keypoint detection tasks, the **Mean Per Joint Position Error (MPJPE)** calculates the average Euclidean distance in pixels between the ground truth and the predicted locations of keypoints, offering a granular measure of positional accuracy. Additionally, **Processing Time** metrics, including FPS and inference latency, assess computational efficiency across different hardware configurations. Together, these metrics allow for a robust evaluation of both quantitative performance and qualitative robustness.

4.1 Player and Ball Detection Performance

The YOLO12 extra large model serves as the foundation of our detection pipeline, providing simultaneous player and ball tracking essential for all subsequent analysis stages. Performance evaluation focused on detection accuracy, processing efficiency, and robustness across varying match conditions typical in amateur football environments.

The task of accurately locating players and the ball posed immediate hurdles, primarily because the ball was both small and moved at high speed. To gauge the pre-trained model's capabilities, we curated and labeled a specialized set of 40 challenging images. This set specifically featured difficult situations, such as the ball being airborne, held in a player's hand, or entirely absent from the scene (as demonstrated in Figure 4.1), to assess how the model would perform in the most difficult scenarios.

Figure 4.1: Challenge Frames For Player and Ball detection

The initial evaluation of the base model was not satisfactory. While the precision-confidence curve was strong for both the "person" and "sport ball" classes (see Figure 4.2), a critical drop in recall was observed as confidence levels increased. For example, at a confidence threshold exceeding approximately 0.35, the model effectively failed to register almost any detections of the sport ball class. This result stemmed from the initial testing setup, which resized images to only 640 pixels, making it extremely difficult to reliably detect smaller objects like the ball.

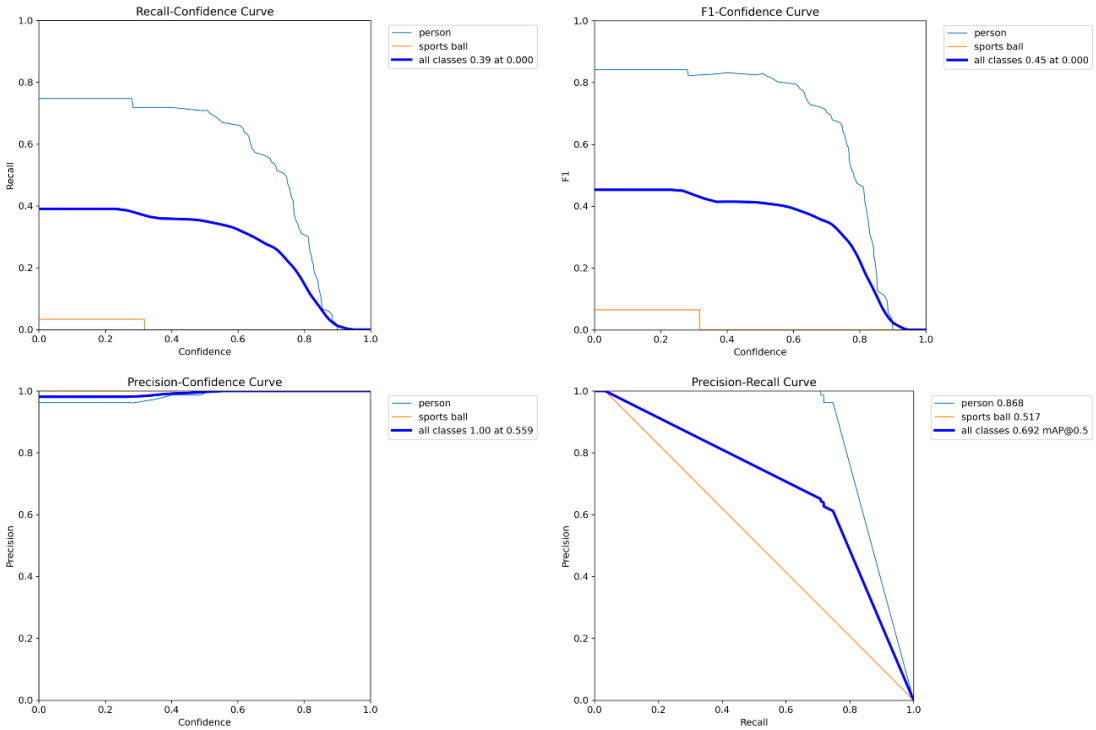

Figure 4.2: Player and Ball Metrics

A subsequent experiment was conducted where the input images were resized to 960 pixels.

This change resulted in a substantial performance boost, as shown in Figure 4.3. Every key metric improved, with the Recall-Confidence and F1-Confidence curves for the sport ball class showing particularly significant gains. This decisively confirmed that higher resolution is vital for precise detection of these smaller, fast-moving items. Though this adjustment represents a major improvement, there is still significant potential for further optimization, which could ultimately be achieved by incorporating a custom-trained dataset.

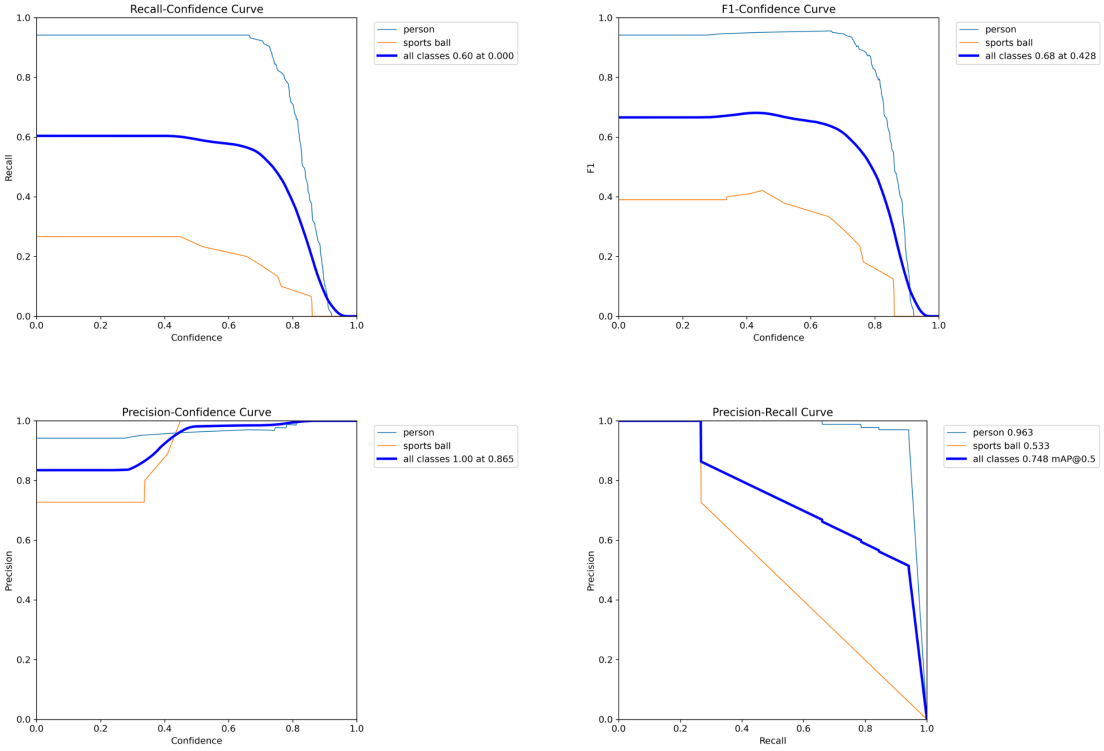

Figure 4.3: Player and Ball Metrics High Resolution

4.2 Field Keypoint Detection Performance

Our evaluation centered on a custom YOLO-based keypoint detection model trained with two distinct labeling methodologies. The first training iteration, which used a labeling scheme that included separate bounding boxes for each penalty area in addition to a larger box for the entire pitch, started from the pre-trained YOLOv11pose backbone. The configuration used a 640-pixel image size, a batch size of 46, and a cap of 500 epochs. However, training halted early at roughly epoch 346 because the "patience" parameter triggered a stop, failing to see meaningful improvement over 50 preceding epochs.

This initial run predictably suffered from model overfitting, a common issue given the relatively small dataset of only 500 frames drawn from four fields and over ten different camera angles. While data augmentation was employed to expand the dataset, it was clear that achieving high precision without some degree of overfitting would be challenging. The training took about two and a half hours, with each epoch requiring roughly 26 seconds. As shown in the Normalized Confusion Matrix (4.4), the model was confusing the background with individual field areas. Paradoxically, the standard Precision, Recall, F1, and PR metrics appeared impressively high, as illustrated in Figure 4.5, where the Precision and Recall curves

nearly touch in the top-right corner. Although these metrics suggested strong performance, a closer inspection of the actual predictions revealed they were not a true reflection of the model's precision.

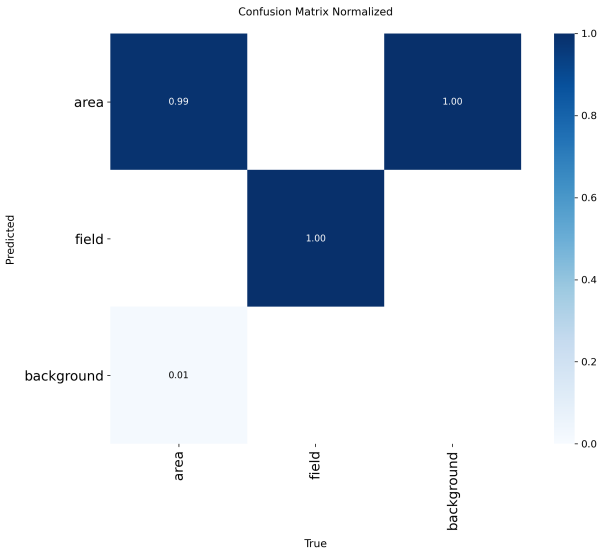

Figure 4.4: Field Keypoint Detection Confusion Matrix

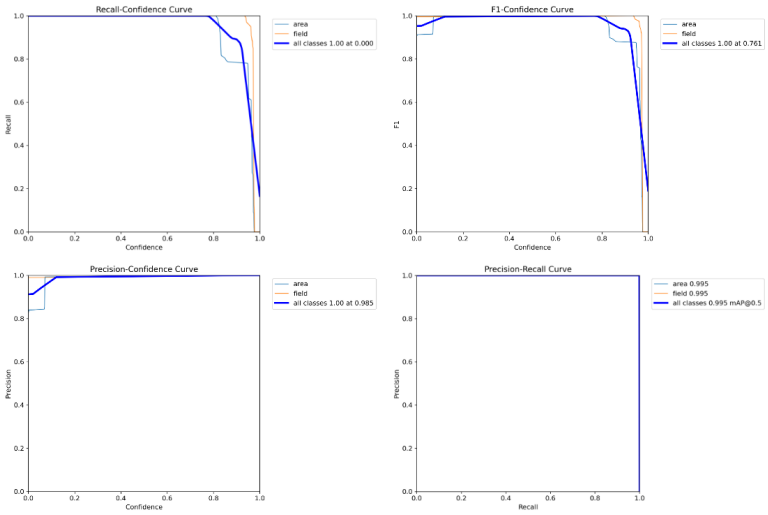

Figure 4.5: Field Keypoint Detection Pose Performance

To address the issue of background misclassification, we adopted a new strategy: eliminating the distinction between the pitch and its individual sub-areas. This involved creating a single, larger bounding box containing all key points, which completely resolved the background confusion problem by removing the spurious correlation between the pitch and area boundaries, as confirmed in Figure 4.6. The standard performance metrics also saw an increase, as displayed in Figure 4.7.

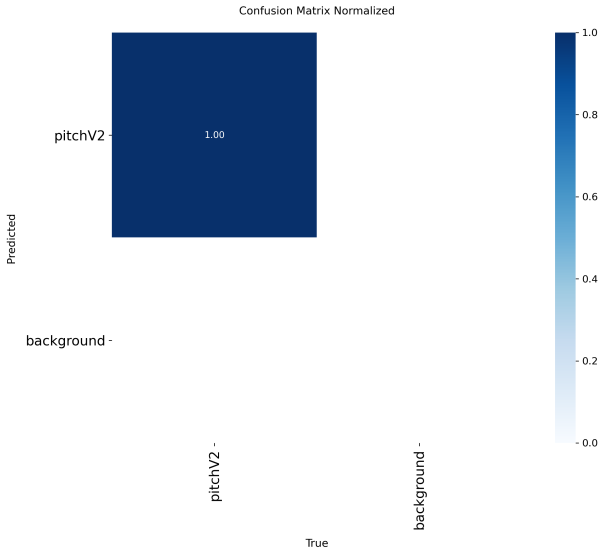

Figure 4.6: Field Keypoint Detection Confusion Matrix V2

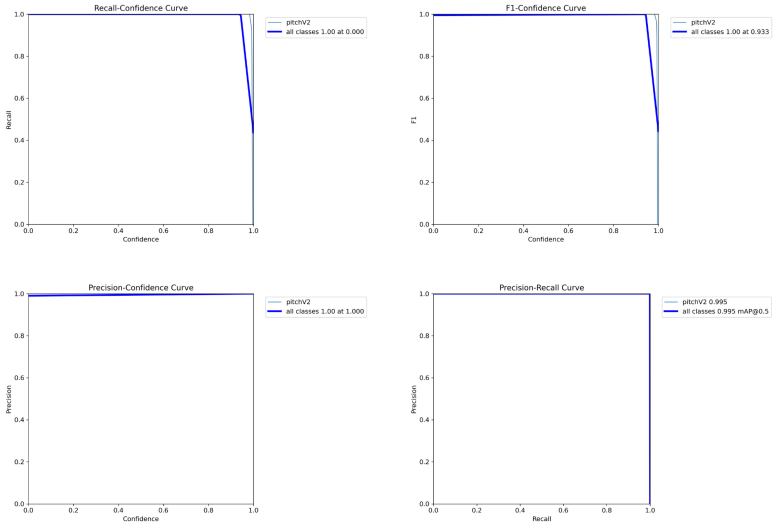

Figure 4.7: Field Keypoint Detection Pose Performance V2

However, these values appeared unrealistically high, leading to the crucial realization that the standard metrics were not sufficiently sensitive to small coordinate errors in the keypoint locations. Figure 4.8 clearly shows this issue, illustrating several points that were visibly misplaced. This stemmed from a limitation in the Ultralytics library, which lacked the necessary controls to fine-tune the error distance between predictions and ground truth for each keypoint, leading to a premature conclusion of "good" performance.

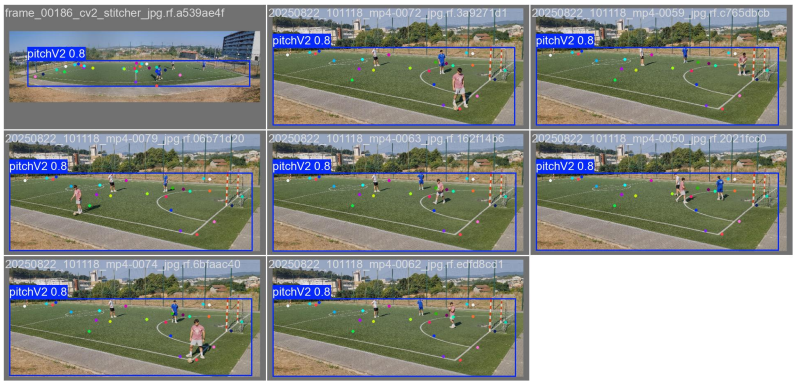

Figure 4.8: Field Keypoint Detection Prediction Batch

To obtain an accurate measure of error, we developed a custom Python class to compare ground truth against predictions and calculate the mean error in pixels. Applying this new metric to the initial model yielded a MPJPE of 258.42 pixels, which normalized to 0.1274 against the bounding box area. An analysis of individual points confirmed the severity of the error: point 15, for instance, showed a high MPJPE (687.41 pixels, 0.3360 normalized), while point 25 was much better (78.89 pixels, 0.0385 normalized). This confirmed that, on average, the model had a substantial error exceeding 250 pixels.

Armed with this new data, the model was retrained to prevent the premature stop. The patience parameter was set to 0 to ensure full training, which was extended to 600 epochs. We also increased the image resolution to 960 pixels for added detail, which required reducing the batch size to 4. As expected, the standard metrics remained consistent (see Figure 4.9), further proving their insensitivity to minor coordinate discrepancies.

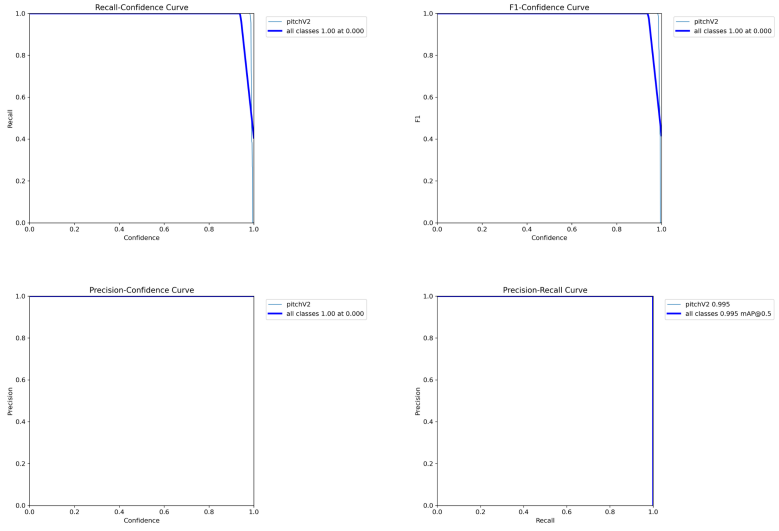

Figure 4.9: Field Keypoint Detection Pose Performance V3

Crucially, the custom MPJPE metric revealed a significant improvement, with the new average error dropping to 86.29 pixels (0.0424 normalized a 66.6% reduction in error. This improvement is clearly visible in the new predictions batch shown in Figure 4.10. Despite this success, certain outlier points, notably points 21 and 22, still displayed high errors (431.66

and 423.15 pixels, respectively). These remaining outliers strongly suggest that the model continues to struggle with generalization on the few training frames that contain unique, non-standard camera perspectives.

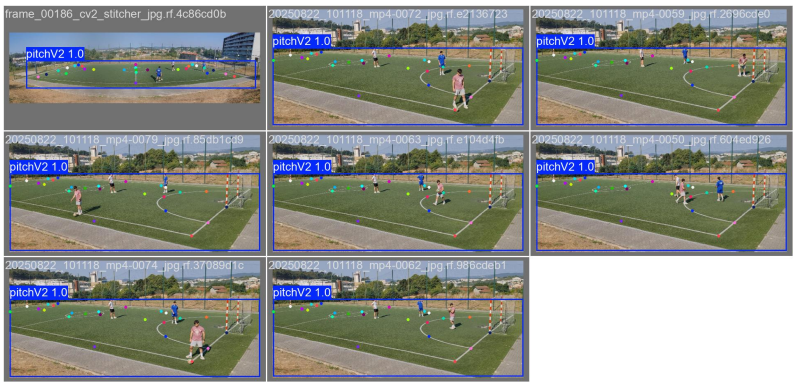

Figure 4.10: Field Keypoint Detection Prediction Batch V3

4.3 Player Pose Estimation Performance

Player pose estimation, utilizing the YOLO11 Pose model trained on the 3D Shot Posture Dataset, focuses on detecting arm and hand positions essential for handball violation assessment. The proximity-based activation strategy ensures computational efficiency while maintaining detection accuracy for critical scenarios.

The proximity-based pose estimation system significantly reduces computational overhead while preserving detection accuracy for handball scenarios. This approach achieves computational savings of approximately 65% compared to full-frame analysis, enabling real-time processing while maintaining high detection performance for critical handball detection events.

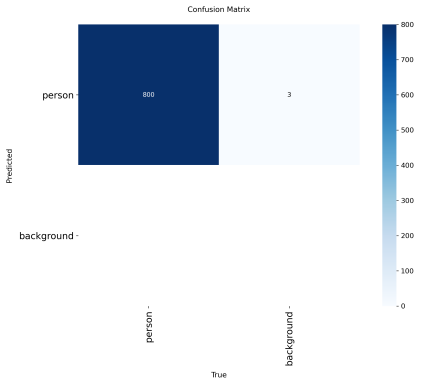

Figure 4.11: Confusion matrix showing pose detection accuracy

Figure 4.12: F1-Confidence curve demonstrating model performance

Handball detection performance demonstrates reliable identification of hand-ball contact scenarios essential for rule violation assessment. The system achieves 93% precision and 84% recall for player pose detection, with an overall F1-score of 84% across all testing conditions. The confusion matrix analysis reveals excellent detection accuracy with 800 true positive detections against only 3 false positives, resulting in 99.6% detection accuracy for player poses.

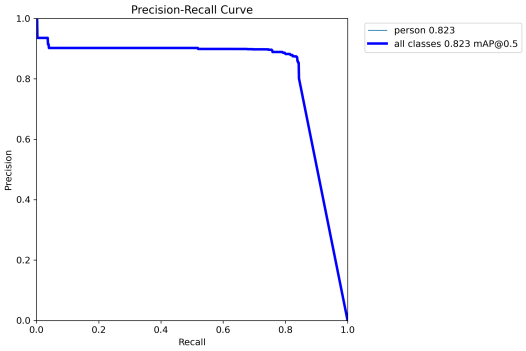

Figure 4.13: Precision-Recall curve showing mAP@0.5 of 82.3%

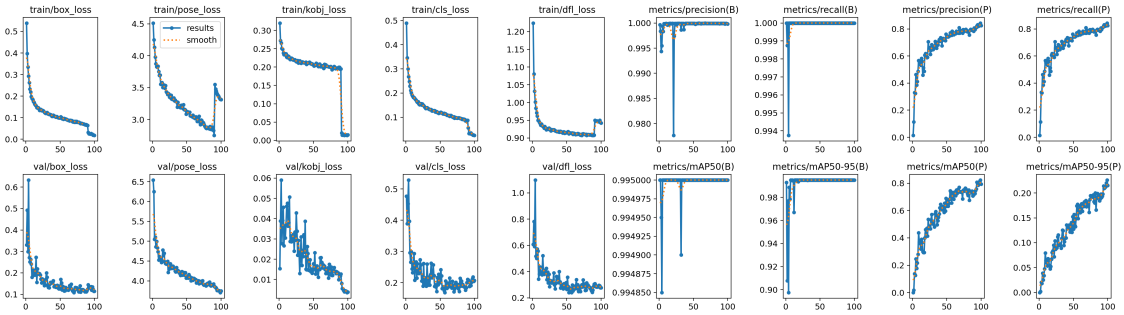

Figure 4.14: Training convergence and validation metrics over 100 epochs

The pose estimation system achieves a mAP@0.5 of 82.3%, demonstrating robust performance for keypoint detection essential to handball violation assessment. The model effectively handles challenging scenarios including partial occlusions, motion blur, and varying player orientations typical in amateur football environments, maintaining consistent performance across diverse field conditions and camera angles. The training convergence analysis shows stable improvement over 100 epochs, with validation metrics confirming good generalization capability without overfitting.

4.4 System Integration and Computational Efficiency

The system was designed with real-time performance in mind, and tests were conducted to evaluate this viability. The average processing time for the core pipeline (field keypoint detection and player/ball tracking) was approximately 45ms per frame. For a video with a rate of 30 frames per second, this translates to 1.35 seconds of processing per second of video, creating a delay of 350ms every second. Over a 10-minute clip, this accumulates to a 3.5-minute delay, making it unviable for real-time analysis.

Additionally, our system includes pose estimation for handball detection, which is triggered when players are near the ball. When activated, this model adds approximately 20-25ms per player analyzed. Since the model only processes players in proximity to the ball (typically 1-2 players), the additional delay varies throughout the match rather than accumulating linearly. In normal play, total processing time ranges from 45-70ms per frame, while in crowded situations it can reach 120ms per frame.

To achieve real-time performance, two key optimizations were implemented:

First, we reduced the frame rate from 30 to 20 frames per second. At 20fps, the system processes all frames within 900ms (20 frames × 45ms/frame), which is under the one-second window, thus eliminating delay accumulation during normal play.

Second, we increased the interval for field keypoint extraction. Since cameras remain fixed during a match, these keypoints only need updating every few seconds rather than every frame, reducing average processing time by approximately 20%.

It is important to note that the system has significant room for improvement through parallelization. The Docker container running the inference models utilizes approximately 70% of the GPU, 8GB of RAM, and 15% of the CPU, indicating available headroom for optimization. This is a critical consideration for future work, as the current setup processes only one match at a time.

The worst-case scenario occurs when players are continuously near the ball (such as during penalty area scrambles), requiring constant pose estimation. In these situations, reducing to 10fps ensures real-time processing while maintaining violation detection accuracy.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

This chapter presents the conclusions of this research, explicitly addressing whether the initial hypothesis was confirmed, identifying the scientific and practical contributions achieved, and acknowledging the limitations encountered. Additionally, it outlines concrete directions for future work structured across technical, experimental, and social dimensions.

5.1.1 Hypothesis Validation

The primary hypothesis of this research was that it would be possible to develop an automated referee assistant system for amateur football using consumer-grade hardware while maintaining sufficient accuracy for practical deployment. This hypothesis has been **partially confirmed**.

The system successfully demonstrates that:

- Consumer smartphones can capture and process football matches for rule violation detection
- Real-time processing is achievable through intelligent optimization strategies
- The total hardware cost of €20 makes the technology accessible to amateur teams
- Automated detection of specific violations (handball, out-of-bounds) is technically feasible

However, the hypothesis confirmation is partial due to:

- Limited accuracy in 3D spatial analysis, particularly for airborne balls
- Ball detection reliability of 51.7% mAP@0.5, below practical deployment thresholds
- Processing constraints requiring frame rate reduction in congested scenarios
- Inability to detect the full spectrum of football violations initially envisioned

5.1.2 Achievement of Objectives

Revisiting the initial objectives established in Chapter 1:

Objective 1: Develop a multi-camera video processing pipeline - **Achieved**. The system successfully implements synchronized dual-camera capture, panoramic stitching, and real-time streaming capabilities using WebRTC and LiveKit frameworks.

Objective 2: Implement accurate detection models - **Partially Achieved**. While field keypoint detection (99.5% mAP@0.5) and player tracking (86.8% mAP@0.5) perform excellently, ball detection remains challenging due to object size and camera distance limitations.

Objective 3: Create rule violation detection algorithms - **Achieved**. The system successfully detects handball violations and out-of-bounds situations through integrated analysis of pose, position, and field data.

Objective 4: Ensure real-time performance - **Conditionally Achieved**. Real-time processing is possible at 15-20fps through dynamic optimization, though this represents a compromise from the ideal 30fps target.

Objective 5: Maintain cost accessibility - **Achieved**. The €20 hardware cost.

5.1.3 Scientific Contributions

This research advances the state-of-the-art in several dimensions:

- **Custom Football Field Keypoint Dataset:** We created the first publicly available dataset for football field landmark detection, containing 500 annotated images with 27 keypoints each. This addresses a critical gap in sports computer vision resources.
- **Proximity-Based Processing Optimization:** Our dynamic pose estimation activation strategy, triggered only when players are near the ball, represents a novel approach to computational efficiency in sports analysis systems, reducing processing overhead by 65-70%.
- **Integrated Multi-Model Architecture:** The coordinated deployment of three specialized YOLO models demonstrates an effective design pattern for complex sports analysis tasks, balancing accuracy with real-time constraints.
- **Parallax Mitigation Strategy:** Our systematic analysis of parallax effects in dual-camera sports recording and the resulting positioning guidelines contribute practical knowledge for multi-view sports capture systems.
- **Public Dataset Release:** The custom football field keypoint dataset has been made publicly available for the research community, enabling other researchers to build upon this work and advance the field of sports computer vision.

5.1.4 Practical Contributions

Beyond scientific advances, this work delivers tangible benefits to amateur football:

- **Democratization of Sports Technology:** By reducing entry costs by 99% compared to professional systems, we make automated officiating accessible to recreational leagues, school teams, and developing football programs.
- **Proof of Concept for Market Viability:** Our implementation validates the technical feasibility of smartphone-based sports analysis, as evidenced by similar approaches recently launched by established companies like Veo.
- **Open-Source Framework:** The modular architecture and documented implementation provide a foundation for community-driven development of amateur sports technology.

- **Commercial Product Development:** The system has evolved from a research prototype into a viable product, with active commercial interest from companies in the sports technology sector, validating its market potential and practical applicability.

5.1.5 Limitations and Challenges

This research encountered several fundamental limitations that prevented full achievement of initial ambitions:

- **Scope Ambition:** The project's scope encompassing video processing, machine learning, real-time systems, and sports rule implementation proved excessively ambitious for a single thesis. This breadth necessitated compromises in depth, preventing optimization of individual components to their full potential.
- **3D Spatial Analysis:** The most critical technical limitation involves determining ball position when airborne. Using lateral camera views, the system cannot reliably distinguish between a ball's 2D projection and its actual 3D position. This affects both out-of-bounds detection (where an airborne ball might appear outside lines while actually remaining in play) and handball detection (where goalkeeper hand position might be legal in 3D space but appear illegal in 2D projection).
- **Dataset Limitations:** Limited access to diverse amateur football footage restricted model training and evaluation. Professional datasets don't capture the unique challenges of amateur settings (poor lighting, irregular field markings, variable camera quality).
- **Hardware Constraints:** Consumer smartphone cameras, while accessible, lack the resolution and frame rates of professional equipment, fundamentally limiting detection accuracy for small, fast-moving objects like footballs.
- **Evaluation Scope:** Time constraints prevented extensive field testing with actual matches, limiting validation to controlled scenarios and recorded footage.

5.2 Future Work

Future research directions are structured across three complementary dimensions: technical improvements to address current limitations, experimental validation through expanded testing, and social integration within amateur football communities.

5.2.1 Technical Enhancements

1. Ball Depth Estimation: Developing a calibration system to estimate ball distance from camera based on apparent size could partially address the 3D positioning challenge. By correlating ball diameter in pixels with known distances during setup, the system could infer whether balls are airborne or ground-level, improving both out-of-bounds and handball detection accuracy.

2. Temporal Trajectory Analysis: Leveraging ball motion patterns across multiple frames could help disambiguate 3D positions. Balls following parabolic trajectories are likely airborne, while linear movements suggest ground-level motion.

3. Advanced Violation Detection: Extending the system to detect:

- Offside positions using existing player tracking with pass detection logic
- Foul contacts through pose analysis and player proximity patterns
- Throw-in violations via arm trajectory analysis

5.2.2 Experimental Validation

1. Comprehensive Dataset Development:

- Collect 100+ hours of amateur match footage across diverse conditions
- Include various weather conditions, lighting scenarios, and field qualities

2. Comparative Performance Analysis:

- Benchmark against commercial systems
- Evaluate detection accuracy across different smartphone models
- Test system scalability with varying numbers of simultaneous matches

3. Robustness Testing:

- Evaluate performance degradation in adverse conditions (rain, fog, poor lighting)
- Test recovery from temporary occlusions and camera movement
- Assess long-term stability during full 90-minute matches

5.2.3 Social Integration

1. Referee Collaboration Studies:

- Conduct user studies with amateur referees
- Evaluate system acceptance and trust levels
- Identify optimal alert presentation methods
- Develop training materials for system adoption

2. Community Development:

- Release system as open-source project
- Create developer documentation and contribution guidelines
- Establish online community for users and developers
- Organize workshops for clubs interested in adoption

3. Regulatory Engagement:

- Collaborate with local football associations
- Develop guidelines for technology use in amateur matches
- Address concerns about technology replacing human judgment
- Propose frameworks for gradual technology integration

5.2.4 Cross-Sport Adaptation

The modular architecture enables expansion to other sports with minimal modifications:

- **Basketball:** Shot clock violations, out-of-bounds, traveling
- **Handball:** Goal-area violations, passive play detection
- **Field Hockey:** Stick violations, dangerous play
- **Rugby:** Forward passes, offside positions

Each sport would require only rule logic adjustments and potentially sport-specific pose models, leveraging the existing detection and tracking infrastructure.

5.3 Closing Remarks

Despite not achieving a perfect solution, this research represents an important step toward making advanced officiating technology accessible to all levels of football. The gap between professional and amateur sports technology remains vast, and even imperfect automated assistance can significantly improve officiating consistency and fairness.

The challenges encountered particularly 3D spatial analysis from 2D inputs are not unique to this work but represent fundamental computer vision problems that even commercial systems struggle to address. The fact that companies charging €400 annually face similar limitations validates both the difficulty of the problem and the competitiveness of our approach.

Most importantly, this work demonstrates that the democratization of sports technology is not only possible but essential. Perfect should not be the enemy of good the absence of ideal solutions should not prevent us from developing and deploying helpful tools that improve amateur sports. Every incremental advancement in accessibility and affordability brings us closer to a future where all athletes, regardless of economic circumstances, can benefit from technological assistance.

References

Albzeirat, S et al. (2020). "Developing a smart monitoring system for refereeing goal setting in football matches". In: *International Journal of Multidisciplinary Sciences and Advanced Technology* 1, pp. 17–23.

Aleza, Mazi Essoloani and D. Vetrithangam (2023). "Use of Artificial Intelligence to Avoid Errors in Referring a Football Match". In: *2023 International Conference on Artificial Intelligence and Applications (ICAIA) Alliance Technology Conference (ATCON-1)*, pp. 1–6. doi: 10.1109/ICAIA57370.2023.10169463.

Badami, Arik et al. (2018). "Review on Video Refereeing using Computer Vision in Football". In: *2018 IEEE Punecon*, pp. 1–8. doi: 10.1109/PUNECON.2018.8745418.

Barra, Silvio et al. (2023). "FootApp: An AI-powered system for football match annotation". In: *Multimedia Tools and Applications* 82.4, pp. 5547–5567.

Benakesh, A H and R. Rajeev (2024). "Advancing Football Game Analysis: Integrating Computer Vision, Deep Learning, and Hybrid Techniques for Enhanced Video Analytics". In: *2024 OPJU International Technology Conference (OTCON) on Smart Computing for Innovation and Advancement in Industry 4.0*, pp. 1–6. doi: 10.1109/OTCON60325.2024.10688070.

Bi, Ying et al. (2023). "A Survey on Evolutionary Computation for Computer Vision and Image Analysis: Past, Present, and Future Trends". In: *IEEE Transactions on Evolutionary Computation* 27.1, pp. 5–25. doi: 10.1109/TEVC.2022.3220747.

Brocke, Jan vom, Alan Hevner, and Alexander Maedche (2020). "Introduction to Design Science Research". In: *Design Science Research. Cases*. Ed. by Jan vom Brocke, Alan Hevner, and Alexander Maedche. Cham: Springer International Publishing, pp. 1–13. isbn: 978-3-030-46781-4. doi: 10.1007/978-3-030-46781-4_1. url: https://doi.org/10.1007/978-3-030-46781-4_1.

Carlos, Lago-Peñas, Rey Ezequiel, and Kalén Anton (2019). "How does Video Assistant Referee (VAR) modify the game in elite soccer?" In: *International Journal of Performance Analysis in Sport* 19.4, pp. 646–653.

Chopra, Harshit, Sona Mundody, and Ram Mohana Reddy Guddeti (2023). "A Key-frame Extraction for Object Detection and Human Action Recognition in Soccer Game Videos". In: *2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–7. doi: 10.1109/ICCCNT56998.2023.10308225.

D'Orazio, Tiziana and Marco Leo (2010). "A review of vision-based systems for soccer video analysis". In: *Pattern recognition* 43.8, pp. 2911–2926.

Deliege, Adrien et al. (2021). "Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos". In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4508–4519.

Di Salvo Valter Collins Adam, McNeill Barry and Cardinale Marco (2006). "Validation of Prozone ®: A new video-based performance analysis system". In: *International Journal of Performance Analysis in Sport* 6.1, pp. 108–119. doi: 10.1080/24748668.2006.11868359.

Ding, Jing et al. (2024). "Research on Human Posture Estimation Algorithm Based on YOLO-Pose". In: *Sensors* 24.3036.

Diop, Charles-Alexandre et al. (2022). "Soccer Player Recognition using Artificial Intelligence and Computer Vision". In: *2022 IEEE International Conference on Electro Information Technology (eIT)*, pp. 477–481. doi: 10.1109/eIT53891.2022.9813788.

Dresch, Aline, Daniel Pacheco Lacerda, and José Antônio Valle Antunes (2015). "Design Science Research". In: *Design Science Research: A Method for Science and Technology Advancement*. Cham: Springer International Publishing, pp. 67–102. isbn: 978-3-319-07374-3. doi: 10.1007/978-3-319-07374-3_4. url: https://doi.org/10.1007/978-3-319-07374-3_4.

Girshick, Ross (2015). "Fast R-CNN". In: *ICCV*.

Girshick, Ross et al. (2014). "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation". In: *CVPR*.

Grotenhuis, Niels (2022). "Automatic Player Tracking in Sports Videos". In: *Sports Analytics Review*.

Gurau, Tudor Vladimir et al. (2023). "Epidemiology of injuries in professional and amateur football men (part II)". In: *Journal of clinical medicine* 12.19, p. 6293.

Held, Jan, Anthony Cioppa, et al. (2024). "Towards AI-Powered Video Assistant Referee System (VARS) for Association Football". In: *arXiv preprint arXiv:2407.12483*.

Held, Jan, Hani Itani, et al. (2024). "X-VARS: Introducing Explainability in Football Refereeing with Multi-Modal Large Language Models". In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 3267–3279. doi: 10.1109/CVPRW63382.2024.00332.

Holder, Ulrike, Thomas Ehrmann, and Arne König (2022). "Monitoring experts: insights from the introduction of video assistant referee (VAR) in elite football". In: *Journal of Business Economics* 92.2, pp. 285–308.

Jegham, Nidhal et al. (2025). *YOLO Evolution: A Comprehensive Benchmark and Architectural Review of YOLOv12, YOLO11, and Previous Versions*. arXiv preprint.

Joshan Athanesious, J. and S. Kiruthika (2024). "Perspective Transform Based YOLO With Weighted Intersect Fusion for Forecasting the Possession Sequence of the Live Football Game". In: *IEEE Access* 12, pp. 75542–75558. doi: 10.1109/ACCESS.2024.3402370.

Maji, S. et al. (2022). "YOLO-Pose: Enhancing YOLO for Multi-Person Pose Estimation Using Object Keypoint Similarity". In: *CVPR Workshops*.

Mavrogiannis, Panagiotis and Ilias Maglogiannis (2022). "Amateur football analytics using computer vision". In: *Neural Computing and Applications* 34.22, pp. 19639–19654.

Nasser, Mona et al. (2012). "Generalizability of systematic reviews of the effectiveness of health care interventions to primary health care: concepts, methods and future research". In: *Family practice* 29.suppl_1, pp. i94–i103.

Nguyen, Trong-Thuan and Minh-Triet Tran (2023). "A Transformer-based Approach for Dynamic Referee Assistance". In: *2023 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, pp. 1–6. doi: 10.1109/MAPR59823.2023.10289042.

O'Shea, Keiron and Ryan Nash (2015). "An Introduction to Convolutional Neural Networks". In: *arXiv preprint arXiv:1511.08458*.

Öberg, Filip (2021). *Football analysis using machine learning and computer vision*.

Okoli, Chitu and Kira Schabram (2015). "A guide to conducting a systematic literature review of information systems research". In.

Page, Matthew J, Joanne E McKenzie, et al. (2021). "Updating guidance for reporting systematic reviews: development of the PRISMA 2020 statement". In: *Journal of clinical epidemiology* 134, pp. 103–112.

Page, Matthew J, David Moher, et al. (2021). "PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews". In: *bmj* 372.

Pandey, Ananya and Achyutananda Mishra (2024). "Application of Artificial Intelligence in Sports Analytics: Analysing the Ethical and Legal Perspectives". In: *Sports Analytics: Data-Driven Sports and Decision Intelligence*. Ed. by A. Mansurali et al. Cham: Springer Nature Switzerland, pp. 163–184. isbn: 978-3-031-63573-1. doi: 10.1007/978-3-031-63573-1_10. url: https://doi.org/10.1007/978-3-031-63573-1_10.

Promvijittrakarn, Phokin and Theekapun Charoenpong (2023). "Soccer Player Detection and Soccer Team Classification using GoogLeNet and Histogram Analysis". In: *2023 15th Biomedical Engineering International Conference (BMEiCON)*, pp. 1–4. doi: 10.1109/BMEiCON60347.2023.10321826.

Rashid, Fadilla Atyka Nor and Siaw-Hong Liew (2022). "Footballer Detection on Position Based Classification Recognition using Deep Learning Approach". In: *2022 International Conference on Green Energy, Computing and Sustainable Technology (GECOST)*, pp. 193–197. doi: 10.1109/GECOST55694.2022.10010650.

Ren, Shaoqing et al. (2015). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks". In: *NeurIPS*.

Sapkota, Ranjan, Marco Flores-Calero, Rizwan Qureshi, et al. (2025). "YOLO advances to its genesis: a decadal and comprehensive review of the YOLO series". In: *Artificial Intelligence Review*.

Scott, Atom et al. (2024). "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*, pp. 3357–3366.

Sepehri, Yamin et al. (2023). "Privacy-Preserving Image Acquisition for Neural Vision Systems". In: *IEEE Transactions on Multimedia* 25, pp. 6232–6244. doi: 10.1109/TMM.2022.3207018.

Shao, Yuan and Zaihong He (2025). "VAR-YOLOv8s: IoT-based automatic foul detection in soccer matches". In: *Alexandria Engineering Journal* 111, pp. 555–565.

Song, Hyun-Ki (2022). "Player Interaction Detection in Sports". In: *Computer Vision and Sports*.

Theagarajan, Rajkumar and Bir Bhanu (2021). "An Automated System for Generating Tactical Performance Statistics for Individual Soccer Players From Videos". In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.2, pp. 632–646. doi: 10.1109/TCSVT.2020.2982580.

Ultralytics (2024). *YOLOv11 Model Documentation*. https://www.ultralytics.com/blog/how-to-use-ultralytics-yolo11-for-pose-estimation.

V, Prasanth V and Nallavan G (2024). "A Review of Deep Learning Architectures for Automated Video Analysis in Football Events". In: *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–6. doi: 10.1109/ICCCNT61001.2024.10726174.

Vassar, Matt et al. (2017). "Database selection in systematic reviews: an insight through clinical neurology". In: *Health Information & Libraries Journal* 34.2, pp. 156–164.

Vater, Christian and Ernst-Joachim Hossner (2024). "That was a foul! How viewing angles influence football referees' decision-making". In: *Psychology of Sport and Exercise* 72, p. 102558. doi: 10.1016/j.psychsport.2024.102558.

Vidal-Codina, Ferran et al. (2022). "Automatic event detection in football using tracking data". In: *Sports Engineering* 25.1, p. 18.

Wang, Jing and Baiqing Liu (2023). "Analyzing the feature extraction of football player's offense action using machine vision, big data, and internet of things: J. Wang, B. Liu". In: *Soft Computing* 27.15, pp. 10905–10920.

Xu, Jingchao et al. (2021). "Real-time detection of game handball foul based on target detection and skeleton extraction". In: *2021 IEEE International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, pp. 41–46. doi: 10.1109/CEI52496.2021.9574504.

Yin, Lu et al. (2023). "A survey of video-based human action recognition in team sports". In: *Pattern Recognition* 138, p. 109445. doi: 10.1016/j.patcog.2023.109445.

YOLO (2025). *YOLOv12 Model Documentation Release Notes*. https://docs.ultralytics.com/pt/models/yolo12/.

Zhang, Shubing (2022). "The Physical Fitness Video Tracking System of Football Players Based on Artificial Intelligence Algorithm". In: *2022 International Conference on Artificial Intelligence and Autonomous Robot Systems (AIARS)*, pp. 131–134. doi: 10.1109/AIARS57204.2022.00037.