



**UNIVERSITY OF PIRAEUS**

**SCHOOL OF INFORMATION AND COMMUNICATION TECHNOLOGIES**

**DEPARTMENT OF DIGITAL SYSTEMS**

**MASTER OF SCIENCE**

**“INFORMATION SYSTEMS AND SERVICES”**

**Recognition of Basketball Referee Signals Using Machine Learning  
and Computer Vision**

By

**Ioannis Georgopoulos**

Submitted

In partial fulfilment of the requirements for the degree of  
Master of Science in “Information Systems & Services” in the specialization of  
“Big Data & Analytics”

at the

**UNIVERSITY OF PIRAEUS**

**MARCH 2026**

Thesis Supervisor: Ilias G. Maglogiannis

Title: Professor

University of Piraeus,. All rights reserved.

Author: Ioannis Georgopoulos ME2204

## ΣΕΛΙΔΑ ΕΓΚΥΡΟΤΗΤΑΣ

**Ονοματεπώνυμο Φοιτητή/Φοιτήτριας:** Ιωάννης Γεωργόπουλος

**Τίτλος Μεταπτυχιακής Διπλωματικής Εργασίας:** Αναγνώριση Σημάτων Διαιτητών

Καλαθοσφαίρισης με Χρήση Μηχανικής Μάθησης και Επεξεργασία Εικόνας

*Η παρούσα Μεταπτυχιακή Διπλωματική Εργασία υποβάλλεται ως μερική εκπλήρωση των απαιτήσεων του Προγράμματος Μεταπτυχιακών Σπουδών “Πληροφορικά Συστήματα & Υπηρεσίες” του Τμήματος Ψηφιακών Συστημάτων του Πανεπιστημίου Πειραιώς και εγκρίθηκε στις ..... [ημερομηνία έγκρισης] από τα μέλη της Εξεταστικής Επιτροπής.*

### Εξεταστική Επιτροπή

Επιβλέπων (Τμήμα Ψηφιακών Συστημάτων, Πανεπιστήμιο Πειραιώς) Καθηγητής Ηλίας Μαγκλογιάννης

Μέλος Εξεταστικής Επιτροπής:

Μέλος Εξεταστικής Επιτροπής:

### ΥΠΕΥΘΥΝΗ ΔΗΛΩΣΗ ΑΥΘΕΝΤΙΚΟΤΗΤΑΣ

*Ο Ιωάννης Γεωργόπουλος, γνωρίζοντας τις συνέπειες της λογοκλοπής, δηλώνω υπεύθυνα ότι η παρούσα εργασία με τίτλο «Αναγνώριση Σημάτων Διαιτητών Καλαθοσφαίρισης με Χρήση Μηχανικής Μάθησης και Επεξεργασία Εικόνας», αποτελεί προϊόν αυστηρά προσωπικής εργασίας και όλες οι πηγές που έχω χρησιμοποιήσει, έχουν δηλωθεί κατάλληλα στις βιβλιογραφικές παραπομπές και αναφορές. Τα σημεία όπου έχω χρησιμοποιήσει ιδέες, κείμενο ή/και πηγές άλλων συγγραφέων, αναφέρονται ευδιάκριτα στο κείμενο με την κατάλληλη παραπομπή και η σχετική αναφορά περιλαμβάνεται στο τμήμα των βιβλιογραφικών αναφορών με πλήρη περιγραφή.*

*Επιπλέον δηλώνω υπεύθυνα ότι η συγκεκριμένη Μεταπτυχιακή Διπλωματική Εργασία έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει αξιολογηθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό.*

*Παράβαση της ανωτέρω ακαδημαϊκής μου ευθύνης αποτελεί ουσιώδη λόγο για την ανάκληση του πτυχίου μου. Σε κάθε περίπτωση, αναληθούς ή ανακριβούς δηλώσεως, υπόκειμαι στις συνέπειες που προβλέπονται τις διατάξεις που προβλέπει η Ελληνική και Κοινοτική Νομοθεσία περί πνευματικής ιδιοκτησίας.*

**Ο/Η ΔΗΛΩΝ/ΟΥΣΑ**

**Ονοματεπώνυμο:** Ιωάννης Γεωργόπουλος

**Αριθμός Μητρώου:** ME2204

## ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Professor Ilias Maglogiannis, for his guidance, valuable feedback, and support throughout the development of this thesis.

I would also like to thank referee Rachiotis for his willingness to share his knowledge and insights regarding basketball referee signals, which proved valuable in understanding the practical aspects of the problem addressed in this work.

Finally, I would like to thank my family and friends for their continuous support and encouragement during my studies.

## TABLE OF CONTENTS

|                                                                                      |    |
|--------------------------------------------------------------------------------------|----|
| ACKNOWLEDGEMENTS .....                                                               | 3  |
| TABLE OF CONTENTS.....                                                               | 4  |
| TABLE OF FIGURES .....                                                               | 7  |
| LIST OF TABLES.....                                                                  | 8  |
| ΠΕΡΙΛΗΨΗ .....                                                                       | 9  |
| ABSTRACT .....                                                                       | 10 |
| 1 INTRODUCTION.....                                                                  | 11 |
| 1.1 Motivation .....                                                                 | 11 |
| 1.2 Problem Statement.....                                                           | 11 |
| 1.3 Research Questions .....                                                         | 12 |
| 1.4 Contributions .....                                                              | 12 |
| 1.5 Thesis Structure .....                                                           | 13 |
| 2 BACKGROUND & LITERATURE REVIEW .....                                               | 14 |
| 2.1 Computer Vision for Human-Centric Understanding in Sports .....                  | 14 |
| 2.2 Gesture Recognition: Modalities and Problem Types .....                          | 15 |
| 2.2.1 Vision-Based and Sensor-Based Approaches.....                                  | 15 |
| 2.2.2 Static and Dynamic Gesture Recognition.....                                    | 16 |
| 2.2.3 Fine Grained Pose Dependent Recognition .....                                  | 16 |
| 2.3 Referee Signal Recognition in Basketball: Existing Approaches and Limitations... | 17 |
| 2.4 Pose Estimation and Fine-Grained Visual Representation .....                     | 18 |
| 2.5 Dataset Bias, Shortcut Learning, and Domain Shift.....                           | 19 |
| 2.6 Summary and Research Gap.....                                                    | 20 |
| 3 PROBLEM DEFINITION & THEORETICAL ANALYSIS .....                                    | 22 |
| 3.1 Problem Statement.....                                                           | 22 |
| 3.2 Referee Signals as a Visual Recognition Task.....                                | 22 |
| 3.3 Broadcast Video Bias and Contextual Leakage .....                                | 23 |
| 3.4 Limitations of Image Only Classification Approaches.....                         | 24 |
| 3.5 Motivation for a Referee-Centric Pipeline .....                                  | 25 |
| 3.6 Decomposition into Consecutive Subtasks.....                                     | 25 |

|       |                                                                       |    |
|-------|-----------------------------------------------------------------------|----|
| 3.7   | Questions and Hypotheses .....                                        | 26 |
| 3.8   | Summary.....                                                          | 26 |
| 4     | SYSTEM ARCHITECTURE & TOOLS .....                                     | 27 |
| 4.1   | Overall Pipeline Architecture .....                                   | 27 |
| 4.2   | Design Logic for a Multi-Stage Approach .....                         | 28 |
| 4.3   | Video Processing and Frame Extraction .....                           | 29 |
| 4.4   | Model Integration and Data Flow Between Stage.....                    | 29 |
| 4.5   | Tools, Frameworks, and Implementation Environment .....               | 30 |
| 4.6   | System Limitation and Practical Considerations.....                   | 31 |
| 4.7   | Summary of Architectural Decisions .....                              | 32 |
| 5     | EXPERIMENTAL METHODOLOGY .....                                        | 33 |
| 5.1   | Alternative Approach: Broadcast-Based Classification (Roboflow) ..... | 33 |
| 5.2   | Data Source Selection & Constraints .....                             | 34 |
| 5.3   | Model 1: Referee Detection.....                                       | 35 |
| 5.4   | Model 2: Signal / No-Signal Classification .....                      | 36 |
| 5.5   | Model 3: Signal Type Classification.....                              | 37 |
| 6     | EXPERIMENTAL RESULTS .....                                            | 39 |
| 6.1   | Evaluation Setup and Metrics .....                                    | 39 |
| 6.2   | Broadcast-Based Classification (Alternative Approach) .....           | 40 |
| 6.2.1 | Data Used.....                                                        | 40 |
| 6.2.2 | Preprocessing and Augmentation.....                                   | 43 |
| 6.2.3 | Model Configuration .....                                             | 44 |
| 6.2.4 | Quantitative Results.....                                             | 44 |
| 6.2.5 | Qualitative Analysis.....                                             | 46 |
| 6.2.6 | Key Takeaway .....                                                    | 47 |
| 6.3   | Model 1 — Referee Detection .....                                     | 49 |
| 6.3.1 | Data Used.....                                                        | 49 |
| 6.3.2 | Preprocessing and Augmentation.....                                   | 49 |
| 6.3.3 | Model Configuration .....                                             | 50 |
| 6.3.4 | Quantitative Results.....                                             | 51 |
| 6.3.5 | Qualitative Results .....                                             | 52 |
| 6.3.6 | Key Takeaway .....                                                    | 54 |

|       |                                                           |    |
|-------|-----------------------------------------------------------|----|
| 6.4   | Model 2 — Signal / No-Signal Classification .....         | 54 |
| 6.4.1 | Data Used .....                                           | 54 |
| 6.4.2 | Preprocessing and Augmentation .....                      | 55 |
| 6.4.3 | Model Configuration .....                                 | 56 |
| 6.4.4 | Quantitative Results .....                                | 57 |
| 6.4.5 | Qualitative Evaluation .....                              | 58 |
| 6.4.6 | Key Takeaway .....                                        | 59 |
| 6.5   | Model 3 — Signal Type Classification .....                | 60 |
| 6.5.1 | Initial Attempt .....                                     | 60 |
| 6.5.2 | Dataset Expansion and Class Grouping Strategy .....       | 61 |
| 6.5.3 | Preprocessing and Augmentation .....                      | 63 |
| 6.5.4 | Model Configuration .....                                 | 64 |
| 6.5.5 | Quantitative Results .....                                | 64 |
| 6.5.6 | Qualitative Evaluation .....                              | 67 |
| 6.5.7 | Key Takeaway .....                                        | 68 |
| 6.6   | End-to-End Pipeline Evaluation on Video Input .....       | 69 |
| 6.6.1 | End-to-End Pipeline Workflow .....                        | 69 |
| 6.6.2 | Parameter Selection and Tuning .....                      | 71 |
| 6.6.3 | Temporal Aggregation and Event Logic .....                | 73 |
| 6.6.4 | End-to-End Qualitative Demonstration .....                | 74 |
| 6.6.5 | Practical Performance Considerations .....                | 76 |
| 6.6.6 | Quantitative Evaluation of the Integrated Pipeline .....  | 78 |
| 6.6.7 | Key Takeaways .....                                       | 80 |
| 7     | DISCUSSION AND CONCLUSIONS .....                          | 82 |
| 7.1   | Overview of the Proposed Pipeline and Main Findings ..... | 82 |
| 7.2   | Methodological Insights .....                             | 83 |
| 7.3   | Limitations .....                                         | 84 |
| 7.4   | Future Work .....                                         | 85 |
| 7.5   | Final Conclusion .....                                    | 86 |
|       | REFERENCES .....                                          | 87 |

## TABLE OF FIGURES

|                                                                                                                                                                                                                            |    |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| Figure 4.1: Overview of the proposed multi-stage pipeline .....                                                                                                                                                            | 27 |
| Figure 6.1: Representative broadcast images used for direct referee signal classification (Roboflow, 2024). .....                                                                                                          | 41 |
| Figure 6.2: Normalized confusion matrix for the broadcast based referee signal classification model evaluated on the test set.. .....                                                                                      | 45 |
| Figure 6.3: Representative qualitative examples from the broadcast based referee signal classification experiment. The top row shows correctly classified instances. The bottom row illustrates typical failure cases..... | 46 |
| Figure 6.4: Sensitivity heatmaps for representative broadcast based classification predictions. For each example, the original broadcast frame is shown alongside the corresponding sensitivity map. ....                  | 48 |
| Figure 6.5: Precision–recall curve for the referee detection model evaluated on the validation set...                                                                                                                      | 52 |
| Figure 6.6: Representative referee detection results on previously unseen EuroLeague game footage. ....                                                                                                                    | 53 |
| Figure 6.7: Confusion matrix for the revised binary signal classifier evaluated on the validation set. .                                                                                                                   | 57 |
| Figure 6.8: Representative qualitative examples from the revised binary signal classifier.....                                                                                                                             | 59 |
| Figure 6.9: Confusion matrix for the multiclass signal classifier evaluated on the validation set. ....                                                                                                                    | 65 |
| Figure 6.10: Confusion matrix for the multiclass signal classifier evaluated on the test set. ....                                                                                                                         | 66 |
| Figure 6.11: qualitative examples from the grouped multiclass signal classifier .....                                                                                                                                      | 67 |
| Figure 6.12: representative frames from the fully integrated pipeline.....                                                                                                                                                 | 75 |
| Figure 6.13: Event level confusion matrix of the integrated pipeline .....                                                                                                                                                 | 79 |

## LIST OF TABLES

|                                                                                                                             |    |
|-----------------------------------------------------------------------------------------------------------------------------|----|
| Table 6-1: Initial class distribution of the broadcast based training dataset prior to preprocessing (Roboflow, 2024) ..... | 42 |
| Table 6-2: Final class distribution of the broadcast based training dataset after preprocessing.....                        | 43 |
| Table 6-3: Referee signal class distribution .....                                                                          | 60 |
| Table 6-4: Distribution of referee signal instances extracted from eight games prior to grouping. ....                      | 62 |
| Table 6-5: Final distribution of grouped referee signal categories across dataset splits. ....                              | 63 |
| Table 6-6: Tunable parameters used .....                                                                                    | 71 |
| Table 6-7: Event level evaluation.....                                                                                      | 79 |

## ΠΕΡΙΛΗΨΗ

Η παρούσα διπλωματική εργασία ερευνά το πρόβλημα της αυτόματης αναγνώρισης σημάτων διαιτητών καλαθοσφαίρισης από τηλεοπτικό υλικό αγώνων. Σε αντίθεση με τις κλασικές έρευνες αναγνώρισης αντικειμένων, τα σήματα των διαιτητών εκφράζονται μέσω μικρών διαφοροποιήσεων στη στάση του σώματος και στη διάταξη των άκρων, γεγονός που καθιστά την αναγνώρισή τους ιδιαίτερα απαιτητική σε ρεαλιστικές συνθήκες μετάδοσης, όπου παρατηρούνται διαφορετικές γωνίες κάμερας, διάφορες παρεμβολές, μέτρια ανάλυση μετάδοσης, μεταβολές της οπτικής κλίμακας και ισχυρή επιρροή του γενικότερου οπτικού πλαισίου. Για την αντιμετώπιση των προκλήσεων αυτών προτείνεται μια αρχιτεκτονική τριών σταδίων, με επίκεντρο τον διαιτητή. Η προσέγγισή της εργασίας, είναι η διάσπαση του προβλήματος σε τρία διαδοχικά στάδια. Αρχικά, ο εντοπισμός των διαιτητών σε πλήρη καρτέ μετάδοσης. Στη συνέχεια, δυαδική ταξινόμηση για διάκριση μεταξύ σήματος και μη σήματος, ενώ τέλος, αναγνώριση και ταξινόμηση του συγκεκριμένου τύπου σήματος. Η προσέγγιση αξιολογείται σε σύγκριση με ένα μοντέλο άμεσης ταξινόμησης ολόκληρων καρτέ, αποδεικνύοντας ότι το γενικότερο οπτικό πλαίσιο ενός αγώνα μπορεί να οδηγήσει σε φαινομενικά υψηλή ακρίβεια, αλλά με περιορισμένη ικανότητα γενίκευσης. Κατασκευάστηκαν προσαρμοσμένα σύνολα δεδομένων από πλήρεις αγώνες υπό ρεαλιστικές συνθήκες μετάδοσης. Τα πειραματικά αποτελέσματα δείχνουν ότι η αρχιτεκτονική εντοπισμού των διαιτητών βελτιώνει την ανθεκτικότητα του συστήματος, ενώ το μέγεθος του συνόλου δεδομένων, η ισορροπία μεταξύ των κλάσεων και οι στρατηγικές αύξησης του συνόλου δεδομένων, επηρεάζουν καθοριστικά τη δυνατότητα γενίκευσης. Επιπλέον, ενσωματώνονται μηχανισμοί χρονικής εξομάλυνσης για τη βελτίωση της σταθερότητας των προβλέψεων σε βίντεο, χωρίς την ανάγκη υπολογιστικά απαιτητικών μοντέλων μάθησης σε επίπεδο βίντεο.

## ABSTRACT

This thesis investigates the problem of automatic basketball referee signal recognition from broadcast video footage. Unlike conventional object recognition tasks, referee signals are expressed through subtle variations in human pose and hands configuration, making them challenging to detect under realistic broadcast conditions characterized by camera motion, resolution, occlusions, scale variation, and contextual bias. To address these challenges, a referee centric, multistage computer vision pipeline is proposed. The system decomposes the task into three sequential components. First, referee detection from full broadcast frames. Second, binary classification to distinguish signal from non-signal gestures and third, multiclass classification to recognize specific signal types. The approach is evaluated against a direct broadcast frame classification baseline, demonstrating that context driven models achieve high apparent accuracy but fail to generalize reliably. Custom datasets were constructed from full game recordings under realistic conditions. Experimental results show that detection-first architectures improve robustness, while dataset scale, class balance, and augmentation strategy critically influence generalization performance. Furthermore, lightweight temporal aggregation mechanisms are integrated to enhance prediction stability during video inference without requiring computationally intensive end to end video models. The findings highlight the importance of architectural decomposition and domain dataset construction in gesture recognition systems operating in real world broadcast environments. The proposed framework provides both methodological insights and practical guidelines for developing structured recognition systems in sports analytics.

# 1 Introduction

The increasing integration of artificial intelligence into sports analytics has enabled automated analysis of complex in game events. Among these, referee hand signals constitute a structured and visually precise form of communication that carries critical decisions during basketball matches. This thesis investigates the feasibility of automatically detecting and classifying such signals from broadcast video, addressing both methodological and practical challenges inherent in real world game footage.

## 1.1 Motivation

Recent advances in computer vision have significantly expanded the scope of automated sports analysis, enabling the detection of players, ball trajectories, tactical formations, and other events, directly from video data. Most existing systems, however, focus primarily on players or objects analysis, while limited attention has been given to the automated interpretation of referee gestures. In basketball, referee signals play a critical role in communicating fouls, violations, scoring decisions, and administrative actions. These gestures are formally standardized, visually structured, and consistently used across professional competitions, making them a compelling target for machine learning image recognition (FIBA- International Basketball Federation, 2024).

Despite their structured nature, referee signals present several challenges for automated recognition. Unlike conventional object recognition tasks, the discriminative information is included in minimal variations of human pose, arm configuration, and relative limb positioning.

From a practical perspective, automated referee signal recognition could enhance sports analytics pipelines by enabling structured event logging, improving game annotation, and supporting performance analysis systems. Furthermore, such a system could operate as a component within larger automated officiating. However, achieving robust performance under realistic broadcast conditions requires architectural decisions that account for domain limitations, few annotated data, and generalization challenges.

Motivated by these considerations, this thesis explores a structured, referee centric approach to signal recognition that decomposes the problem into detection, binary classification, and multiclass classification stages. The objective is not only to assess whether referee signals can be recognized from broadcast footage, but also to investigate how pipeline design choices influence robustness and generalization in a real world setting.

## 1.2 Problem Statement

The objective of this thesis is to develop and evaluate a computer vision system capable of recognizing basketball referee hand signals directly from broadcast video footage. The task requires

identifying referees within full game frames and classifying their gestures under realistic conditions, including camera motion, occlusions, multiple people, and limited signals.

A direct classification of full broadcast frames is insufficient due to strong contextual bias and the small visual footprint of referee gestures. Therefore, the problem is formulated as a structured pipeline consisting of referee detection, binary signal discrimination (signal vs. no signal), and multiclass signal classification. The central challenge lies in achieving reliable generalization across games while operating under practical data and hardware constraints.

### 1.3 Research Questions

To investigate the feasibility of automated referee signal recognition from broadcast basketball footage, this thesis addresses the following research questions:

1. Can referee hand signals be reliably recognized from standard broadcast video using a structured multi stage computer vision pipeline?
2. Does a referee centric approach improves robustness and generalization compared to direct broadcast frame classification?
3. How do dataset scale, class balance, and augmentation strategy influence the performance and generalization of binary and multiclass signal classifiers?
4. Can lightweight temporal aggregation mechanisms enhance prediction stability without requiring fully video based deep learning models?

These questions guide the methodological design, experimental evaluation, and analytical discussion presented in the subsequent chapters.

### 1.4 Contributions

This thesis contributes a structured approach to the problem of basketball referee signal recognition from broadcast video. First, it proposes and implements a referee centric, multi stage computer vision pipeline that decomposes the task into detection, binary signal discrimination, and multiclass classification. This architectural choice is motivated by the limitations of direct broadcast frame classification and is evaluated throughout the experimental analysis.

Second, the thesis introduces custom datasets derived from full game recordings, including annotated referee detection data and cropped referee images for downstream classification tasks. The process emphasizes realistic broadcast conditions and practical data limitations rather than controlled laboratory settings.

Third, an empirical comparison is conducted between direct broadcast classification and the proposed detection first approach, highlighting the impact of contextual bias and domain shift on generalization performance. In addition, the study analyzes how dataset scale, class imbalance, and

augmentation strategies influence both binary and multiclass recognition results. Finally, lightweight temporal aggregation mechanisms are integrated into the inference pipeline to improve prediction stability in video scenarios without relying on fully end to end video learning architectures.

## 1.5 Thesis Structure

The remainder of this thesis is organized as follows:

- Chapter 2 presents the theoretical background and reviews related work in gesture recognition, human pose based analysis, and sports video understanding.
- Chapter 3 analyzes the problem domain and introduces the conceptual design choices underlying the proposed approach.
- Chapter 4 describes the overall system architecture, implementation environment, and technical tools used throughout the project.
- Chapter 5 details the dataset construction process, including data collection, preprocessing, annotation strategy, and augmentation procedures.
- Chapter 6 presents the experimental setup and results for each stage of the proposed pipeline, including detection, binary classification, multiclass classification, and integrated inference.
- Chapter 7 discusses the findings, limitations, and implications of the results, and outlines potential directions for future research.

This structure reflects the progression from theoretical grounding to practical implementation and critical evaluation of the proposed system.

## 2 Background & Literature Review

### 2.1 Computer Vision for Human-Centric Understanding in Sports

Computer vision has increasingly been applied to the automated analysis of sports events, enabling tasks such as player tracking, action recognition, event detection, and tactical analysis. Advances in deep learning have significantly improved the robustness and scalability of such systems, particularly in broadcast scenarios where large volumes of visual data are available. Within this context, sports analytics has emerged as a subfield that leverages visual data to extract structured information from complex, dynamic environments.

A central theme in modern computer vision is human centric visual understanding, which focuses on interpreting human actions, gestures, and interactions from images and videos. Human activity analysis encompasses multiple levels of abstraction, ranging from low level motion patterns to high level activities (Aggarwal & Ryoo, 2011). Gestures constitute a fine grained category within this taxonomy, typically characterized by structured limb configurations and temporally bounded movements that carry specific semantic meaning (Aggarwal & Ryoo, 2011).

In sports environments, most vision systems concentrate on players, ball trajectories, and scoring events, while comparatively less attention has been given to officials. Referee signals, however, represent a structured and standardized form of human gesture communication that directly encodes decisions affecting game flow. Unlike generic action recognition tasks, referee signals are intentionally performed and visually distinctive, yet they occur in dynamically changing environments.

The recognition of such signals introduces several challenges typical of human centered visual understanding. Broadcast footage often includes varying camera angles, occlusions, motion blur, illumination changes, and background distractions from spectators and players. Furthermore, the discriminative information for referee signals is embedded primarily in upper body pose and limb configuration rather than in global scene context (Moeslund & Granum, 2001). As a result, successful recognition requires models capable of capturing visual cues while maintaining robustness to environmental variability.

From a methodological perspective, human activity recognition has evolved from handcrafted feature based approaches to deep learning architectures capable of learning hierarchical visual representations directly from data (Poppe, 2010). Traditional approaches often relied on motion descriptors, silhouettes, or handcrafted features, while modern systems employ convolutional neural networks and, in video settings, spatiotemporal architectures (Poppe, 2010) (Tran, Bourdev, Fergus, Torresani, & Paluri). Despite these advances, real world deployment in sports settings remains challenging due to domain variability and limited labeled data for specialized tasks such as referee signal interpretation (Torralba & Efros, 2011).

In this thesis, referee signal recognition is framed as a human centric visual recognition problem within the broader domain of sports analytics. The focus is not on generic action understanding, but on structured gesture interpretation under broadcast conditions. This framing establishes the conceptual foundation for the following discussion of gesture recognition methodologies and the architectural decisions adopted in the proposed multi stage pipeline.

## 2.2 Gesture Recognition: Modalities and Problem Types

Gesture recognition has been extensively studied within the broader domain of human activity analysis and human computer interaction. A gesture can be defined as a structured movement or posture of the body, typically involving the hands and arms, that conveys semantic meaning. Within the taxonomy of human activity understanding, gestures are generally considered fine grained, temporally bounded actions characterized by specific limb configurations and motion patterns (Aggarwal & Ryoo, 2011).

### 2.2.1 Vision-Based and Sensor-Based Approaches

Two principal options dominate gesture recognition research. Sensor based approaches and vision based approaches.

Sensor based systems rely on wearable devices such as accelerometers, gyroscopes, or electromyography (EMG) sensors attached to the user (Yeh, Pan, & Hu, 2017). These methods capture motion signals directly from the body and are often robust to visual occlusion and illumination changes. For example, (Yeh, Pan, & Hu, 2017) proposed a sensor based system for basketball referee signal recognition using surface electromyography and accelerometer data combined with Deep Belief Networks, achieving high classification accuracy under controlled conditions. Similarly, wearable device approaches have also been explored using accelerometer data and sequence alignment methods such as Dynamic Time Warping (Bedico, et al., 2020).

However, sensor-based approaches require instrumentation of the subject, which limits scalability and real world deployment in professional sports settings. In official basketball games, referees are not equipped with wearable sensing systems designed for gesture recognition. Consequently, sensor solutions, while technically effective, are often impractical for broadcast based analysis.

Vision based approaches, in contrast, rely on images or video captured by cameras. These systems are compatible with existing broadcast infrastructures. Vision based gesture recognition typically involves a sequence of steps including region of interest extraction, feature representation, and classification. Earlier approaches relied on handcrafted features such as Histograms of Oriented Gradients (HOG), Local Binary Patterns (LBP), or silhouette based descriptors, combined with classifiers

such as Support Vector Machines (SVM) or Random Forests (Žemgulys J. , Raudonis, Maskeliūnas, & Damaševičius, 2018). More recent work has adopted deep learning architectures capable of learning discriminative visual representations directly from raw image data (Žemgulys J. , Raudonis, Maskeliūnas, & Damaševičius, 2020) (Wang, HoKun, & Zhou, 2024).

In the context of sports, vision based approaches must operate under challenging conditions including dynamic camera motion, varying viewpoints, partial occlusion, and background noise. These constraints are particularly relevant for referee signal recognition, where gestures occur in fast paced environments and are often tied within complex visual scenes.

### 2.2.2 Static and Dynamic Gesture Recognition

Gesture recognition problems can also be categorized as static or dynamic (Poppe, 2010). Static gestures correspond to single, distinguishable body postures maintained for a short duration, whereas dynamic gestures involve temporally ordered sequences of movements. Many sign language recognition systems, for example, rely on modeling temporal dependencies across frames using sequential models such as Hidden Markov Models or recurrent neural networks (Tran, Bourdev, Fergus, Torresani, & Paluri).

In sports officiating, referee signals often contain both static and dynamic components. A signal may begin with a transitional motion and lead to a clearly defined final posture that conveys the semantic decision. While fully temporal modeling can capture motion trajectories, it also increases annotation complexity and computational requirements. Consequently, a number of referee signal recognition studies have approached the problem using frame based or short sequence representations extracted from video streams (Žemgulys J. , Raudonis, Maskeliūnas, & Damaševičius, 2018) (Žemgulys J. , Raudonis, Maskeliūnas, & Damaševičius, 2020).

The distinction between static and dynamic modeling is therefore not binary but contextual. In broadcast settings, where camera viewpoints are not optimized for motion capture and temporal annotations are costly, frame level recognition combined with system level temporal aggregation constitutes a practical alternative to end to end video modeling (Torralba & Efros, 2011).

### 2.2.3 Fine Grained Pose Dependent Recognition

A defining characteristic of referee signal recognition is that information lays primarily in upper body pose configuration, particularly arm orientation and hand position. Unlike coarse action recognition tasks (e.g., running, jumping), referee signals require fine grained discrimination between visually similar limb configurations. This aligns the task more closely with fine grained gesture recognition than with generic activity classification.

Modern computer vision systems often address such problems through pose estimation frameworks that explicitly model human skeletal structure (e.g., multi-person 2D pose estimation

methods such as (Cao, Hidalgo Martinez, Simon, Wei, & Sheikh, 2017)). Pose based representations can increase robustness to background variation by focusing on body geometry rather than global scene appearance. However, pose estimation itself introduces additional computational complexity and may suffer under severe occlusion or low resolution broadcast conditions.

For this reason, gesture recognition in sports environments often involves a trade-off between explicit pose modeling and appearance based feature learning. The choice of representation must balance robustness, computational feasibility, and the constraints forced by available training data.

### 2.3 Referee Signal Recognition in Basketball: Existing Approaches and Limitations

Compared to general human action recognition, research specifically targeting basketball referee signal recognition remains relatively limited. Existing studies have primarily approached the problem using classical computer vision pipelines based on handcrafted feature extraction and conventional machine learning classifiers.

One of the early vision based approaches to basketball referee signal recognition employed Histograms of Oriented Gradients (HOG) for feature extraction combined with Support Vector Machines (SVM) for classification (Žemgulys J. , Raudonis, Maskeliūnas, & Damaševičius, 2018). In this framework, referee images extracted from video were preprocessed, edge detected, and represented using gradient descriptors before classification. Despite the relatively small dataset and controlled experimental setup, high classification accuracy was reported for a limited set of gesture classes.

An extended version of this work further included Local Binary Patterns (LBP) features and evaluated both SVM and Random Forest classifiers on referee signals extracted from real time basketball videos (Žemgulys J. , Raudonis, Maskeliūnas, & Damaševičius, 2020). The study demonstrated that handcrafted texture and gradient descriptors can achieve strong performance under constrained conditions. However, the dataset remained relatively small, and gestures were considered in isolated settings with manually defined regions of interest.

Several important characteristics emerge from these prior studies:

1. Handcrafted Feature Dependence

Most existing methods rely on manually designed descriptors such as HOG or LBP (Žemgulys J. , Raudonis, Maskeliūnas, & Damaševičius, 2020). While these features capture edge orientation and local texture patterns effectively, they are limited in their ability to generalize. Modern deep learning approaches learn hierarchical representations directly from data and may better capture pose variations (Poppe, 2010) (Wang, HoKun, & Zhou, 2024).

2. Limited Dataset Scale and Diversity

The datasets used in previous referee signal studies are typically small and restricted to a limited number of gesture classes. Viewpoint diversity, occlusion, and broadcast variability are often underrepresented. As a result, reported accuracies may not reflect performance under real world broadcast conditions (Torralba & Efros, 2011).

### 3. Isolated Gesture Assumptions

Many approaches treat gesture recognition as a classification task performed on pre cropped images containing the referee. This assumption simplifies the problem but does not address the challenge of reliably localizing referees in unconstrained scenes. In practical broadcast scenarios, the system must first detect the referee before classifying the signal.

### 4. Limited Consideration of Temporal Integration

Although gestures are temporal, prior work often performs classification at the frame level or on short, manually segmented clips (Žemgulys J. , Raudonis, Maskeliūnas, & Damaševičius, 2018). While this simplifies modeling, it does not fully exploit temporal consistency or address ambiguity between visually similar static postures.

More recent work has also explored real-time referee gesture recognition with spatio-temporal deep learning architectures (Wang, HoKun, & Zhou, 2024). Overall, the literature demonstrates that referee signal recognition is feasible using both classical and modern methods. However, several limitations remain evident and motivate the exploration of more structured and scalable approaches. This gap provides the foundation for the multi stage pipeline adopted in this thesis, which addresses referee localization, signal presence detection, and signal type classification as distinct but interconnected subproblems.

## 2.4 Pose Estimation and Fine-Grained Visual Representation

Recognition of referee signals depends on the accurate interpretation of body configuration, particularly the spatial arrangement of the arms, forearms, hands and fingers. Unlike other action categories such as running or jumping, referee gestures differ primarily in minimal variations of limb orientation and relative joint positioning. Consequently, the task aligns closely with fine grained visual recognition.

One notable line of research in human computer vision addresses this challenge through pose estimation, where the goal is to detect and localize anatomical keypoints (e.g., shoulders, elbows, wrists) from images or video. Multi-person 2D pose estimation methods such as OpenPose (Cao, Hidalgo Martinez, Simon, Wei, & Sheikh, 2017) introduced Part Affinity Fields to associate detected keypoints into skeletal representations in real time. Such frameworks enable downstream reasoning about human posture independent of background texture or global scene appearance.

Pose based representations offer several conceptual advantages for gesture recognition. By abstracting away from raw pixel intensities and focusing on joint geometry, they can reduce sensitivity to illumination changes, clothing variations, and background clutter (Moeslund & Granum, 2001). In structured gesture domains, such as sign language interpretation or officiating signals, modeling of skeletal configuration may enhance distinction between visually similar gestures that differ primarily in joint angles or arm elevation. However, the integration of pose estimation into broadcast analysis introduces trade-offs. First, pose detection itself can be sensitive to occlusion, motion blur, and low resolution, conditions frequently encountered in sports footage. Second, additional processing stages increase computational requirements, which may be critical in real time applications. Third, pose based pipelines require reliable keypoint detection across diverse camera angles, which is not guaranteed.

As a result, gesture recognition systems in sports often operate along a spectrum between pose modeling and appearance based learning. Appearance based approaches, including convolutional neural networks trained directly on cropped referee images, learn pose sensitive features without requiring an explicit skeletal representation (Wang, HoKun, & Zhou, 2024). While potentially less interpretable, such methods may prove more robust when keypoint detection is unstable or when resolution constraints limit precise joint localization.

In the context of basketball referee signal recognition, the central challenge lies not only in identifying a human figure, but in distinguishing fine grained hand-fingers configurations under broadcast variability. Whether through explicit skeletal modeling or learned appearance representations, successful systems must capture distinctive pose information while remaining resilient to environmental noise.

## 2.5 Dataset Bias, Shortcut Learning, and Domain Shift

While advances in deep learning have significantly improved visual recognition performance, recent research reveals a critical limitation of data driven models. Their tendency to exploit correlations present in training datasets rather than learning the intended semantic concepts. This phenomenon, commonly referred to as *shortcut learning*, describes the behavior of neural networks that rely on easily exploitable cues instead of robust, causally relevant features (Geirhos, et al., 2020).

In image classification tasks, shortcut learning often manifests as reliance on background context, texture statistics, or re-occurring objects rather than the target object itself. For example, a model trained to recognize animals may implicitly associate specific backgrounds (e.g., water for boats, grass for cows) with particular labels. Although such models can achieve high in-distribution accuracy, their performance may degrade substantially under distribution shifts where contextual cues differ.

This issue is closely related to dataset bias and domain shift, where the properties of the training data differ from those encountered during deployment (Torralba & Efros, 2011). In broadcast sports footage, contextual factors such as camera angle, court markings, team uniforms, scoreboard overlays, and crowd presence may correlate with specific referee signals in the training set. A model trained directly on broadcast frames may therefore learn scene level associations rather than focusing on the referee’s gesture.

The challenge is particularly visible in fine grained gesture recognition. Because referee signals are visually subtle and occur within complex scenes, contextual cues can be easier for a model to exploit than the intended distinctive features of limb configuration. High classification accuracy on a broadcast dataset does not imply that the model has learned to interpret referee gestures semantically, it may instead reflect sensitivity to incidental visual regularities.

These considerations underscore the importance of architectural strategies aimed at reducing reliance on contextual shortcuts. Approaches that explicitly isolate the referee from the surrounding scene, focus on body centered representations, or decouple detection and classification can mitigate the risk of context learning. By limiting the visual input to regions containing the relevant human subject, such systems encourage models to learn features more directly aligned with gesture semantics. In summary, dataset bias and shortcut learning represent fundamental risks in broadcast based visual recognition systems. Awareness of these is essential when designing models for referee signal interpretation.

## 2.6 Summary and Research Gap

This chapter positioned basketball referee signal recognition within the broader domains of computer vision, gesture recognition, and sports analytics. Gesture recognition was discussed in terms of sensor-based versus vision-based and static versus dynamic, highlighting the particular limitations of broadcast based analysis in professional sports environments.

Existing work on basketball referee signals demonstrates that the problem is technically feasible using both handcrafted feature pipelines and classical classifiers. However, prior studies have generally relied on relatively small datasets, controlled cropping assumptions, and limited variability in recording conditions. Furthermore, many approaches treat gesture recognition as a standalone classification problem, without addressing referee localization in unconstrained scenes.

The discussion of pose based representations emphasized that referee signal interpretation is fine grained and dependent on hands configuration. At the same time, the analysis of dataset bias and shortcut learning shows the risk that high performing models may exploit contextual correlations rather than learning gesture semantics, particularly when trained directly on broadcast frames.

Taken together, the literature reveals a gap between controlled gesture recognition and robust systems capable of generalizing across various game conditions. There remains a need for structured approaches that explicitly isolate the referee within the scene, focus on body centered visual information and mitigate shortcut learning by reducing reliance on global contextual cues.

### 3 Problem Definition & Theoretical Analysis

This chapter provides a conceptual analysis of the problem addressed in this thesis, focusing on the challenges in recognizing basketball referee signals from broadcast video data. This chapter examines referee signaling as a visual recognition task, the bias introduced by broadcast camera footage, and the limitations of image only classification approaches. Through this analysis, the chapter motivates the need for a referee-centric, multistage pipeline and establishes the theoretical foundations that guide the architecture and methodology presented in the following chapters.

#### 3.1 Problem Statement

Automatic recognition of basketball referee signals from video is a challenging computer vision problem due to the visual, contextual, and temporal characteristics of broadcast sports footage. Referee signals are typically expressed through hand and arm gestures that occupy a relatively small region of the frame, are often partially visually blocked, and exhibit substantial variability in execution across game situations. These challenges are magnified by the properties of broadcast video, where camera viewpoints are optimized for following the gameplay, rather than officiating actions, resulting in frequent zoom changes, motion blur, and limited focus on referees.

From a computational perspective, the task extends beyond simple gesture classification. While some referee signals can be identified from a single static posture, others are fundamentally temporal and require motion context to be fully interpreted. At the same time, limitations related to data availability, annotation effort, and broadcast camera behavior limit the feasibility of applying fully temporal video-based models. As a result, a tension exists between the temporal nature of certain referee signals and the primarily frame-based information accessible in real world broadcast data and datasets.

The core problem addressed in this thesis is therefore the design of a robust and interpretable approach for recognizing referee signals from broadcast basketball video under realistic limitations. Specifically, the objective is to determine whether a referee-centric, multi stage vision pipeline can reduce contextual bias, mitigate dependency on background cues, and enable reliable signal recognition despite limited resolution, partial temporal information, and the variability in broadcast footage.

#### 3.2 Referee Signals as a Visual Recognition Task

Basketball referee signals are a specialized form of visual communication, expressed primarily through hand and arm gestures that reveal officiating decisions to players, officials, and spectators. From a computer vision perspective, recognizing such signals differs fundamentally from conventional object recognition tasks, as the information is embedded in human pose, limb configuration, and, in

some cases, motion patterns rather than in static object appearance (Aggarwal & Ryoo, 2011). The visual footprint of these gestures is typically small relative to the overall gesture, and their interpretation depends on spatial relationships between body parts.

An additional challenge comes from the variability in human performed gestures. The same referee signal may be executed with differences in arm height, orientation, camera angle, speed, or duration, depending on the individual referee, game context, and momentary conditions. At the same time, different signals may share visually very similar postures, leading to inter-class uncertainty when observed in isolation. These factors complicate the definition of clear visual boundaries between signal classes, particularly when relying on single frame representations.

Referee signals are not uniform in their temporal characteristics. While certain calls are communicated through near static poses that can be captured reliably in individual frames, others involve dynamic motion sequences whose meaning emerges over time. In such cases, a single frame may be insufficient to clarify the signal without access to prior or subsequent motion. This partial temporal nature places referee signal recognition at the intersection of image-based gesture classification and action recognition, introducing ambiguity when temporal context is limited or absent (Tran, Bourdev, Fergus, Torresani, & Paluri).

Finally, the broadcast setting forces further limitations on the visual task. Camera framing prioritizes the ball and players rather than referees, often resulting in non-optimal viewpoints, obstructions, or motion blur precisely when a signal is performed. Therefore, the visual information available for signal recognition is both incomplete and noisy, reinforcing the need to carefully consider how referee signals are framed as a recognition problem under real-world broadcast conditions.

### 3.3 Broadcast Video Bias and Contextual Leakage

Broadcast sports video is produced with the primary objective of enhancing the viewing experience for spectators and not to support analytical or machine learning tasks. As a result, with broadcast footage manifest systematic biases in camera placement, framing, and visual emphasis that directly affect the information available to a vision model. These biases include consistent court player positions, recurring camera angles, and tightly coupled relationships between game situations and visual context. While such regularities are beneficial for human viewers, they introduce misleading factors for data driven models.

In the context of referee signal recognition, these broadcast specific patterns can lead to contextual leakage, where a model learns to associate background cues with signal labels rather than relying on the referee's gesture itself (Geirhos, et al., 2020). For example, certain referee calls tend to occur in predictable game situations, which may correspond with characteristic player or referee

formations. When trained on static broadcast images, a classifier may therefore achieve high accuracy by exploiting these correlations, even if the referee is only weakly represented in the frame.

This phenomenon poses a significant threat to generalization. Models that rely on contextual shortcuts are likely to perform well on test data drawn from the same distribution as the training set, but fail when applied to footage from different games, leagues, or broadcast styles (Geirhos, et al., 2020). Such behavior weakens the validity of performance metrics obtained from conventional train-test splits and hides whether the model has genuinely learned to recognize referee signals or merely memorized dataset specific context.

Recognizing and addressing broadcast video bias is therefore essential for framing the problem of referee signal recognition. Without explicit mechanisms to control or reduce contextual leakage, image based models risk learning correlations that do not transfer beyond the datasets that were used. This insight motivates the need for approaches that explicitly focus on the referee as the primary source of visual information, rather than treating the full broadcast frame as a reliable representation of a gesture.

### 3.4 Limitations of Image Only Classification Approaches

Image only classification approaches treat referee signal recognition as a static visual problem, assuming that all information required for correct classification is contained within a single frame. While this assumption simplifies model design and data annotation, it introduces fundamental limitations when applied to broadcast basketball footage. In many cases, individual frames do not capture the full content of a referee signal, particularly when the gesture unfolds over time or when intermediate poses are visually ambiguous.

In addition to temporal ambiguity, image only models are open to background dominance and contextual overfitting. When the referee occupies a small portion of the frame, classification models will prioritize background features that are more visually correlated with the target labels. As discussed previously, this can lead to shortcut learning, where contextual elements replace the intended signal, resulting in misleadingly high performance and poor generalization beyond the training environment (Geirhos, et al., 2020).

These limitations highlight a mismatch between the nature of referee signaling and the assumptions used on static image classification. While image only approaches may appear effective under controlled conditions or within defined datasets, they struggle to provide robust and interpretable solutions when faced with the variability and contextual bias in real-world broadcast video.

### 3.5 Motivation for a Referee-Centric Pipeline

The limitations of image only, full frame classification motivate a shift to referee-centric analysis. In broadcast basketball footage, the referee is the primary source of information relevant to officiating decisions, yet often represents only a small component of the overall image. Treating the entire frame as input therefore introduces substantial irrelevant information, increasing the likelihood that models rely on contextual cues rather than the intended gestural content.

A referee-centric approach addresses this issue by forcing the visual input to the region of interest which is the referee. Conceptually, this mirrors the way humans interpret referee signals. Attention is directed toward the official's posture and gestures while the surrounding game context is largely ignored. By isolating the referee, the recognition task becomes more focused, reducing background noise.

Apart from reducing bias, a referee-centric approach enables a more structured decomposition of the problem. Instead of attempting to solve detection, gesture presence, and signal classification simultaneously, these components can be addressed sequentially. This separation of steps allows each subtask to be optimized according to its own objectives and limitations, while filtering irrelevant information. As a result, the overall system becomes more interpretable, as errors can be traced to specific stages in the pipeline.

### 3.6 Decomposition into Consecutive Subtasks

Building on the referee-centric perspective, the problem of referee signal recognition can be decomposed into a sequence of simpler subtasks. Rather than treating signal recognition as a single end-to-end classification problem, this decomposition helps to distinct the challenges associated with detection, signal presence, and signal classification.

The first subtask involves identifying and isolating the referee within the broadcast frame. This step focuses exclusively on referee detection, regardless of whether a signal is being performed. By separating the referee from the surrounding scene, subsequent stages operate on inputs that are more closely aligned with the task, therefore reducing background dominance and contextual interference.

The second subtask concerns determining whether the detected referee is performing a signal. Most appearances of a referee do not correspond to officiating gestures, and attempting to classify signal types without first establishing a signal presence introduces unnecessary noise, which is filtered out with this binary classification step.

The final subtask consists of classifying the specific type of referee signal when a signal is present. At this stage, the visual input is both focused and filtered, allowing the classification problem to be addressed under more favorable conditions. This sequential decomposition is setting the stage for the methodology presented in the following chapters.

### 3.7 Questions and Hypotheses

Based on the analysis presented in this chapter, this thesis presents a set of research questions that aim to evaluate the effectiveness and limitations of referee-centric approaches for signal recognition in broadcast basketball video. These questions are reflecting both the practical challenges of the task and the considerations related to bias, generalization, and model interpretability.

The first research question examines whether isolating the referee from the broader broadcast context can reduce reliance on background cues and decrease contextual bias in signal recognition. Specifically, it investigates whether a referee-centric approach leads to more robust performance across different games and recording conditions.

The second research question addresses the role of task decomposition in managing complexity and ambiguity. It explores whether decomposing the problem into consecutive subtasks as, referee detection, signal presence identification, and signal type classification, results in improved interpretability.

Finally, the third research question considers the trade-offs forced by temporal limitations. Given that certain referee signals are dynamic, this question examines whether a frame based pipeline, can provide a viable compromise, compared to action recognition, when working with broadcast footage.

Together, these research questions establish the framework for the investigations presented in later chapters, linking the challenges identified in this chapter to the evaluations that follow.

### 3.8 Summary

This chapter examined the problem of basketball referee signal recognition from a theoretical perspective, focusing on the characteristics of referee gestures, the limitations forced by broadcast video, and the sources of bias that affect image based learning approaches. Referee signals are a challenging visual recognition task involving small, and temporal gestures that are often weakly represented in broadcast frames. Particular attention was given to the role of contextual leakage, highlighting the risks of relying on full frame image classification.

To address these challenges, the chapter motivated a referee-centric approach of the problem and proposed a sequential decomposition into distinct subtasks. This framework proposes the isolation of visual information and aligns the recognition process with the nature of officiating gestures. The insights developed in this chapter provide the foundation for the system architecture presented in the following chapter, where these concepts are translated into a concrete implementation.

## 4 System Architecture & Tools

This chapter introduces the system architecture which supports the proposed referee-centric signal recognition pipeline. Building on the problem presented and conceptual analysis introduced in Chapter 3, the focus here shifts from *what* is being addressed to *how* the solution is realized. Rather than adopting a single end to end model, the system is designed as a multi-stage pipeline in which each component addresses a defined subtask within the overall problem. This decomposition reflects both the natural structure of the task, progressing from locating referees to determining whether a signal is present and finally identifying its type, and the constraints of broadcast basketball footage. The chapter provides a high level overview of the pipeline, explains the rationale behind key architectural choices, and describes the tools and implementation environment used to support this design.

### 4.1 Overall Pipeline Architecture

The proposed system follows a sequential, multi stage pipeline, Figure 4.1, designed to isolate the information necessary for reliable referee signal recognition. Starting from raw broadcast video, the pipeline transforms high dimensional, context rich frames into focused representations centered on the referee. This design mirrors the logical structure of the task itself, identifying who is relevant in the scene, determining whether an actual signal is occurring, and finally classifying which signal is being performed.

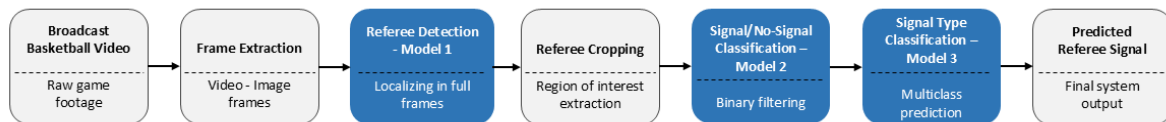


Figure 4.1: Overview of the proposed multi-stage pipeline

At a high level, the pipeline consists of three primary learning components coming after video preprocessing. First, broadcast game footage is converted into individual image frames at a fixed sampling rate, producing a manageable input representation. These frames are then processed by Model 1, a referee detection model responsible for detecting the referees within full game images. Detected referees are cropped and forwarded to Model 2, a binary classifier that filters frames based on the presence or absence of a signaling action. Only frames classified as containing a signal are then passed to Model 3, which performs multiclass classification to identify the specific type of referee signal.

Each stage in the pipeline operates independently and consumes the outputs of the previous stage, forming a feed-forward architecture. This separation allows each model to focus on a narrowly

defined task, reducing reliance on global scene context and reducing the risk of shortcut learning. Moreover, it enables intermediate outputs, such as detected bounding boxes or binary signal decisions, to be inspected and analyzed in isolation, supporting interpretability and systematic error analysis. By breaking down the overall problem into distinct interconnected stages, the pipeline establishes a clear data flow from raw video to final signal prediction.

## 4.2 Design Logic for a Multi-Stage Approach

The decision to adopt a multi-stage system architecture was driven by both the characteristics of referee signal recognition and the limitations of broadcast basketball imagery. Instead of framing the task as a single end to end classification problem, the proposed approach breaks down the task into an order of simpler, more controlled subtasks. This choice reflects the observation that referee signals occupy only a small portion of broadcast frames, while the surrounding context, players, court markings, audience, and score overlays, dominates the visual content.

End to end classification models trained directly on full broadcast images are prone to exploiting contextual correlations that are not directly related to referee gestures. As discussed in Chapter 3, such models may achieve high apparent accuracy by relying on cues, including player formations, camera angles, or other game state indicators, rather than learning representations of the referee’s body posture and hand configuration. A multi-stage pipeline decreases this risk by explicitly isolating the referee early in the process and filtering the available input to the next in order classifiers.

From a systems perspective, this design also enables each model to operate on a task relevant input distribution. The referee detection stage focuses exclusively on spatial localization, independent of signaling gestures. The binary signal detection stage then addresses a critical decision, whether a gesture is present, without the any added complexity. Only after this filtering step is the most challenging task, signal type classification, applied to a reduced and consistent set of images. This hierarchical reduction of complexity supports more stable training and clearer error attribution across stages.

Finally, this architecture offers practical advantages in terms of interpretability and experimentation. Intermediate outputs can be inspected independently, allowing targeted analysis of failures and informed adjustments at specific stages without retraining the entire system. This is particularly valuable in applied computer vision settings, where data availability, annotation effort, and computational resources are not limitless. Collectively, these considerations motivate the adoption of a multi-stage pipeline as a solution to referee signal recognition in broadcast basketball footage.

### 4.3 Video Processing and Frame Extraction

The input to the proposed system consists of broadcast basketball video, which must be transformed into a set of images suitable for image based learning models, as the pipeline operates on individual frames and not in continuous video streams. The workflow begins with the selection and preparation of full game broadcast recordings. Non-informative segments, such as timeouts or halftime breaks, are removed to ensure that extracted frames correspond to active gameplay. This process reduces redundancy and limits the frames that do not contain referees or relevant actions, therefore improving the efficiency of processing.

Following the video trimming, frames are extracted at a fixed sampling rate. Rather than using every available frame, which would introduce high redundancy and unnecessary computational load, a lower frame rate is applied. This strategy maintains a sufficient visual variation to capture gestures while sustaining a manageable dataset size. Each extracted frame is then resized to a uniform resolution to ensure compatibility across models and to standardize input dimensions throughout the pipeline. By the end of the preprocessing stage, the raw video input has been transformed into a sequence of uniform formatted image frames. These frames serve as the input for the referee detection model, enabling consistent data flow and reproducible experimentation across all stages of the system.

### 4.4 Model Integration and Data Flow Between Stage

The proposed pipeline is designed as a sequential system in which each model consumes the outputs of the previous stage, establishing a clear and deterministic flow from raw frames to final signal predictions. This integration strategy ensures that each component operates within a defined role and that information passed between stages is both minimal and relevant.

Following frame extraction, full game images are first processed by the referee detection model. For each input frame, this model produces spatial bounding boxes corresponding to detected referees. These bounding boxes are then used to crop referee-centered images, while discarding the majority of the surrounding areas. The cropped images constitute a new intermediate dataset that serves as the input to the following classification stages. No additional information from the original frame is retained beyond the cropped referee areas.

The second stage of the pipeline applies a binary classification model to the cropped referee images. Its role is to act as a mechanism to filter out frames in which the referee is not performing any signal. Frames classified as containing no signal are dropped at this point and do not propagate further through the flow. This exclusion step significantly reduces the volume of data reaching the final classification stage and limits the influence of ambiguous or non-relevant samples.

Only referee crops that pass the binary filtering stage are forwarded to the multiclass classification model. This final model assigns each input image to one of the predefined referee signal categories. By restricting its input to images which only contain a signaling action, the multiclass classifier operates on a narrower input distribution, allowing it to only focus on visual distinctions between signals.

Essentially, the pipeline does not use collective training, shared weights, or feedback connections between models. Each stage is trained and executed independently, and errors are not propagated backward across components. While this design prevents end to end optimization, it simplifies experimentation, and enables targeted analysis of performance at individual stages. As a result, the integration strategy supports both practical deployment considerations and methodological transparency.

#### 4.5 Tools, Frameworks, and Implementation Environment

The implementation of the proposed system relies on a combination of computer vision frameworks, data annotation tools, and video processing utilities selected to support development, reproducibility, and efficient experimentation. Instead of introducing custom architectures or highly specialized software, the system is built using widely adopted tools that provide robust support for both training and inference across the different stages of the pipeline.

Deep learning models within the pipeline are implemented using a modern object detection and image classification framework that supports both detection and classification tasks within its ecosystem (Jocher & Ultralytics, 2023). This choice supports consistency across stages and simplifies model integration, as detection and classification components share similar and compatible input formats. The framework also provides built-in utilities for evaluation and hardware acceleration, enabling systematic experimentation under specific computational resources.

Dataset annotation and preprocessing are supported through an external annotation platform that enables efficient bounding box labeling, dataset versioning, and augmentation workflows. This tool is used to manage both detection annotations and classification datasets, allowing consistent preprocessing and controlled augmentation strategies across models. Export functionality ensures that datasets can be seamlessly transferred into the training environment without manual restructuring.

Video preprocessing and frame extraction are handled using dedicated multimedia utilities designed for reliable manipulation of large video files. These tools support trimming, format conversion, and frame sampling at specified rates, forming an essential part of the data preparation pipeline. Their use ensures that the transformation from raw video to image frames is deterministic and reproducible.

All experiments are conducted in a local execution environment configured to leverage GPU acceleration where available. The system is implemented using notebook based workflows, allowing iterative development, inspection of intermediate outputs, and fast debugging. Importantly, training configurations, hyperparameters, and evaluation are not detailed in this section and are instead introduced in Chapter 5, preserving a clear separation between design and experimental methodology.

## 4.6 System Limitation and Practical Considerations

The design of the proposed pipeline is shaped not only by the structure and nature of the referee signal recognition task, but also by a set of practical constraints related to data, computation, and annotation effort. These constraints are treated as design features and play a central role in motivating the architectural decisions described in this chapter.

One key constraint emerges from the nature of broadcast basketball footage. Referees typically occupy a small portion of the frame, and their visibility is highly dependent on camera angle, zoom level, and obstruction by players. As a result, only a fraction of extracted frames contain clear and usable signal information. This motivates early stage filtering and reinforces the need for a detection first approach that isolates referees before attempting any interpretation of their actions.

Computational limitations further influence system design. Training and inference are performed under finite GPU memory and system RAM constraints, which restrict the feasibility of large scale end to end models or dense video processing. To address this, the pipeline adopts frame-based inference and streaming prediction strategies, ensuring that images are processed sequentially rather than loaded into memory in large batches. This approach enables scalable processing of full game footage while avoiding memory bottlenecks.

Annotation effort represents an additional practical consideration. Labeling of referee signals is time consuming, making large, fully annotated datasets difficult to obtain. By structuring the pipeline to progressively reduce the number of candidate frames, the system limits the volume of data requiring manual review and labeling at later stages. This selective refinement supports more efficient use of annotation resources and allows higher quality labels to be assigned where they are most impactful.

Finally, the nature of the system reflects a trade off between optimality and controllability. While collective optimization of all components might result in marginal performance gains under ideal conditions, such approaches complicate training, hide error sources, and increase sensitivity to dataset bias. In contrast, the chosen architecture prioritizes transparency, robustness, and adaptability.

## 4.7 Summary of Architectural Decisions

This chapter has presented the architectural foundation of the proposed referee-centric signal recognition system, emphasizing structure, data flow, and the design logic rather than experimental detail. The system is organized as a multi-stage pipeline that transforms raw broadcast video into progressively more focused representations. Key decisions include the decomposition of the task into three sequential learning stages, the use of frame based processing for broadcast video, and the separation of model responsibilities. Referee detection, signal presence detection, and signal type classification are treated as independent but also interconnected components. With the overall system architecture established, the next chapters detail the experimental methodology used to train and evaluate each component of the pipeline.

## 5 Experimental Methodology

This chapter describes the experimental methodology that was followed to investigate the problem of automatic basketball referee signal recognition under real world broadcast conditions. Building upon the analysis and architecture introduced in the previous chapters, the focus here is on how the proposed approaches were implemented, trained, and evaluated. Empirical performance, will be presented separately in Chapter 6.

Two distinct methodological directions are considered. First, an alternative broadcast based classification approach is examined, in which static broadcast images are used directly for signal recognition. This approach serves both as a baseline and as a point of comparison, highlighting the limitations of image only classification when applied to broadcast footage. Second, the chapter details the proposed referee-centric, multi stage pipeline, which breaks down the task into three sequential models: referee detection, signal presence classification, and signal type classification.

The experimental design reflects the practical limitations described earlier, including limited labeled data, broadcast biases, and the partial temporal nature of referee gestures. As a result, methodological choices emphasize data realism, modularity, and interpretability, rather than end to end optimization. Each model is therefore trained and evaluated independently, allowing failure modes and generalization behavior to be analyzed at each stage of the pipeline.

### 5.1 Alternative Approach: Broadcast-Based Classification (Roboflow)

As an initial methodological exploration, an alternative approach based on static broadcast image classification was implemented and evaluated. In this setup, referee signals are inferred directly from individual broadcast frames, without explicitly detecting referees or isolating gesture specific regions. This approach reflects a common formulation in image classification tasks, where each image is treated as an independent sample and mapped directly to a predefined class label.

The dataset used for this approach consisted of broadcast images extracted from basketball game footage (Roboflow, 2024). Each image captured the full court view as produced by the broadcast feed, including players, referees, court markings, overlays, and audience. No spatial constraints were forced to ensure that the referee or the signaling gesture occupied a dominant and consistent region within the frame. As a result, the classification task relied on the model’s ability to identify relevant visual cues within complex scenes.

Dataset preparation focused primarily on label consistency and data cleanliness, rather than on data augmentation. Images associated with multiple signal labels were removed to avoid ambiguity, while frames without an assigned signal were explicitly mapped to a *no-signal* class. Duplicate images were also eliminated to prevent unintentional overrepresentation of near-identical samples within the

training data. Although multiple variants of the same frame (e.g., color and grayscale versions) were present in the original dataset, deduplication was applied to ensure a balanced and non-redundant training set.

Model training was performed locally using a lightweight image classification architecture, initialized from pretrained weights. All images were resized to a fixed resolution to ensure consistency with the model’s input requirements. No additional data augmentation was applied during training, allowing the experimental setup to reflect the raw statistical properties of the broadcast dataset as closely as possible.

This broadcast-based classification approach is included in this work as a baseline point of comparison, instead of the primary solution. Its simplicity differs from the referee-centric, multi stage pipeline described in the next sections, and it provides a useful reference for examining the limitations of static image classification when applied to broadcast sports footage.

## 5.2 Data Source Selection & Constraints

The selection of data sources played a significant role in forming the methodology of this work. As discussed in the previous section, an initial exploration based on static broadcast image classification was feasible from an implementation perspective, but revealed fundamental limitations related to the nature of broadcast footage. These limitations motivated the adoption of a video based, referee-centric data acquisition strategy, despite its higher practical and technical cost.

Broadcast video is produced with the primary goal of maximizing viewer engagement rather than supporting visual analysis. Camera framing, zoom level, and viewpoint are optimized to follow the ball and key gameplay moments, resulting in frequent variations in referee visibility, scale, and orientation. In many frames, referees occupy a small portion of the image or are missing, while other elements such as players, spectators, overlays, and advertisements dominate the scene. As a result, datasets derived from broadcast frames pair referee signals with contextual and game specific cues that are not directly related to officiating actions (Torralba & Efros, 2011).

To address these constraints, full basketball games were recorded directly and treated as the primary data source for the proposed pipeline. Recording entire games enabled complete control over frame extraction and ensured that all referee actions, including transitional gestures, were retained. This approach allowed the dataset to reflect the natural temporal continuity of officiating behavior, even though individual frames were ultimately processed by the models.

Furthermore, the use of recorded game footage also enabled the evaluation of the proposed system under a continuous video input. Unlike static image datasets, recorded video provides stream of frames, allowing the full multi stage pipeline to operate with real time or live game scenarios. This design choice reflects a system oriented perspective, in which referee detection, signal filtering, and

signal classification are applied sequentially to incoming frames rather than to isolated images. While referee focused crops could alternatively be obtained through offline annotation of existing image datasets, such an approach would not reflect the operational conditions of a live basketball game. By adopting a video based input, the setup validates that the proposed pipeline is compatible with streaming inference and continuity, even though individual models are trained on static frames.

From the recorded video footage, frames were extracted at a fixed sampling rate ensuring sufficient variation in referee pose and gesture progression, while avoiding excessive redundancy between similar frames. All extracted frames were resized to a consistent resolution to maintain compatibility across models and to simplify processing.

The decision to rely on recorded game footage introduced several practical constraints. Manual annotation effort increased substantially compared to using pre-existing labeled image datasets, particularly for referee detection and signal labeling. Storage requirements and preprocessing time also grew due to the volume of extracted frames. Additionally, training and inference needed to be designed with hardware limitations in mind, favoring lightweight models and memory efficient processing strategies.

Despite these challenges, the recorded video based approach was selected not only because it enabled a clear separation between referee related information versus a broadcast context, but also as it reflects the operational conditions of a live basketball game. As a consequence, the methodological choices described in the following sections prioritize data realism and controlled visual focus over convenience, reflecting the core objectives of this thesis.

### 5.3 Model 1: Referee Detection

The first stage of the proposed pipeline focuses on detecting referees in full frame basketball game images. This step is essential for isolating referee information and enabling the signal classification to operate on referee only inputs. The detection stage explicitly localizes referees and serves as the entry point of the multi-stage system.

Training data for the referee detection model was constructed from frames extracted from recorded basketball games. Referees were manually annotated using bounding boxes, resulting in a dataset tailored specifically to the visual characteristics of professional basketball officiating. This annotation strategy ensured that the detection task remained focused solely on referee localization, independent of any signaling activity.

Data augmentation was applied with the objective of improving robustness while preserving realism. Augmentations were selected conservatively to reflect variations encountered in real game footage, such as minor changes in orientation, brightness, or image quality. Extreme geometric

transformations or aggressive visual distortions were avoided, as they would not correspond to realistic broadcast or recorded-game conditions and could negatively impact generalization.

The detection model was implemented using a single stage object detection architecture based on the YOLOv8 framework (Jocher & Ultralytics, 2023). This family of models was selected due to its favorable trade-off between detection accuracy and inference speed, making it suitable for deployment within a sequential video processing pipeline. Training was performed on resized frames with a fixed input resolution to ensure consistency across all stages of the system.

During inference, the trained detector was applied frame by frame to recorded game footage. Detected referee bounding boxes were used to automatically generate cropped referee images, which were stored and passed to subsequent classification stages. To accommodate the large volume of video frames and to prevent excessive memory usage, inference was designed to operate in a streaming manner, processing frames sequentially rather than loading entire videos into memory. The performance of this model, as well as its role in enabling reliable referee-centric cropping, is evaluated in detail in Chapter 6.

#### 5.4 Model 2: Signal / No-Signal Classification

The second stage of the proposed pipeline addresses the task of distinguishing between signaling and non-signaling referee images. Given cropped referee images produced by the detection stage, this model performs a binary classification that determines whether a referee is actively making an officiating signal or is in a neutral, transitional, or non-signaling pose. This step acts as a critical filter, ensuring that only frames containing meaningful signaling activity are forwarded to the final multiclass classification stage.

The dataset for this model was constructed using referee crops generated by Model 1 from recorded game footage. Cropped images were manually reviewed and assigned to one of two classes: signal or no-signal. This labeling process accounted for the ambiguity presented in officiating gestures, including transitional movements, partially completed signals, and visually unclear poses. As a result, the binary formulation was chosen to separate frames with clear signaling intent from all other referee appearances, rather than attempting fine classification at this stage.

To avoid biasing the classifier toward visual cues, particular attention was paid to class balance and visual diversity. The dataset was selected to contain comparable numbers of signaling and non-signaling examples, drawn from multiple games and a variety of in game contexts. Augmentation was applied conservatively, with the goal of preserving realistic referee appearance while introducing limited variation in pose and lighting. This approach was intended to prevent overfitting while maintaining alignment with real world video input conditions.

The binary classifier was implemented using an image classification model initialized from pretrained weights within the YOLOv8 framework (Jocher & Ultralytics, 2023). Training was performed on resized referee crops with a fixed input resolution to ensure consistency across experiments. The model was trained independently from the detection stage, allowing its behavior and generalization characteristics to be analyzed in isolation. This intermediate classification stage reflects a design choice within the pipeline. By separating signal presence detection from signal type recognition, the system reduces the complexity of the multiclass task and limits error propagation. The effectiveness and limitations of this binary classifier, particularly with respect to generalization across games and recording conditions, are examined in Chapter 6.

### 5.5 Model 3: Signal Type Classification

The final stage of the proposed pipeline focuses on multiclass classification of referee signal types. Given referee crops that have been filtered by the previous binary classifier, this model aims to assign each signaling instance to a specific officiating category. This stage represents the most fine grained and visually demanding task in the pipeline, as many basketball referee signals differ only in slight variations of hand position, arm orientation, or body posture.

The dataset for multiclass signal classification was constructed by manually assigning signal labels to referee crops identified as signaling instances. Signal categories were defined to reflect the most frequently occurring and visually distinguishable gestures observed in the recorded games. Due to the natural imbalance in signal frequency during gameplay, the resulting dataset resulted in uneven class distributions, with certain signal types appearing substantially less often than others (He & Garcia, 2009).

Given the limited size of the available dataset, particular care was taken in framing the classification task. Signal classes were grouped conservatively to balance optimal specificity and statistical feasibility, avoiding excessive distinctions that could not be reliably supported by the data. This grouping strategy reflects a practical trade-off between recognition detail and model robustness, consistent with the limitations identified earlier in the thesis.

The multiclass classifier was implemented using a lightweight image classification architecture within the YOLOv8 framework. As with the previous stages, training was performed on resized referee crops with a fixed input resolution. The model was trained independently from the binary classifier, allowing its behavior to be evaluated separately and preventing error coupling during training. This final classification stage completes the referee-centric pipeline by transforming isolated signaling instances into labels. However, due to dataset size limitations and the visual similarity between certain signal types, this task poses significant challenges for generalization. The quantitative performance of

the multiclass classifier, as well as an analysis of its strengths and limitations, is presented in the following Chapter.

## 6 Experimental Results

This chapter presents a detailed experimental evaluation of the proposed referee-centric pipeline for basketball referee signal recognition. Unlike the previous chapters, which focused on formulating the problem, system design, and methodological choices, the emphasis here is placed on results derived from the experimentation. Each component of the pipeline, from broadcast based image classification to referee detection and the following signal classification, is evaluated using quantitative metrics, dataset specific statistics, and qualitative visual evidence. Model configurations, dataset sizes, preprocessing and augmentation strategies, and observed limitations are reported explicitly in order to provide a transparent and reproducible assessment of performance. In addition, this chapter examines both successful outcomes and failure cases, with particular attention to generalization behavior and error propagation across pipeline stages. The results presented here form the basis for the interpretation and discussion provided in the following chapter.

### 6.1 Evaluation Setup and Metrics

This section describes the evaluations adopted for assessing the performance of the proposed system and its individual components. Given the multi stage nature of the pipeline, different evaluation strategies and metrics are employed depending on the task addressed by each model. All evaluations are conducted at the frame level, in order to isolate model behavior.

For referee detection (Model 1), performance is evaluated using standard object detection metrics, including precision, recall, and mean Average Precision (mAP) at an Intersection over Union (IoU) threshold of 0.5. Precision measures the proportion of detected referee bounding boxes that correspond to true referees, while recall quantifies the proportion of referees in the frame that are successfully detected. The mAP@0.5 metric provides a summary of detection accuracy by combining localization and classification performance across confidence thresholds.

For signal versus no-signal classification (Model 2) and signal type classification (Model 3), classification performance is evaluated using accuracy, precision, recall, and confusion matrices (Sokolova & Lapalme, 2009). Accuracy is reported as an overall indicator of correct predictions, while confusion matrices are used to analyze class-wise behavior, misclassification patterns, and sources of error. In the multiclass setting, particular attention is paid to confusion between visually similar signal categories and to the impact of class imbalance on performance.

Evaluation datasets are constructed using held-out images that were not seen during training, following the train-validation–test splits described in Chapter 5. Where applicable, additional qualitative evaluations are performed on images and video frames drawn from unseen games to assess generalization beyond the training distribution. Confidence thresholds used during inference are kept

consistent with those applied during experimental tuning, and no test time augmentation is employed. Finally, it is emphasized that the results presented in this chapter reflect component level behavior under realistic operating conditions, rather than optimized benchmark performance. This choice aligns with the objective of evaluating the feasibility and limitations of a referee-centric pipeline applied to real basketball game footage.

## 6.2 Broadcast-Based Classification (Alternative Approach)

The purpose of this experiment is to evaluate a direct broadcast based image classification approach as an alternative to the proposed referee-centric pipeline. In this, full broadcast frames are classified directly into referee signal categories without explicitly detecting or isolating the referee. This approach serves as a baseline and comparison point, as similar strategies are commonly adopted in image classification tasks when explicit object localization is not enforced. The goal of this subsection is not to optimize broadcast-based performance, but to assess whether such an approach can reliably learn referee signal gestures or whether it instead exploits contextual cues present in broadcast footage. The findings from this experiment are used to motivate the necessity of the multi stage, referee focused pipeline introduced in the previous chapters.

### 6.2.1 Data Used

The broadcast-based classification experiment employs a publicly available dataset composed of images extracted from basketball game footage (Roboflow, 2024). The dataset consists of full frame broadcast images and was not collected or annotated as part of this work. As such, it reflects a diverse set of games, camera angles and arenas, providing a realistic visual environment.

Each image is labeled directly with a referee signal class, forming a single label multiclass classification task. No bounding boxes or annotations are provided, and referees are not explicitly localized within the frames. As a result, the referees typically occupy a small and variable portion of the image, while players, court markings, spectators, and broadcast graphics dominate the visual content. Representative examples of the broadcast based input images are shown in Figure 6.1.

The dataset is divided into predefined training, validation, and test splits containing 6082, 581, and 291 images respectively. Given that the experimental analysis and dataset preparation steps were applied exclusively to the training split, the following discussion focuses on the composition and processing of the training data.



Figure 6.1: Representative broadcast images used for direct referee signal classification (Roboflow, 2024).

An inspection of the original training set revealed substantial class imbalance across referee signal classes. High frequency classes such as *free throw attempt*, *foul*, and *three-point attempt* appeared orders of magnitude more often than rare signals such as *technical foul* or *three free throws*. The initial class distribution of the training set is summarized in Table 6.1.

Further examination of the training annotations identified several labeling inconsistencies. Specifically, 5486 images were associated with exactly one referee signal label, 581 images contained no observable referee signal, and 15 images were assigned multiple, conflicting signal labels. The ambiguously labeled images were removed from the training set, while the images without observable signals were retained and explicitly assigned to a newly introduced *no signal* class. No labeling issues of this type were observed in the validation or test splits.

Table 6-1: Initial class distribution of the broadcast based training dataset prior to preprocessing (Roboflow, 2024).

| Signal Class                 | Number of Images |
|------------------------------|------------------|
| Free throw attempt           | 1590             |
| Foul                         | 1041             |
| Three-point attempt          | 879              |
| Two-point shot attempt       | 539              |
| Free throw made              | 321              |
| Two free throws              | 266              |
| Substitution                 | 246              |
| Three-point made             | 198              |
| Timeout                      | 98               |
| Jump ball / held ball        | 81               |
| One free throw               | 80               |
| Traveling                    | 66               |
| Unsportsmanlike foul         | 42               |
| End of possession            | 27               |
| Kicked ball / foot violation | 21               |
| Technical foul               | 15               |
| Three free throws            | 6                |

In addition, the dataset was found to contain systematic duplication in the form of grayscale variants (Roboflow, 2024). For each original color image, two visually identical grayscale copies were present, effectively tripling the number of samples without introducing new visual information. To mitigate redundancy while preserving class diversity, a controlled deduplication strategy was applied. For high frequency classes, a single randomly selected variant was retained per image. For mid frequency classes, two variants were retained and for low frequency classes, all available variants were kept. This strategy reduced redundancy while maintaining representation for rare signal categories. The resulting training set comprises 18 classes, including the *no signal* category, with a revised class distribution reported in Table 6.2. Although class imbalance remains, the final dataset provides a more controlled and transparent basis for evaluating broadcast based signal classification without altering the underlying visual characteristics of the data.

Table 6-2: Final class distribution of the broadcast based training dataset after preprocessing.

| Class Index | Signal Class                 | Number of Images |
|-------------|------------------------------|------------------|
| 0           | Free throw attempt           | 528              |
| 1           | No signal                    | 386              |
| 2           | Two-point attempt            | 360              |
| 3           | Foul                         | 348              |
| 4           | Three-point attempt          | 291              |
| 5           | Free throw made              | 214              |
| 6           | Two free throws              | 178              |
| 7           | Substitution                 | 166              |
| 8           | Three-point made             | 130              |
| 9           | Timeout                      | 98               |
| 10          | One free throw               | 80               |
| 11          | Jump ball / held ball        | 69               |
| 12          | Traveling                    | 54               |
| 13          | Unsportsmanlike foul         | 42               |
| 14          | End of possession            | 27               |
| 15          | Kicked ball / foot violation | 21               |
| 16          | Technical foul               | 15               |
| 17          | Three free throws            | 6                |

### 6.2.2 Preprocessing and Augmentation

Prior to training, all broadcast images were resized to a fixed spatial resolution of  $640 \times 640$  pixels, consistent with the default input configuration of the YOLOv8 classification framework (Jocher & Ultralytics, 2023). Apart from this resizing operation, no additional preprocessing or data augmentation techniques were applied. In particular, images were not cropped, masked, normalized, or modified to emphasize the referee region.

This decision was made deliberately in order to preserve the original visual characteristics of broadcast footage, including contextual elements such as players, court markings, spectators, and on-screen graphics, as illustrated in Figure 6.1. By avoiding artificial transformations, the experiment aims to evaluate whether a direct broadcast based classification model can learn referee signal gestures from raw broadcast input alone, without any prior spatial bias or augmentation driven regularization.

### 6.2.3 Model Configuration

For the broadcast based classification experiment, a lightweight image classification model based on the YOLOv8 framework was employed. Specifically, the YOLOv8s-cls variant was selected as a representative small scale classifier, balancing computational efficiency with sufficient model capacity for multiclass image classification tasks.

The model was initialized using pretrained weights (yolov8s-cls.pt), leveraging transfer learning from large-scale image datasets. Training was performed for 30 epochs with a fixed input resolution of  $640 \times 640$  pixels. A batch size of 16 was used during training, chosen to fit within the available GPU memory constraints, while the number of data loading workers was limited to 2 in order to avoid excessive memory usage during training on a local Windows based system. All experiments were executed on a single GPU.

Unless otherwise specified, default optimization settings provided by the YOLOv8 framework were retained, and no architectural modifications or custom loss functions were introduced. This configuration reflects a standard application of a modern image classification model to broadcast footage, allowing the results to be interpreted without effects from extensive hyperparameter tuning or model customization.

### 6.2.4 Quantitative Results

The broadcast-based classification model achieves a top 1 accuracy of 0.897 on the held-out test set, indicating strong overall predictive performance when evaluated purely in terms of classification correctness. At face value, this result suggests that the model is capable of discriminating between a wide range of referee signal categories from full broadcast frames.

A more detailed view of model behavior is provided by the normalized confusion matrix shown in Figure 6.2, where columns correspond to ground truth classes and rows to predicted classes. It should be noted that two signal categories, *end of possession* and *technical foul*, were not present in the test split. To enable computation of the confusion matrix, a single placeholder image was included for each of these classes. Predictions associated with these categories are therefore not considered meaningful and are excluded from interpretation.

For the remaining classes, the confusion matrix shows strong diagonal dominance, with many categories achieving high correct classification rates. Signals such as *free throw attempt*, *jump ball*, *kicked ball*, *timeout*, and *traveling* are classified correctly in all or nearly all test instances. High frequency classes including *foul* and *three-point attempt* also demonstrate strong performance, although minor confusion is observed with related categories.

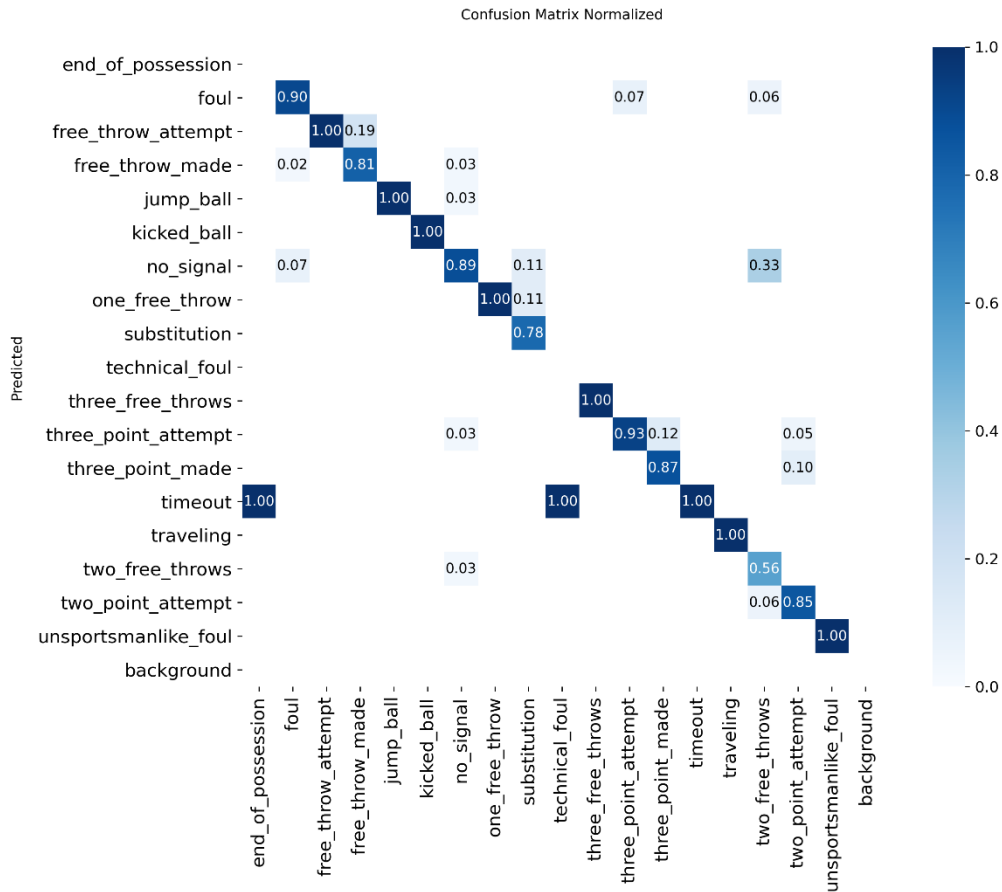


Figure 6.2: Normalized confusion matrix for the broadcast based referee signal classification model evaluated on the test set..

Misclassifications are not evenly distributed, but instead follow consistent and interpretable patterns. Confusion is most pronounced between visually and contextually similar events, such as *free throw made* versus *free throw attempt*, *two-point attempt* versus *three-point attempt* or *three-point made*, and *two free throws* being misclassified as *no signal* or *foul*. These errors suggest that the model relies on contextual cues associated with game flow, player positioning, and scoring situations, rather than visual distinctions between referee gestures.

After excluding the two dummy images that were only introduced to preserve class-folder consistency, the Roboflow baseline model was evaluated again on 290 real test images. The model this time achieved a standard accuracy of 90.3% and a balanced accuracy of 91.2%. Balanced accuracy was included because the dataset was imbalanced and provides a more reliable estimate of average class-wise performance. However, some classes contained very few test samples, so the balanced accuracy should be interpreted together with the per class support values.

Overall, while the quantitative results indicate high test accuracy and stable performance for many signal categories, the observed confusion patterns point to systematic limitations in distinguishing gesture specific information from broadcast images alone. These findings motivate

further qualitative analysis and provide a quantitative basis for evaluating the limitations of direct broadcast based referee signal classification.

### 6.2.5 Qualitative Analysis

Figure 6.3 presents a small set of representative qualitative examples showing both correct predictions and systematic failure cases observed in the broadcast based classification experiment. These examples complement the quantitative results by providing insight into how predictions are formed under realistic broadcast conditions.



Figure 6.3: Representative qualitative examples from the broadcast based referee signal classification experiment. The top row shows correctly classified instances. The bottom row illustrates typical failure cases.

The correctly classified images shown in the top row of Figure 6.3 demonstrate that accurate predictions can be achieved even when referee gestures are only weakly visible. In the *free throw attempt* example, the referee appears distant and partially blocked, however the model predicts the

correct class. Similarly, in the *three-point attempt* example, raised arms are visible but fine grained gesture details, such as three fingers configuration, are not observable. In both cases, the surrounding game context, including player positioning, appears sufficient for the model to arrive at the correct classification.

The bottom row of Figure 6.3 shows typical failure cases that align with the confusion patterns identified in Section 6.2.4. In the *two free throws* example, the model predicts *no signal* despite the presence of a visibly raised referee arm, suggesting that gesture information alone is insufficient when contextual cues are ambiguous or weak. In the *free throw made* example, the model confuses the action with *free throw attempt*, a misclassification that is expected given the near identical broadcast context and minimal visual distinction between these two events at the frame level.

Overall, these examples strengthens the observation that broadcast based classification decisions are strongly influenced by global scene context rather than explicit referee gestures. While such cues are often sufficient for common and structurally distinctive game events, they lead to systematic errors when semantic distinctions or subtle gestures are required. This qualitative behavior motivates the need for a referee centric pipeline, which is explored in the subsequent sections.

#### 6.2.6 Key Takeaway

The visualizations presented in Figure 6.4 provide direct evidence of how the broadcast based model arrives at its predictions. Across multiple correctly classified examples, including *free throw attempt* and *foul*, the regions identified as most influential are predominantly located around the basket area, player formations, and general court layout, while the referee’s gesturing arm receives little or no attention. This indicates that accurate predictions can be produced even when gesture specific visual information is not actively used by the model.

A closer examination of individual frames further highlights the limitations of this behavior. In the *substitution* example misclassified as *foul*, the model’s attention is concentrated on a raised arm belonging to a non-referee individual near the bench area, suggesting sensitivity to generic human pose cues rather than referee signaling patterns. On the contrary, a correctly classified *foul* instance demonstrates attention to the referee’s upper body, illustrating that gesture focus can occur, but only inconsistently and without a stable pattern across predictions.

Taken together, these observations confirm that the broadcast based classifier lacks a reliable mechanism for grounding predictions in referee gestures. Instead, decision making fluctuates between contextual scene cues and opportunistic pose like features, leading to both correct and incorrect classifications. This inconsistency motivates the transition to a referee-centric pipeline, where explicit referee detection is introduced to constrain the visual input and promote more semantically meaningful gesture recognition in the subsequent models.

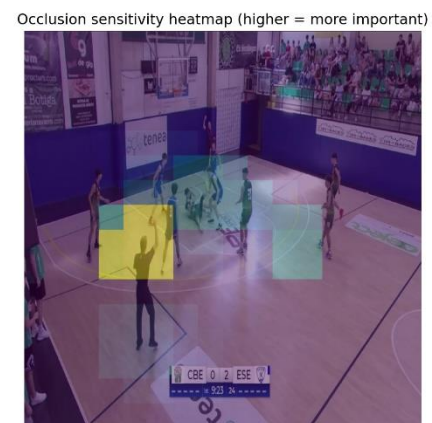
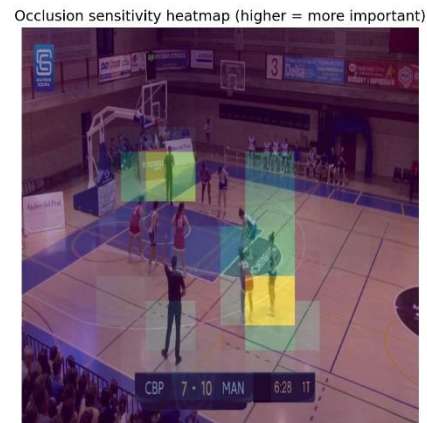
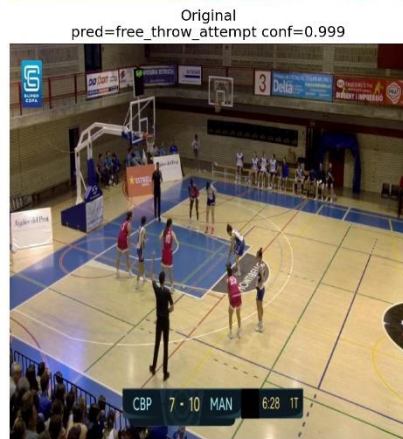
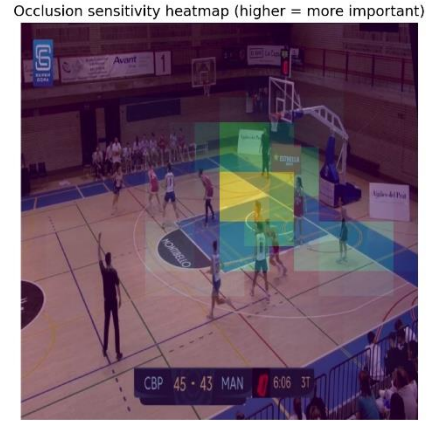


Figure 6.4: Sensitivity heatmaps for representative broadcast based classification predictions. For each example, the original broadcast frame is shown alongside the corresponding sensitivity map.

### 6.3 Model 1 — Referee Detection

The objective of Model 1 is to reliably detect and localize referees in full basketball game frames captured from broadcast video recordings. Unlike the broadcast based classification approach evaluated in Section 6.2, this model explicitly focuses on identifying referees within complex scenes. Accurate referee detection constitutes a foundational step in the proposed pipeline, as all subsequent classification stages depend on the quality and consistency of the extracted referee regions. The scope of this subsection is to evaluate the performance of the referee detection model under realistic operating conditions, including variations in camera angle, referee pose, and background activity. Emphasizing on detection reliability instead of fine grained gesture interpretation, as the primary role of Model 1 is to enable downstream processing by isolating referees from full frame video input.

#### 6.3.1 Data Used

The referee detection model was trained and evaluated using a manually annotated dataset derived from recorded Euroleague basketball games. The dataset consists of image frames extracted from three full games, selected to capture variation in teams, venues, camera angles, and broadcast conditions. Specifically, the games used were *FC Barcelona vs Olympiakos* (Euroleague season 2024–25), *Fenerbahce vs Olympiakos* (Euroleague season 2024–25), and *Real Madrid vs Olympiakos* (Euroleague season 2023–24).

From these games, a total of 390 image frames were selected and manually annotated, with approximately 130 images per game. Each image was labeled with bounding boxes enclosing all visible referees, forming a single class object detection task. Annotation was performed using the Roboflow platform, and only the *referee* class was considered, as the objective of Model 1 is solely to localize referees within full frame images and not to distinguish between different signal types.

All annotated images were resized to a fixed resolution of  $640 \times 640$  pixels to match the input requirements of the detection model. The dataset was split into training and validation subsets following standard practice, with 310 images allocated to training and 80 to validation set used for monitoring performance during training. Importantly, the images used for training and validation of Model 1 are distinct from the additional game footage on which the trained model was later deployed for referee crop extraction in downstream pipeline stages. This dataset provides a focused and controlled basis for evaluating referee detection performance under realistic broadcast like conditions, while remaining compact to allow manual annotation and detailed inspection.

#### 6.3.2 Preprocessing and Augmentation

Prior to training, all annotated images were resized to a fixed spatial resolution of  $640 \times 640$  pixels, consistent with the input requirements of the YOLOv8 detection architecture (Jocher & Ultralytics, 2023). No additional manual preprocessing steps were applied beyond that.

To improve robustness to viewpoint variation, lighting changes, and partial occlusions commonly encountered in broadcast footage, a comprehensive set of data augmentation techniques was applied during training using the Roboflow platform. The augmentation strategy included the following transformations:

- Geometric augmentations:
  - Random rotations in the range of  $-10^\circ$  to  $+10^\circ$
  - Horizontal and vertical shear transformations between  $-10\%$  and  $+10\%$
  - Random cropping of up to  $10\%$  of the image area
- Photometric augmentations:
  - Brightness adjustment in the range of  $-15\%$  to  $+15\%$
  - Saturation adjustment in the range of  $-15\%$  to  $+15\%$
  - Exposure adjustment in the range of  $-5\%$  to  $+5\%$
- Noise and blur:
  - 1-pixel Gaussian blur
  - $0.1\%$  random noise

These augmentations were applied exclusively to the training split in order to preserve a clean validation set. As a result of augmentation, the effective training dataset size increased from the original 310 annotated images to 933 training images, while the validation set remained fixed at 80 images, getting a total of 1,013 images used during model development. The relatively extensive augmentation strategy employed for referee detection reflects the high visual variability of referees in broadcast footage, including changes in pose, scale, camera angle, and illumination. This contrasts with the more conservative augmentation strategies adopted in later classification models, where preserving fine grained gesture details is more critical.

### 6.3.3 Model Configuration

Referee detection was implemented using the YOLOv8 medium (YOLOv8m) object detection architecture, selected to provide a balance between detection accuracy and computational efficiency. The model was initialized with pretrained weights (yolov8m.pt), enabling transfer learning from large scale object detection datasets and facilitating convergence given the limited size of the annotated referee dataset.

Training was performed locally for 100 epochs using an input resolution of  $640 \times 640$  pixels, consistent with the preprocessing and augmentation strategy described in Section 6.3.C. A batch size of 16 was employed, representing a compromise between GPU memory constraints and training stability on the available hardware. All experiments were executed using a single GPU.

The training process utilized the dataset configuration specified in a YAML file, defining the training and validation splits and the single *referee* class. Apart from the number of training epochs and batch size, default optimization parameters provided by the YOLOv8 framework were retained, and no architectural modifications or custom loss functions were introduced. This configuration reflects a standard yet sufficiently expressive detection setup for locating referees in complex broadcast scenes, while avoiding excessive model complexity that could prevent generalization or reproducibility.

#### 6.3.4 Quantitative Results

The referee detection model demonstrates strong and consistent performance on the validation set, indicating reliable localization of referees in full frame basketball footage. Using standard object detection metrics, the model achieves high precision and recall values, reflecting a favorable balance between minimizing false detections and successfully identifying the majority of referees present in the images. The below values indicate that the detector produces mostly accurate referee bounding boxes while maintaining high coverage, which is essential for downstream processing in the proposed pipeline. Accuracy can be further summarized by the mean Average Precision (mAP) confirming robust localization and classification performance under commonly used evaluation criteria. When evaluated under stricter localization requirements, the model reflects the increased difficulty of achieving precise bounding box alignment for small and distant referees in broadcast style images.

- Precision: 0.93
- Recall: 0.95
- Mean Average Precision (mAP): 0.96 at an IoU threshold of 0.5
- mAP: 0.67 averaged over IoU thresholds from 0.5 to 0.95

The precision–recall trade-off curve for the referee detection model is illustrated in Figure 6.5, which shows stable performance across a wide range of confidence thresholds. The shape of the curve indicates that high recall can be achieved without a substantial loss in precision, supporting the suitability of the detector for generating referee centric image crops for downstream classification models. Overall, these quantitative results demonstrate that Model 1 provides a reliable foundation for the proposed multi stage pipeline, showing consistent referee detection under realistic broadcast conditions.

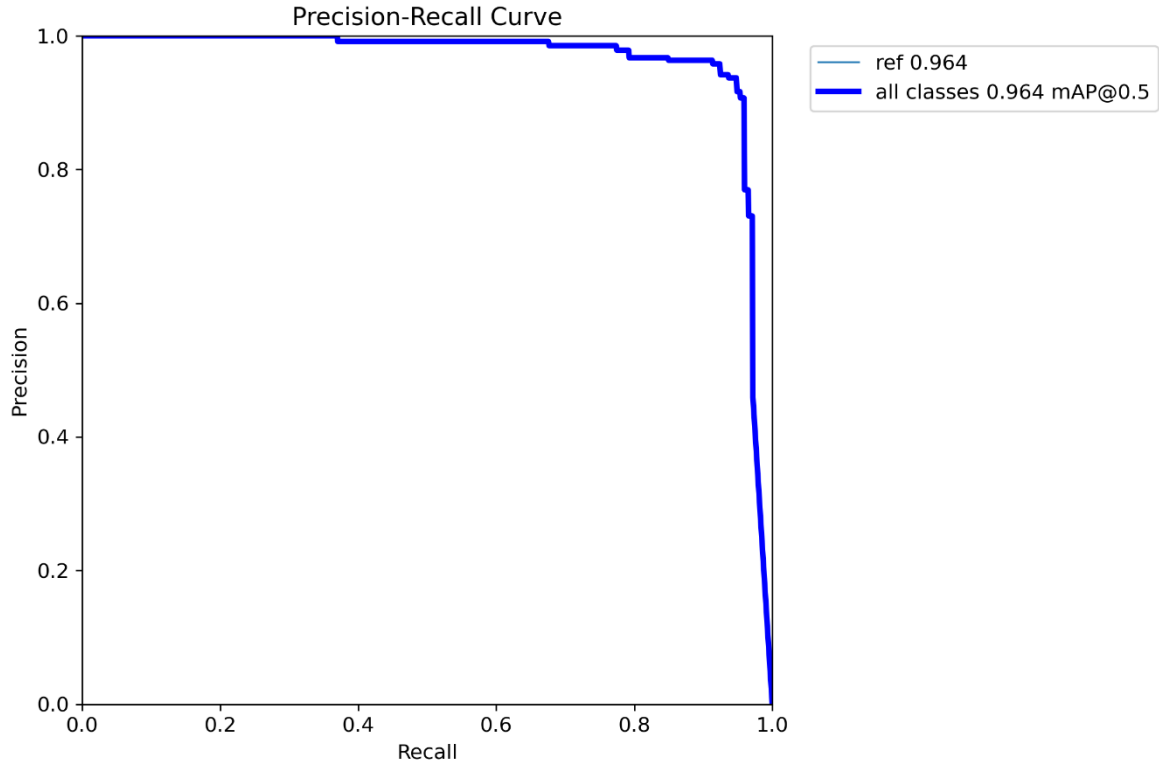


Figure 6.5: Precision–recall curve for the referee detection model evaluated on the validation set.

### 6.3.5 Qualitative Results

Representative inference-time detections on full game footage not used during training are shown in Figure 6.6. The selected examples illustrate the ability of the detector to localize referees under diverse and realistic broadcast conditions, including different arenas, camera perspectives, and game situations.

Across the presented frames, referees are successfully detected in both wide angle and court level views, despite variations in scale and viewpoint. In several cases, multiple referees are present within the same frame and are correctly localized simultaneously. The examples also demonstrate robustness to partial occlusions, such as obstruction by players, broadcast graphics overlays, or proximity to frame boundaries. Notably, in one instance a referee positioned behind the scoreboard overlay is still detected, and in another, a referee partially occluded by a player in a court level camera view is successfully localized.

The qualitative results further show that detection is not limited to static postures. Referees engaged in active signaling are identified reliably, even when arm extensions are partially cut by the bounding box. Minor localization inaccuracy is observed in some cases, such as slight cropping of extended limbs, however, these deviations do not compromise the identification of the referee or the suitability of the detection for subsequent cropping and classification.

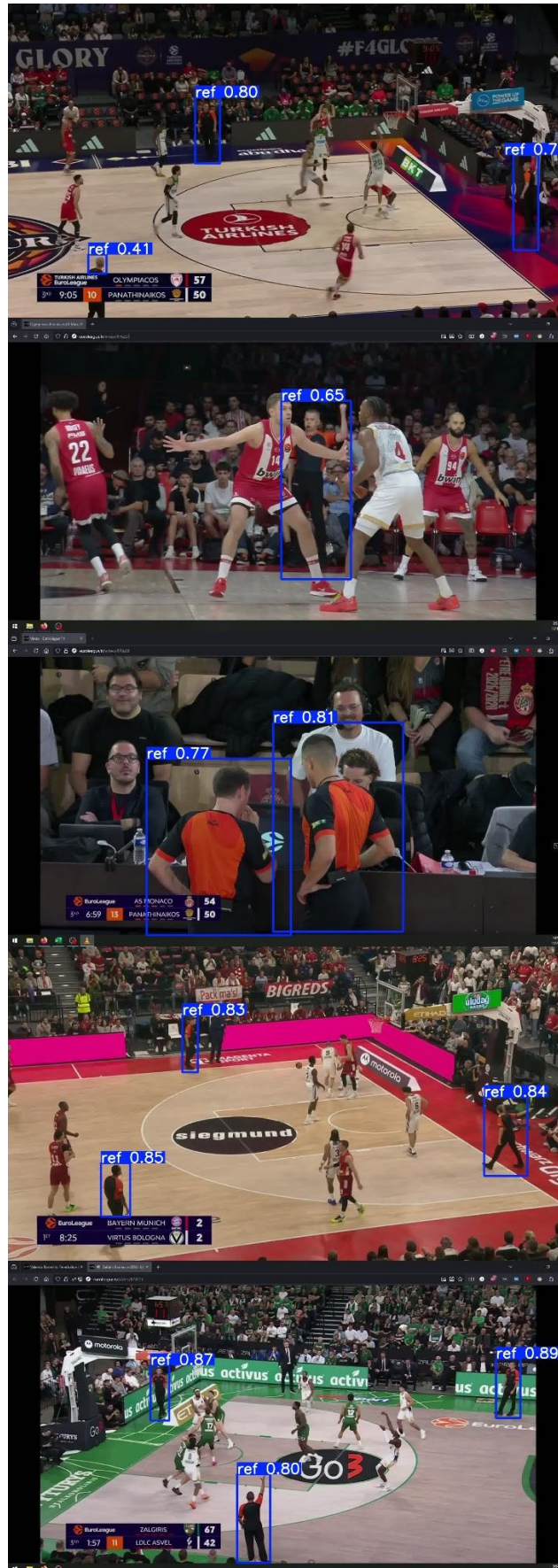


Figure 6.6: Representative referee detection results on previously unseen EuroLeague game footage.

Overall, the qualitative evidence supports the quantitative findings presented in Section 6.3.4, confirming that Model 1 provides consistent and reliable referee localization across heterogeneous broadcast scenarios. This establishes a stable visual foundation for the processing stages that follow.

#### 6.3.6 Key Takeaway

The results of Model 1 demonstrate that referees can be reliably localized in full broadcast frames with high precision and recall, even under challenging real game conditions. Quantitative evaluation shows strong detection performance, while qualitative inspection confirms robustness to variations in camera angle, scale, partial occlusion, and multi referee scenarios. Although minor bounding box inaccuracy is occasionally observed, localization quality is sufficient for extracting referee centered image crops without materially affecting downstream processing. Importantly, by constraining the following analysis to detected referee regions rather than entire broadcast frames, Model 1 directly addresses the contextual bias observed in the broadcast based classification approach. The detector therefore functions not only as a preprocessing step, but as a structural mechanism that shifts the visual focus of the system toward referees. This transition establishes the foundation for the signal specific classification tasks evaluated in the following sections.

### 6.4 Model 2 — Signal / No-Signal Classification

The objective of Model 2 is to determine whether a detected referee is actively performing a signal or not. Building upon the referee localization provided by Model 1, this stage operates on cropped referee images and creates a binary classification task classifying *signal* from *no signal* instances. By restricting the visual input to referee regions only, the model aims to reduce reliance on global broadcast context and focus on referee gestures.

#### 6.4.1 Data Used

##### 6.4.1.1 Initial Dataset Configuration (Baseline)

The initial version of Model 2 was trained using referee crops extracted from four games, yielding a balanced dataset of 930 signal and 930 no-signal instances (1,860 total images). These images were manually verified following inference from Model 1 and uploaded to Roboflow for dataset management and augmentation.

After applying augmentation, the effective dataset size increased to 4455 training images, with 186 validation and 186 test samples. The resulting binary classifier achieved a test accuracy of 0.98, with balanced precision and recall across both classes. While these results indicated strong in-distribution performance, qualitative inspection during inference on additional footage revealed limited generalization capability, particularly in borderline or ambiguous signaling scenarios.

This discrepancy between test set accuracy and unseen footage motivated a reassessment of dataset composition and augmentation strategy.

#### 6.4.1.2 Expanded Dataset for Improved Generalization

To address the observed generalization limitations, the dataset was expanded by applying Model 1 inference across a total of eight games, including both previously used and newly recorded fixtures from the 2025–26 Euroleague season. All detected referee crops were manually reviewed and labeled as *signal* or *no signal*, with particular attention given to ambiguous and transitional frames.

In total, 3649 cropped referee images were uploaded to Roboflow for the revised dataset. The data were split into 2554 training, 548 validation, and 547 test images. Following augmentation of the training set, class balance was maintained, resulting in 2550 signal and 2558 no-signal training instances.

This expanded dataset increases visual diversity across arenas, teams and camera angles, providing a more realistic basis for evaluating binary signal detection performance under varied broadcast conditions.

#### 6.4.2 Preprocessing and Augmentation

##### 6.4.2.1 Initial Configuration

In the initial version of Model 2, referee crops were resized to the model’s default classification input resolution. To increase variability and reduce overfitting, moderate augmentation was applied during training, including:

- Rotation:  $\pm 10^\circ$
- Brightness adjustment  $\pm 15\%$
- Saturation adjustment  $\pm 10\%$
- Additive noise up to 0.1%

These transformations were applied exclusively to the training set. While this configuration produced strong test set accuracy, later qualitative evaluation suggested that the model showed limited robustness to posture variations and signal frames.

##### 6.4.2.2 Revised Configuration

In the revised dataset, preprocessing and augmentation were adjusted to preserve gesture specific information and reduce artificial distortions. All referee crops were resized to  $224 \times 224$  pixels, maintaining aspect ratio and applying zero padding (black borders) where necessary. This ensured consistent input dimensions without geometric stretching that could distort arm orientation or hand position.

Augmentation was deliberately reduced to avoid excessive transformation of gesture features. The revised augmentation strategy included:

- Rotation  $-5^{\circ}$  to  $+5^{\circ}$
- Brightness adjustment  $-5\%$  to  $+5\%$

No saturation or additive noise was applied in the revised configuration. This more conservative augmentation approach was adopted to preserve visual cues associated with referee signaling, particularly in borderline or transitional frames. The goal was to improve generalization by exposing the model to realistic variation without introducing synthetic distortions that could obscure gesture interpretation.

### 6.4.3 Model Configuration

#### 6.4.3.1 Initial Configuration

The initial binary signal classifier was implemented using the YOLOv8 medium classification model (YOLOv8m-cls) initialized with pretrained weights (Jocher & Ultralytics, 2023). Training was conducted for 100 epochs using an input resolution of  $640 \times 640$  pixels, consistent with the preprocessing configuration applied in the first dataset version. A batch size of 8 was used to mitigate memory constraints during local training, and image caching was disabled to prevent excessive RAM usage. No architectural modifications or custom loss functions were introduced, and default optimization parameters provided by the YOLOv8 framework were retained. This configuration yielded high accuracy. However, inference on additional footage revealed limitations in generalization, motivating a revision of both dataset composition and model configuration.

#### 6.4.3.2 Revised Configuration

In the revised version, the classifier was simplified by adopting the YOLOv8 smaller classification model (YOLOv8s-cls). This reduction in model capacity was intended to discourage overfitting to dataset specific images while maintaining sufficient power for binary recognition. Training was performed for 80 epochs with an input resolution of  $224 \times 224$  pixels, aligned with the revised preprocessing strategy. A batch size of 16 was used, and image caching remained disabled to ensure stable memory usage during training. The number of loading workers was set to 4, and all other optimization parameters were kept at framework defaults.

By reducing input dimensionality, moderating model capacity, and simplifying augmentation, the revised configuration tried to improve robustness and encourage the model to focus on referee posture features rather than background artifacts or dataset specific biases.

## 6.4.4 Quantitative Results

### 6.4.4.1 Initial Baseline Model

The initial version of Model 2 achieved a test accuracy of 0.98 on its validation set, with balanced precision and recall across both classes (*signal* and *no signal*). The confusion matrix indicated symmetric performance, with only three misclassifications out of 186 samples. While these results suggested strong binary discrimination capability, following deployment on additional game footage revealed limited robustness in borderline and transitional referee postures. This discrepancy motivated the dataset expansion and configuration revisions described in previous sections.

### 6.4.4.2 Revised Model — Validation Performance

For the revised configuration, the model achieved a top-1 validation accuracy of 0.99. The corresponding confusion matrix, computed on the validation split (548 samples), is shown in Figure 6.7.

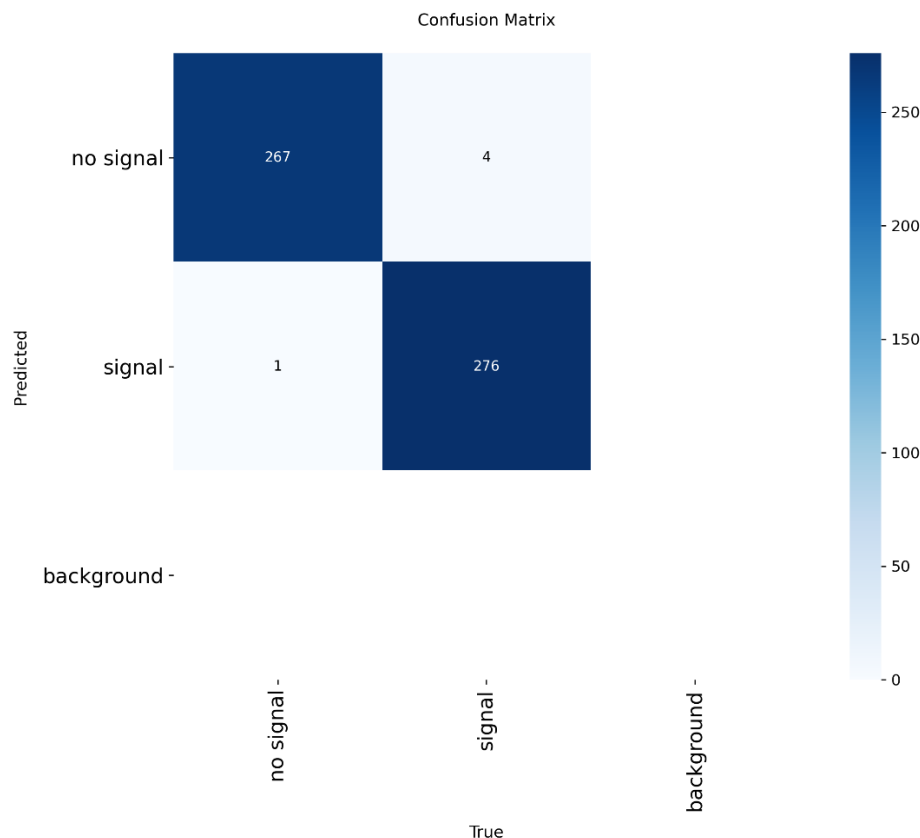


Figure 6.7: Confusion matrix for the revised binary signal classifier evaluated on the validation set.

Out of 548 validation samples:

- 267 *no signal* instances were correctly classified
- 276 *signal* instances were correctly classified

- 4 *signal* instances were misclassified as *no signal* (false negatives)
- 1 *no signal* instance was misclassified as *signal* (false positive)

These results indicate very low misclassification rates for both classes. The slightly higher number of false negatives suggests a very small tendency toward conservative predictions. However, the overall balance between sensitivity and specificity remains strong. The normalized confusion matrix further confirms high class recall for both categories.

#### 6.4.4.3 Revised Model — Test Performance

When evaluated on the test set (547 samples), the revised model achieved a top-1 test accuracy of 0.96. This represents a modest decrease relative to validation performance, reflecting a small but expected generalization gap. The reduction suggests that while the model performs nearly perfectly on validation data, minor performance degradation occurs on completely unseen test samples. Nevertheless, a test accuracy of 0.96 demonstrates robust binary classification capability under diverse conditions, especially given the increased dataset variability introduced in the revised configuration. The combination of high validation performance and strong test accuracy indicates that the simplified architecture and moderated augmentation strategy improved stability compared to the initial baseline model. Overall, the revised Model 2 provides reliable separation between *signal* and *no signal* referee postures, forming a stable intermediate stage in the multi-step pipeline before proceeding to signal type classification.

#### 6.4.5 Qualitative Evaluation

Figure 6.8 presents representative qualitative examples from the revised binary classifier, showing both correct predictions and typical misclassification patterns.

The top row shows correctly classified *signal* instances, where referees show clearly raised arms consistent with standard signaling gestures. Despite variation in camera angle and body orientation, the classifier reliably identifies these postures as signaling actions. The middle row presents correctly classified *no signal* examples, characterized by neutral standing posture without pronounced arm extension. These cases demonstrate that the model does not associate the presence of a referee with signaling behavior.

The bottom row illustrates representative misclassifications. In both examples, the model predicts *signal* for instances that are labeled as *no signal*. Closer inspection reveals arm positioning that may resemble transitional signaling cues. In one case, the referee raises a hand toward the mouth while holding a whistle; in the other, the referee holds the ball with slight arm extension. Although these postures do not correspond to formal signals, they introduce visual ambiguity that may resemble signaling gestures.



*Figure 6.8: Representative qualitative examples from the revised binary signal classifier.*

These qualitative observations align with the quantitative results presented in Section 6.4.6. While the revised model demonstrates strong binary discrimination capability, borderline and transitional postures remain a source of minor classification error. Nevertheless, such cases represent a smaller fraction of overall predictions and reflect visual ambiguity rather than systematic bias.

#### 6.4.6 Key Takeaway

The results of Model 2 demonstrate that binary discrimination between *signal* and *no signal* referee gestures is feasible when classification is restricted to referee centered image crops. The revised dataset configuration and moderated augmentation strategy led to strong validation

performance and stable generalization to an independent test set, with only minor degradation relative to evaluation. Qualitative analysis confirms that correct predictions are associated with distinguishable posture patterns, particularly raised arm gestures for signaling and neutral stance for non-signaling instances. The observed misclassifications arise primarily from transitional or borderline postures in which partial arm elevation resembles signaling behavior. These errors reflect visual ambiguity rather than systematic failure. Overall, Model 2 successfully reduces the complexity of the problem by filtering out non-signaling referee frames prior to multiclass gesture classification.

## 6.5 Model 3 — Signal Type Classification

### 6.5.1 Initial Attempt

The initial formulation of Model 3 aimed to perform multiclass classification of referee signals across 10 distinct gesture categories, derived from the signaling instances identified by Model 2. The dataset comprised 237 manually labeled images extracted from referee detections, distributed across the following classes as shown in Table 6.3.

*Table 6-3: Referee signal class distribution*

| Signal Class                    | Number of Occurances |
|---------------------------------|----------------------|
| direction of play out of bounds | 15                   |
| one free throw                  | 44                   |
| foul                            | 45                   |
| jump ball                       | 19                   |
| kick ball                       | 3                    |
| player number                   | 31                   |
| two free throws                 | 34                   |
| stop the clock                  | 24                   |
| three point attempt             | 13                   |
| three point successful          | 11                   |

Due to the limited number of available samples, the dataset was split into 191 training images, 22 validation images, and 24 test images. To compensate for lack of data, augmentation techniques were applied, including rotation ( $\pm 10^\circ$ ), brightness variation ( $\pm 10\%$ ), minor exposure and shear adjustments (5%), and low-level additive noise (0.1%). After augmentation, the total dataset size increased to 619 images, with 573 used for training.

The model was implemented using the YOLOv8 small classification architecture (YOLOv8s-cls) with default optimization parameters. Images were resized to  $640 \times 640$  pixels, and adaptive contrast equalization was applied during preprocessing.

Despite these measures, the resulting performance was unstable. The confusion matrix revealed substantial class confusion. Given the small validation and test splits, performance metrics were highly sensitive to individual misclassifications, limiting the reliability of quantitative evaluation. These results indicate that the original problem formulation, ten gesture categories trained on a severely limited dataset, was insufficient for robust multiclass learning. The combination of class imbalance, visual similarity among gestures, and inadequate sample size resulted in inconsistent predictions and poor generalization capacity.

This outcome motivated a reformulation of the signal type classification task, focusing on dataset expansion and semantic grouping of visually similar gesture categories, as described in the following subsection.

#### 6.5.2 Dataset Expansion and Class Grouping Strategy

Following the diagnostic failure of the initial multiclass formulation, the signal type classification task was reformulated with two primary objectives. To substantially increase the dataset size and to reduce fine grained class ambiguity through gestures grouping.

##### 6.5.2.1 Signal Extraction from Eight Games

Referee signal crops were generated by applying Model 1 (referee detection) followed by Model 2 (binary signal filtering) to eight full basketball games. The first seven games were designated for training and validation, while the eighth game was entirely held for testing to ensure strict separation between training and evaluation data.

In total, 1,822 signal instances were collected across the eight games. The original signal categories and their total counts are summarized in Table 6.4. It should be noted that the categories *Two Arms Up*, *Arm Up*, and *Uncertain* were not directly mapped to referee signals at the time of extraction. These instances corresponded to gestures where finger semantics were unclear, hand orientation was not identifiable, or the signaling motion was incomplete. Rather than discarding these samples, they were retained for regrouping based on visual arm configuration, as described in the following subsection.

Table 6-4: Distribution of referee signal instances extracted from eight games prior to grouping.

| Original Signal Class    | Total Instances |
|--------------------------|-----------------|
| Stop Clock Open Palm     | 120             |
| Stop Clock Fist – Foul   | 191             |
| Scoring 1 Point          | 97              |
| Scoring 2 Points         | 18              |
| Scoring Attempt 3 Points | 143             |
| Direction of Play        | 76              |
| Number Reporting         | 78              |
| Holding Foul             | 2               |
| Unsportsmanlike Foul     | 21              |
| Jump Ball                | 15              |
| Substitution             | 15              |
| Kick Ball                | 2               |
| Cancel Score             | 2               |
| <i>Two Arms Up</i>       | 156             |
| <i>Arm Up</i>            | 539             |
| <i>Uncertain</i>         | 347             |
| <b>Total</b>             | <b>1,822</b>    |

Although this expanded dataset significantly improved sample availability compared to the initial attempt (237 images), strong visual similarity remained among several gesture categories, particularly those involving raised arms.

#### 6.5.2.2 Class Grouping

To address this issue, visually similar gestures were mapped into four broader categories based on arm configuration and body posture rather than signal type semantics. The final grouped classes were defined as follows:

- One\_arm\_up: Stop Clock Open Palm, Stop Clock Fist – Foul, Scoring 1 Point, Scoring 2 Points, Scoring Attempt 3 Points, Arm Up.
- Both\_arms\_up: Successful\_Scoring\_Attempt\_3 Points, Two Arms Up
- Directional\_point: Direction of Play
- Number\_reporting: Number Reporting

This grouping strategy was motivated by visual Distinguishability. Gestures with a single raised arm share common structural features that are often difficult to distinguish at image level, whereas two arm gestures and directional pointing exhibit distinct posture configurations.

#### 6.5.2.3 Dataset Split and Balancing

Games 1–7 were used for training and validation, while Game 8 was reserved exclusively for testing. This split ensured that evaluation was conducted on completely unseen game footage, reducing the risk of overfitting. Before balancing, the training set showed significant class imbalance, a common challenge in classification tasks where certain classes are naturally underrepresented (He & Garcia, 2009), as the high frequency we observe in one arm up gestures. To mitigate this issue:

- The majority class (*one\_arm\_up*) was downsampled from 780 to 350 instances.
- Minority classes (*directional\_point* and *number\_reporting*) were oversampled to 150 instances each.
- The *both\_arms\_up* class was minimally adjusted to 150 instances.

The resulting balanced training distribution was:

- *one\_arm\_up*: 350
- *both\_arms\_up*: 150
- *directional\_point*: 150
- *number\_reporting*: 150

Validation and test sets retained their natural distributions and were not artificially balanced.

The final train/validation/test distribution is summarized in Table 6.5.

*Table 6-5: Final distribution of grouped referee signal categories across dataset splits.*

| Dataset Split                 | One Arm Up | Both Arms Up | Directional Point | Number Reporting | Total |
|-------------------------------|------------|--------------|-------------------|------------------|-------|
| <b>Train (Balanced)</b>       | 350        | 150          | 150               | 150              | 800   |
| <b>Validation (Game 1–7)</b>  | 87         | 17           | 7                 | 8                | 119   |
| <b>Test (Game 8 – Unseen)</b> | 174        | 36           | 10                | 11               | 231   |

#### 6.5.3 Preprocessing and Augmentation

All referee crops were preprocessed using an aspect ratio preserving resize operation to fit within a square input frame. Images were resized to 224 × 224 pixels by scaling the longer side to the target size and applying zero-padding (black borders) to the remaining region. This padding based resizing avoided geometric stretching, thereby preserving arm proportions and gesture structure, which are central to signal discrimination.

To improve robustness while minimizing distortion, only conservative augmentations were applied during training. Specifically, augmentations were introduced through small random rotations ( $\pm 3^\circ$ ), limited translation (1%), and slight scaling (3%). In addition, mild photometric variation was applied via hue (0.01) and saturation/value adjustments (0.15 each). More aggressive augmentation strategies that can alter spatial semantics, such as mosaic, mixup, cutmix, and random erasing, were explicitly disabled. Finally, label smoothing (0.05) was applied to reduce overconfidence and improve generalization.

#### 6.5.4 Model Configuration

The grouped multiclass signal classifier was implemented using the YOLOv8 small classification architecture (YOLOv8s-cls) initialized with pretrained weights (Jocher & Ultralytics, 2023). The model was trained for 80 epochs with an input resolution of  $224 \times 224$  pixels, consistent with the padding preprocessing strategy described in Section 6.5.3.

Training was performed with a batch size of 32 and 4 data loading workers, using early stopping with a patience of 15 epochs to reduce overfitting. Unless otherwise specified by the configuration parameters above, default optimization settings provided by the YOLOv8 framework were retained, and no architectural modifications were introduced.

This configuration was selected to balance computational efficiency with sufficient capacity for classification across the four grouped signal categories, while maintaining consistent input characteristics with the revised binary classifier in Model 2.

#### 6.5.5 Quantitative Results

##### 6.5.5.1 Validation Performance

The multiclass classifier achieved a top-1 validation accuracy of 0.88 on the validation set derived from Games 1–7. The corresponding confusion matrix is presented in Figure 6.9.

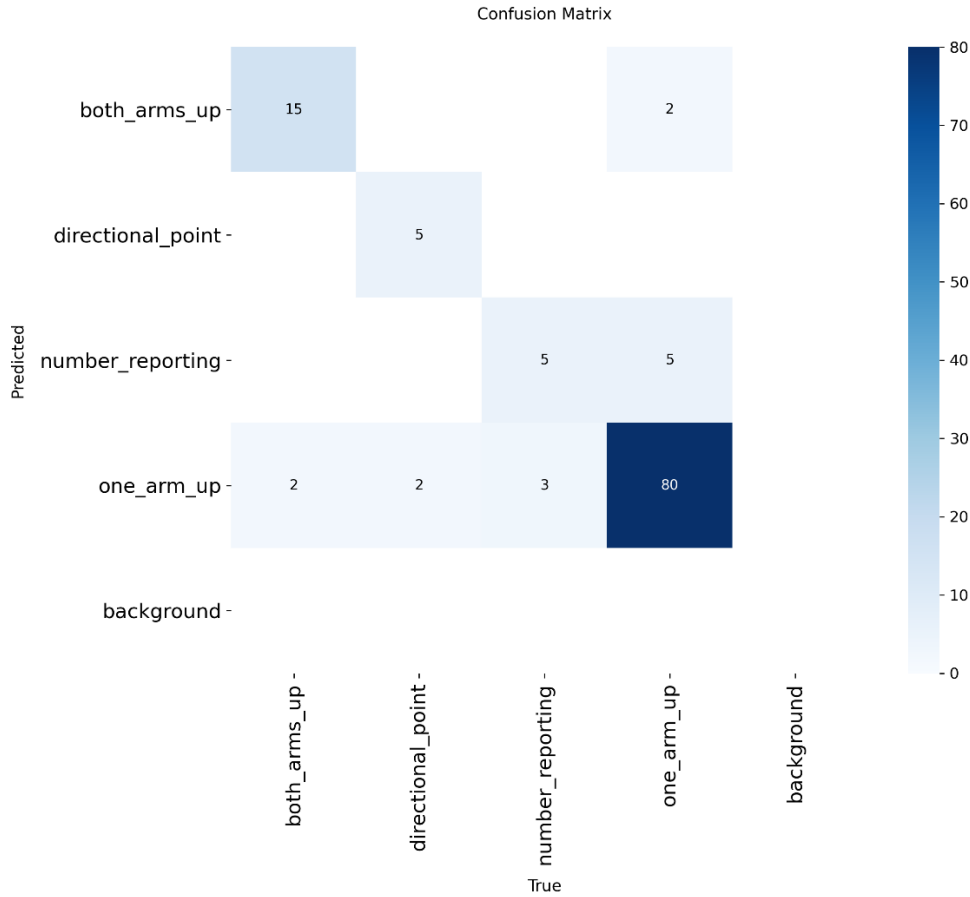


Figure 6.9: Confusion matrix for the multiclass signal classifier evaluated on the validation set.

Across the four grouped categories, performance was strongest for the majority class *one\_arm\_up*, with 80 correct predictions. The classes *both\_arms\_up* and *directional\_point* were also correctly identified in most cases (15 and 5 correct predictions, respectively), while *number\_reporting* exhibited moderate confusion with the *one\_arm\_up* category.

The primary validation errors occurred between visually similar arm configurations. In particular:

- Some *both\_arms\_up* instances were predicted as *one\_arm\_up*.
- Several *number\_reporting* instances were misclassified as *one\_arm\_up*.
- A small number of *one\_arm\_up* gestures were confused with other categories.

These patterns reflect visual overlap between grouped classes, particularly when hand positioning, finger semantics, or camera angle reduce details.

#### 6.5.5.2 Test Performance on Unseen Game

Evaluation on the completely unseen Game 8 test set (231 samples) resulted in a **top-1 test accuracy of 0.81**, indicating a generalization gap of approximately seven percentage points relative to validation performance. The test confusion matrix is shown in Figure 6.10.

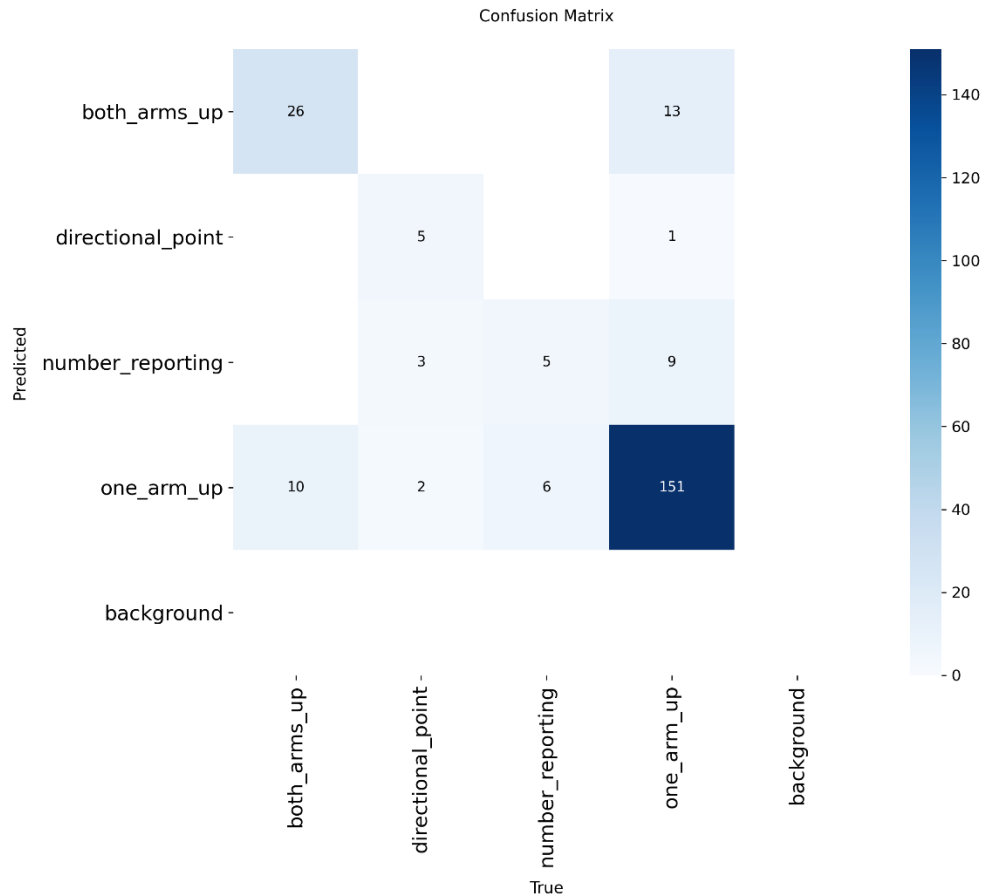


Figure 6.10: Confusion matrix for the multiclass signal classifier evaluated on the test set.

A total of 187 out of 231 samples were correctly classified. As observed during validation, the majority class *one\_arm\_up* achieved the highest number of correct predictions (151 instances), while minority classes experienced greater relative confusion. The dominant error pattern on the test set involved misclassification of all other instances as *one\_arm\_up*. Specifically:

- 10 *both\_arms\_up* instances were predicted as *one\_arm\_up*.
- 6 *number\_reporting* instances were predicted as *one\_arm\_up*.
- 2 *directional\_point* instances were predicted as *one\_arm\_up*.

This behavior suggests that in ambiguous cases, the classifier defaults toward the structurally simpler one arm up configuration. Such confusion is consistent with the visual similarity among raised arm gestures and the reduced frequency of the minority classes in the original (unbalanced) distribution.

Despite the observed gap between validation and test set evaluation, the model maintains meaningful multiclass classification capability under strict game separation. Considering the complexity of posture based gesture recognition from broadcast footage and the class imbalance in the natural test distribution, a test accuracy of 0.81 demonstrates moderate but practically usable performance. Overall, the results indicate that grouping substantially improved stability compared to

the initial ten class formulation; however, fine grained differentiation among similar arm configurations remains challenging.

#### 6.5.6 Qualitative Evaluation

Figure 6.11 presents representative qualitative examples from the grouped multiclass classifier. Both correct predictions and characteristic confusion patterns are illustrated.



Figure 6.11: qualitative examples from the grouped multiclass signal classifier

The top row shows correctly classified examples of *one\_arm\_up* and *number\_reporting*. In the first case, the referee exhibits a clearly extended single arm above shoulder level, forming a strong vertical pose that is visually distinctive. In the second case, finger articulation associated with number reporting is sufficiently visible to allow correct classification. These examples demonstrate that the model is capable of discriminating grouped posture categories when visual cues are clearly expressed.

The remaining rows illustrate typical misclassification patterns observed in the confusion matrix. A common error involves confusion between *one\_arm\_up* and *number\_reporting*. In certain number-reporting gestures, referees raise a single arm while displaying digits with their fingers. When finger configuration is occluded or insufficiently resolved, the gesture may resemble a generic single arm raise, leading the model to predict *one\_arm\_up*. On the other hand, single arm gestures with visible finger articulation may be interpreted as number reporting actions.

Another frequent confusion arises between *both\_arms\_up* and *one\_arm\_up*. In several misclassified instances, both arms are raised, but asymmetrically, one arm fully extended while the other is positioned at head level or partially occluded by camera angle. Under such conditions, the visual symmetry characteristic of two arm gestures becomes less apparent, and the classifier defaults to the dominant single arm category.

Finally, directional point gestures are occasionally misclassified as *one\_arm\_up* when the pointing motion is captured at an intermediate stage or when the arm orientation does not clearly show horizontal directionality. In single frame recognition, transitional motion phases may visually resemble raised arm gestures rather than extended directional pointing.

Overall, the qualitative analysis confirms that most classification errors come from structural overlap among grouped categories and from the limitations of single frame image interpretation. The model performs reliably when gestures exhibit clear configurations but encounters difficulty in ambiguous or transitional frames where postures are partially expressed or visually compressed by camera perspective.

#### 6.5.7 Key Takeaway

The reformulated multiclass signal classifier demonstrates that grouping substantially improves stability compared to the initial ten class formulation. By increasing dataset size and combining visually similar gestures into four categories, the model achieves 0.88 validation accuracy and 0.81 test accuracy on a completely unseen game.

The primary source of error lies in ambiguity among raised arm configurations, particularly between *one\_arm\_up*, *both\_arms\_up*, and *number\_reporting*. These confusions reflect limitations to image based recognition of gesture semantics. Nevertheless, the classifier exhibits meaningful discrimination capability, especially for clearly expressed posture categories.

Importantly, Model 3 operates downstream of both referee detection (Model 1) and binary signal filtering (Model 2), therefore focusing exclusively on referee-centered signaling instances. While gesture interpretation remains challenging, the grouped classification approach provides a structured and computationally feasible solution within the constraints of broadcast footage and image only learning.

This concludes the development of the three stage architecture. The following section evaluates the integrated end-to-end pipeline, including confidence thresholds, inference behavior, and deployment level performance across full game footage.

## 6.6 End-to-End Pipeline Evaluation on Video Input

### 6.6.1 End-to-End Pipeline Workflow

The final system integrates the three trained models into a unified, video processing pipeline designed to operate on continuous broadcast footage. The input to the system is a basketball game video segment, and the output is an annotated video stream in which referees are detected in real time and signaling events are highlighted and classified.

The pipeline operates sequentially as follows:

1. Frame Acquisition

Video frames are read sequentially using OpenCV (Bradski, 2000). The implementation supports recordings at 60 frames per second (fps), although broadcast streams typically operate at approximately 25–30 fps. All temporal parameters in the pipeline are therefore defined relative to the video frame rate.

2. Model 1: Referee Detection and Tracking

Each frame is processed by the trained YOLOv8 detection model (Model 1) (Jocher & Ultralytics, 2023). Detection is performed with persistent tracking enabled (*persist=True*), allowing consistent track IDs to be assigned to referees across consecutive frames. This tracking mechanism is essential for maintaining per referee temporal state (e.g. consecutive frame counting and voting logic).

Detected referees are enclosed by bounding boxes. During normal (non-signal) conditions, these boxes are rendered in green in the output video.

3. Bounding Box Expansion and Cropping

For each detected referee, the bounding box is expanded by a fixed padding of 16 pixels on all sides before cropping. This padding ensures that arm and hand regions, critical for signal recognition, are not truncated. The padded crop is clipped to image boundaries and passed to the classification models.

#### 4. Model 2: Binary Signal Gating

The cropped referee image is processed by the binary classifier (Model 2), which predicts either *signal* or *no signal*. A confidence threshold ( $SIG\_THRESH = 0.8$ ) is applied to the top-1 probability. Only predictions labeled as *signal* with confidence exceeding this threshold are considered positive detections.

To prevent false activations due to single frame noise, a signal event is triggered only if a referee produces at least nine consecutive frames classified as *signal* ( $CONSEC\_FRAMES\_REQUIRED = 9$ ). At 30 fps, this corresponds to approximately 0.3 seconds of sustained signal evidence.

#### 5. Sticky Event Activation

Once the consecutive frame requirement is satisfied, a *sticky event* is activated for the corresponding referee. During this period, which lasts 1.5 seconds, the referee bounding box is rendered in red, and the text “SIGNAL” is displayed. The sticky window ensures that detected events remain visible for human interpretation and provides a temporal window for multiclass voting.

#### 6. Model 3: Multiclass Voting within a Temporal Window

During the sticky event, the grouped multiclass classifier (Model 3) is executed within a restricted voting window. Voting begins 0.5 seconds after trigger activation ( $VOTE\_START\_SEC$ ) to avoid unstable transitional frames and stops 0.2 seconds before the sticky window ends ( $VOTE\_END\_SEC$ ) to exclude return to neutral positions.

To reduce computational overhead while maintaining stability, Model 3 is executed every second frame during the voting interval ( $MODEL3\_STRIDE = 2$ ). Each prediction contributes one vote to the corresponding class count for that referee.

#### 7. Label Stabilization and Locking

A class label is displayed only after a minimum number of votes has been collected ( $MIN\_VOTES\_TO\_SHOW = 6$ ). Once ten votes are collected ( $LOCK\_AFTER\_VOTES = 10$ ), the label may be permanently locked if the leading class exceeds the runner up by at least three votes ( $LOCK\_MARGIN = 3$ ). This mechanism ensures visually stable output. If no stable class has yet been determined, the system displays only the generic “SIGNAL” tag.

#### 8. Overlay Rendering

The final annotated frame is written to the output video. Referees without active signals remain highlighted in green, while signaling referees appear in red with the corresponding class label (e.g. “SIGNAL: directional\_point”).

This architecture enables temporal reasoning while preserving computational efficiency. Detection, gating, voting, and visualization are all performed online, frame by frame, without full video buffering. Although the pipeline was evaluated on short representative video segments, its design is compatible with continuous live broadcast input.

#### 6.6.2 Parameter Selection and Tuning

The end-to-end pipeline includes several tunable parameters that control spatial context, temporal stability, and classification confidence. These parameters were adjusted to balance responsiveness, robustness against false positives, and computational efficiency when processing continuous video input. Table 6.6 summarizes the tunable parameters and their roles within the integrated system.

*Table 6-6: Tunable parameters used*

| Parameter              | Model(s) Affected      | Description                                    | Default Value  |
|------------------------|------------------------|------------------------------------------------|----------------|
| PADDING                | Model 1 → Models 2 & 3 | Bounding box expansion before cropping         | 16 pixels      |
| USE_TRACKING           | Model 1 + Pipeline     | Enables persistent track IDs across frames     | TRUE           |
| SIG_THRESH             | Model 2                | Minimum confidence for signal classification   | 0.8            |
| CONSEC_FRAMES_REQUIRED | Model 2 + Event Logic  | Required consecutive signal frames to trigger  | 9 frames       |
| STICKY_SECONDS         | Event Logic            | Duration of signal display                     | 1.5 seconds    |
| VOTE_START_SEC         | Model 3                | Delay before voting begins after trigger       | 0.5 seconds    |
| VOTE_END_SEC           | Model 3                | Voting cutoff before sticky ends               | 0.2 seconds    |
| MODEL3_STRIDE          | Model 3                | Frequency of Model 3 execution during voting   | Every 2 frames |
| MIN_VOTES_TO_SHOW      | Model 3                | Minimum votes before displaying class label    | 6 votes        |
| LOCK_AFTER_VOTES       | Model 3                | Votes required before attempting label lock    | 10 votes       |
| LOCK_MARGIN            | Model 3                | Required vote margin to permanently lock label | 3 votes        |

- Spatial Context (Bounding Box Padding)

The padding parameter (16 pixels) expands the detected referee bounding. This expansion ensures that raised arms and hand gestures are fully contained within the crop, particularly in wide broadcast views. Smaller padding values occasionally truncated arms, whereas larger values introduced excessive background noise. A moderate padding of 16 pixels provided a balanced trade off.

- Binary Signal Gating Threshold

The Model 2 confidence threshold ( $SIG\_THRESH = 0.8$ ) represents the primary precision–recall trade off in the system. Lower thresholds increased sensitivity but introduced false triggers in transitional frames. Higher thresholds reduced false positives but occasionally suppressed legitimate short duration signals. A value of 0.8 provided stable event detection while maintaining acceptable responsiveness.

- Consecutive Frame Requirement

The requirement of nine consecutive signal predictions ( $CONSEC\_FRAMES\_REQUIRED = 9$ ) ensures temporal consistency before triggering an event. At 30 fps, this corresponds to approximately 0.3 seconds of sustained signal evidence. This mechanism filters out isolated misclassifications and significantly reduces false activations caused by brief posture fluctuations.

- Sticky Display Duration

The sticky window of 1.5 seconds provides a visually interpretable duration for detected events. This parameter does not affect detection accuracy but improves usability by preventing flickering overlays and allowing sufficient time for class stabilization.

- Multiclass Voting Window

The Model 3 voting window begins 0.5 seconds after trigger activation and ends 0.2 seconds before the sticky window concludes. This design intentionally excludes:

- Early transition frames (gesture initiation)
- Late return to neutral frames (ending gesture)

Such transitional frames were observed to introduce instability in single-frame classification. Restricting the voting window improved class consistency. Additionally, Model 3 is executed every second frame ( $MODEL3\_STRIDE = 2$ ) to reduce computational load while preserving sufficient temporal resolution for stable majority voting.

- Label Stabilization and Locking

To prevent label instabilities during ambiguous sequences, a two stage stabilization mechanism is applied:

- A minimum of six votes must be collected before displaying a class label.
- After ten votes, the predicted class may be permanently locked if it exceeds the runner up by at least three votes.

This locking prevents unstable switching between visually similar classes such as *one\_arm\_up* and *number\_reporting*.

Overall, parameter tuning focused on achieving stable event detection without sacrificing responsiveness. The chosen configuration reflects a balance between precision, temporal smoothness, and computational feasibility for real time video processing.

### 6.6.3 Temporal Aggregation and Event Logic

Although each individual model in the pipeline operates on single frames, the integrated system incorporates structured temporal reasoning to improve robustness and reduce instability. This design addresses three primary challenges of frame based gesture recognition: (i) short misclassifications, (ii) ambiguous transition frames, (iii) label flickering.

- Consecutive Frame Signal Triggering

The first temporal mechanism is applied at the binary gating stage (Model 2). A signal event is triggered only after nine consecutive frames are classified as *signal* with confidence above the predefined threshold. This consecutive frame requirement suppresses isolated false positives caused by momentary signal misclassification. Without this mechanism, single frame noise could activate false signal events. By enforcing temporal continuity, the system ensures that only sustained gesture evidence leads to event activation.

- Sticky Event Window

Once triggered, a sticky window of 1.5 seconds maintains the signal state even if subsequent frames briefly fail the Model 2 threshold. This prevents hurried disappearance of the signal overlay during minor fluctuations in posture or confidence. The sticky window serves two purposes. Improves visual interpretability of the output video and provides a structured interval for multiclass voting (Model 3).

- Restricted Voting Window for Multiclass Classification

Multiclass voting is intentionally restricted to a central portion of the sticky interval. Voting begins 0.5 seconds after activation and stops 0.2 seconds before the sticky window ends. This design excludes frames where the arm may still be rising and frames where the referee returns to a normal state. Observation showed that these transitional frames often produced inconsistent multiclass predictions. By focusing voting on the temporally stable core of the gesture, the system increases classification consistency.

- Majority Voting and Strided Inference

During the voting window, Model 3 is executed every second frame. Each prediction contributes one vote toward a class counter. This execution reduces computational overhead while maintaining sufficient temporal sampling for stable majority estimation. The final class label is determined by the majority of accumulated votes rather than by a single frame decision. This voting strategy mitigates the impact of occasional misclassifications and naturally smooths noisy predictions.

- Label Locking and Margin Enforcement

To further enhance stability, the pipeline employs a margin locking mechanism. After a minimum number of votes is collected, the predicted label may be permanently locked if the leading class exceeds the runner up by a predefined margin. This mechanism prevents rapid changes between visually similar classes, particularly in borderline cases. Once locked, the label remains fixed for the remainder of the sticky window.

The combination of consecutive frame gating, sticky windows, restricted voting intervals, majority aggregation, and margin based locking transforms frame level models into a temporally logical event detection system. The pipeline's structured temporal logic provides practical robustness under broadcast video constraints. The approach remains computationally lightweight while significantly reducing false triggers and label instability compared to false per frame classification.

#### 6.6.4 End-to-End Qualitative Demonstration

Figure 6.12 presents representative frames from the fully integrated pipeline operating on a representative broadcast video segment. The screenshots show the behavior of the system under realistic game conditions.



Figure 6.12: representative frames from the fully integrated pipeline

In the top frame, three referees are visible simultaneously. One referee remains inactive and is enclosed by a green bounding box, indicating normal detection without signal activation. The other two referees are in the early stage of a foul call. At this point, the consecutive frame requirement has been satisfied, and the sticky window has been triggered. Their bounding boxes are rendered in red, and the generic label “`SIGNAL`” is displayed. The specific signal class has not yet appeared because the Model 3 voting mechanism is still accumulating sufficient votes for stable classification. This example demonstrates the temporal gating logic.

The middle frame shows a stabilized multiclass prediction. Two referees are visible. The referee on the right is inactive (green bounding box), while the referee on the left is reporting a player number to the table. The system correctly displays a red bounding box and the label “`SIGNAL: number_reporting`.” This frame demonstrates successful integration of detection, binary gating, multiclass voting, and label locking.

The bottom frame shows a directional call. One referee remains inactive (green), while the other is signaling direction of possession. The bounding box is red and labeled “`SIGNAL: directional_point`.” Notably, the arm posture in this frame is captured near the end of the signaling event and is not fully extended. Nevertheless, due to the temporal voting and label locking, the correct class remains displayed. This behavior shows the benefit of excluding early and late transitional frames from the voting window, thereby preventing label changes during gesture initiation or completion.

Across these examples, the pipeline demonstrates the following characteristics:

- Simultaneous multi referee tracking and independent per track event logic.
- Separation between detection (green) and confirmed signal events (red).
- Stable multiclass labeling through majority voting and margin-based locking.
- Smooth visual output without rapid flickering or label flipping.

These qualitative results confirm that the integrated architecture functions sufficiently under realistic broadcast conditions. While individual frame level predictions may show ambiguity, the structured temporal logic transforms the three stage model stack into a stable event detection system capable of producing interpretable video overlays.

#### 6.6.5 Practical Performance Considerations

Beyond classification accuracy, the integrated pipeline was evaluated with respect to computational efficiency, memory usage, and runtime stability when processing continuous video input.

##### 1. Frame-by-Frame Processing and Memory Efficiency

The system operates in a strictly frame by frame manner without buffering the full video in memory. Each frame is processed sequentially and immediately written to the output stream

after annotation. This design ensures stable memory consumption independent of video duration and avoids RAM saturation, which was previously encountered during early experimentation.

## 2. Computational Load Distribution

The computational burden is distributed across the three models:

- Model 1 (Detection + Tracking) runs on every frame.
- Model 2 (Binary Classification) runs on each detected referee.
- Model 3 (Multiclass Classification) runs only during activated sticky windows and only within a restricted voting interval.

Importantly, Model 3 is executed every second frame during voting ( $MODEL3\_STRIDE = 2$ ), effectively reducing multiclass inference calls by approximately 50% within the voting window. This execution significantly lowers computational overhead while preserving sufficient temporal sampling for stable majority voting.

## 3. Latency Considerations

The pipeline introduces intentional temporal latency due to consecutive frame gating and voting delay. The nine frame consecutive requirement corresponds to approximately 0.3 seconds at 30 fps, while the voting window begins 0.5 seconds after trigger activation. Consequently, class labels may appear slightly after the initial physical gesture is displayed. This delay represents a deliberate design trade off that favors stability and precision over quick response.

## 4. Stability Under Multi Referee Conditions

The use of persistent tracking IDs allows independent state management per referee. Each tracked referee maintains its own:

- Consecutive-frame counter
- Sticky activation timer
- Multiclass vote history
- Label lock state

This logic prevents interference between referees when multiple signals occur simultaneously. As demonstrated in the qualitative examples, the system successfully handles parallel detection and independent event triggering.

## 5. Limitations and Engineering Trade offs

While computationally efficient and stable, the pipeline remains dependent on:

- Accurate initial referee detection (Model 1)
- Reliable binary gating (Model 2)

- Sufficient visual clarity for grouped classification (Model 3)

Failure at earlier stages propagates downstream. In addition, the absence of motion limits the system’s ability to exploit temporal dynamic. Nevertheless, the implemented architecture achieves a practical balance between robustness, interpretability, and computational feasibility under broadcast video constraints.

#### 6.6.6 Quantitative Evaluation of the Integrated Pipeline

While the previous sections demonstrated the operation of the proposed pipeline through qualitative examples, an additional quantitative evaluation was conducted to assess the performance of the complete end-to-end system on previously unseen broadcast footage. Unlike the individual model evaluations presented in previous sections 6.3–6.5, this experiment evaluates the combined behavior of all three models operating sequentially under realistic conditions.

For this purpose, a EuroLeague game that had not been used during the training of any of the three models was selected. A total of 56 referee gesture events were manually annotated by recording the start and end timestamps of each signal together with its corresponding ground truth class. The annotated events included the four gesture categories recognized by the proposed pipeline: *one arm up*, *both arms up*, *directional point*, and *number reporting*.

The complete pipeline was then executed on the same video using the final parameter configuration described in the previous sections. During inference, the system automatically recorded the temporal boundaries of each detected event, the predicted gesture class, the temporal voting statistics, and the confidence values produced by Models 2 and 3. The predicted events were then matched manually with the annotated actual events based on their temporal overlap in order to evaluate the performance of the integrated system at an event level.

Figure 6.13 presents the resulting event level confusion matrix. In contrast to a conventional multi class classification problem, two additional prediction outcomes were considered. The generic signal category corresponds to events for which the system detected the presence of a referee signal but the temporal voting mechanism did not gather sufficient evidence to assign a stable gesture class. The missed category represents manually annotated events for which no corresponding pipeline detection was produced.

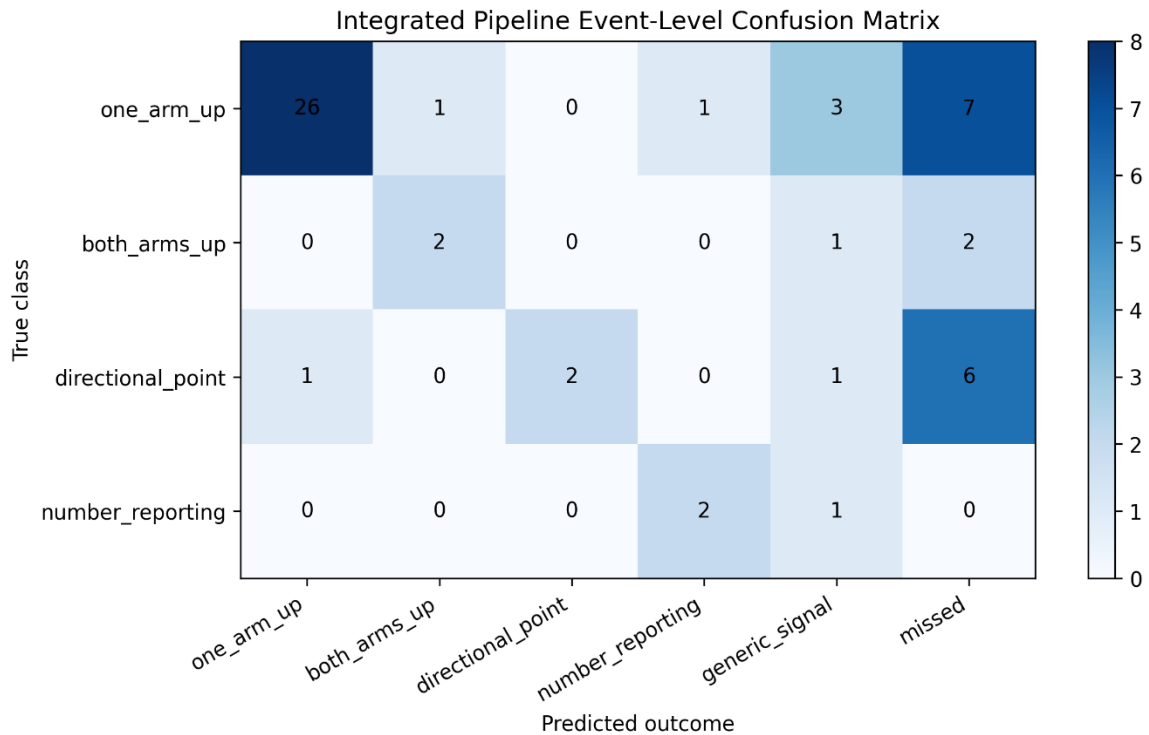


Figure 6.13: Event level confusion matrix of the integrated pipeline

Table 6.7 summarizes the overall event level evaluation. Out of the 56 manually annotated referee gestures, the proposed pipeline correctly recognized 32 events (57.1%), while only three events (5.4%) were assigned an incorrect gesture class. Six events (10.7%) were detected as referee signals but remained unclassified due to insufficient temporal voting confidence, resulting in a generic signal output. Finally, 15 events (26.8%) were missed entirely by the integrated system.

Table 6-7: Event level evaluation

| Evaluation Result | Result Count # | Result Count % |
|-------------------|----------------|----------------|
| Correct           | 32             | 57.1%          |
| Generic signal    | 6              | 10.7%          |
| Missed            | 15             | 26.8%          |
| Wrong class       | 3              | 5.4%           |
| Grand Total       | 56             | 100.0%         |

The confusion matrix further indicates that incorrect gesture classification occurred relatively infrequently compared to missed detections. This observation suggests that, once a stable gesture label was produced, the multiclass classifier generally distinguished the four gesture categories reliably. Most failures originated from the earlier stages of the pipeline, where challenging viewing angles, partial referee occlusions, rapid camera transitions, or short gesture durations prevented the system from producing a stable event suitable for classification.

During inference, the pipeline produced a total of 116 event detections throughout the evaluated video segment. After manual matching with the ground truth annotations, 37 detections corresponded to annotated referee gestures, whereas the remaining 79 detections did not have a matching annotated event. Among the 79 unmatched detections, 58 never reached the confidence threshold required to lock a gesture class and therefore remained as generic signal outputs. The remaining 21 unmatched detections – false positives - produced a stable gesture classification despite not corresponding to any manually annotated event. This behavior demonstrates that the temporal voting mechanism suppresses many uncertain classifications by retaining a generic signal indication instead of forcing a potentially incorrect gesture label.

#### 6.6.7 Key Takeaways

The final integrated pipeline demonstrates that referee signal recognition in broadcast basketball footage can be effectively addressed through a structured, multi stage architecture rather than a single end to end model.

Several key conclusions emerge from the system-level evaluation:

1. Detection Centric Design Is Crucial

Isolating referees from the broadcast frame (Model 1) proved to be a necessary prerequisite for reliable gesture recognition. Direct broadcast frame classification was shown earlier in this chapter to rely heavily on contextual shortcuts. In contrast, the detection strategy enforces gesture centric reasoning by constraining the visual input to referee regions.

2. Binary Gating Stabilizes the Pipeline

The introduction of a dedicated signal versus no-signal classifier (Model 2) functions as a gate before multiclass interpretation by, setting confidence threshold, consecutive frame filtering and introducing sticky activation logic. This intermediate binary stage reduces unnecessary multiclass inference calls and prevents class activations during non-signal referee motion.

3. Temporal Voting Is More Effective Than Frame Level Decisions

Instead of relying on single frame multiclass predictions, the system employs a delayed voting window, the majority aggregation logic and the margin label locking. This strategy compensates for transitional poses, partial occlusions, and momentary ambiguities. The qualitative demonstrations, together with the quantitative event level evaluation presented in Section 6.6.6, indicate that even when individual frames are imperfect, the aggregated decision remains stable and interpretable.

4. Class Grouping Was Necessary for Practical Performance

The initial attempt at multiclass recognition with numerous low sample categories resulted in unstable and inconsistent predictions. The subsequent grouping of visually similar gestures

into four semantically coherent categories produced substantially improved generalization and class separability. This confirms that dataset size, inter-class similarity, and broadcast resolution impose practical limits on granularity in referee gesture recognition.

#### 5. The Pipeline Is Technically Feasible on Broadcast Video

The system successfully processes broadcast video segments, producing:

- Continuous referee tracking
- Clear visual distinction between inactive and signaling states
- Stable multiclass labels
- Smooth, non-flickering overlays

While not optimized for strict real time officiating assistance, the architecture demonstrates technical feasibility for automated tagging.

Across all experimental stages, from direct broadcast classification to the final integrated multi-model pipeline, the results indicate that:

- Referee-centric modeling outperforms context driven shortcuts.
- Structured temporal logic improves robustness without requiring full video based deep learning.
- Data curation and class design are as critical as model selection.

The developed pipeline represents a reasonable, interpretable, and computationally manageable solution for referee signal recognition in broadcast basketball footage.

## 7 Discussion and Conclusions

### 7.1 Overview of the Proposed Pipeline and Main Findings

This chapter interprets the experimental findings presented in Chapter 6 within the broader context of computer vision based gesture recognition. The objective is not to revisit the quantitative results, but to critically assess the effectiveness of the proposed multi stage referee-centric pipeline, examine its methodological implications, and identify its limitations and future potential.

The central contribution of this thesis is the design and evaluation of a three stage system for automatic recognition of basketball referee hand signals from broadcast video. The pipeline consists of referee detection using YOLOv8 (Jocher & Ultralytics, 2023), binary classification of cropped referee images into signal versus no-signal, and multiclass signal type classification with temporal aggregation and voting mechanisms. This design was motivated by observations regarding contextual bias in broadcast footage and by the conceptual limitations of full frame image classification.

A broadcast based classifier was first implemented to evaluate whether signal recognition could be achieved directly from full frame images. Although this approach achieved competitive, further analysis revealed that the model relied heavily on contextual cues such as player positioning, court layout, and game flow rather than the referee's gesture itself. This phenomenon aligns with the concept of shortcut learning, where models exploit false correlations that are predictive within a specific dataset but fail to generalize under distributional shift (Geirhos, et al., 2020). The baseline therefore demonstrated that high performance on static broadcast frames does not necessarily indicate robust gesture understanding.

In contrast, the referee-centric pipeline explicitly isolates the referee responsible for the signal before performing classification. Model 1 (referee detection) showed reliable localization performance across full game frames, enabling automated extraction of referee crops while maintaining computational efficiency through streaming inference. Model 2 (signal versus no-signal classification) initially achieved very high test accuracy but exhibited poor generalization, revealing the impact of dataset realism and augmentation choices. Subsequent retraining with more diverse examples improved robustness. Model 3 (signal type classification), constrained by limited data, required class grouping and balanced training strategies to achieve stable performance.

An additional key finding concerns the importance of temporal aggregation. Instead of adopting end to end video models, temporal consistency was enforced at the system level through consecutive frame logic, majority voting, and vote locking mechanisms. This design significantly reduced prediction instability while preserving interpretability.

Overall, the results demonstrate that referee signal recognition from broadcast video is feasible when approached through a structured, referee-centric pipeline. At the same time, the experiments

reveal the sensitivity of gesture recognition models to domain shift, dataset construction decisions, and contextual bias. Collectively, the results suggest that referee signal recognition is not merely a classification problem but a structured perception task that benefits from referee isolation and realistic dataset design. The experiments reinforce the notion that evaluation metrics alone are insufficient to assess model robustness. Qualitative inspection, cross-game testing, and inference are essential for understanding real world behavior.

## 7.2 Methodological Insights

The results underline the importance of problem framing. Treating referee signal recognition as a full frame image classification task initially appeared efficient but proved conceptually misaligned with the objective of gesture understanding. Broadcast footage contains dense contextual information, players, scoreboard overlays, audience background, and court markings, that can dominate the learning process. The experimental comparison demonstrated that explicitly isolating the signal performer fundamentally changes the nature of the learning task. By restructuring the problem into detection followed by classification, the model is constrained to focus on gesture relevant features rather than spurious correlations.

The experiments emphasize the critical role of dataset realism. Augmentation is often employed to improve robustness and increase effective dataset size. However, excessive or poorly calibrated augmentation may distort the visual distribution in ways that reduce validity. In this project, strong augmentation contributed to inflated validation performance while masking weaknesses. The retraining of Model 2 with more representative and manually reviewed examples demonstrated that distribution precision is more valuable than synthetic diversity in small sized datasets. Particularly in binary classification tasks involving subtle pose differences, the quality and realism of negative examples proved decisive.

Furthermore, the pipeline highlights the benefits of modular system design. Separating detection, binary filtering, and multiclass classification provided several advantages:

- Clear interpretability at each stage.
- Independent retraining and improvement of components.
- Reduced propagation of contextual bias.
- Computational flexibility suitable for mid-range hardware.

This modularity also enabled targeted debugging. For example, generalization issues in Model 2 could be analyzed without blending them with detection errors. In contrast, a single end to end model would have obscured the source of failure.

The handling of temporal information at the system level represents a deliberate methodological choice. While many gesture recognition approaches rely on video based architectures,

such models require dense temporal annotation, higher computational cost, and more complex training pipelines. Instead, this thesis incorporated temporal reasoning through lightweight aggregation strategies applied to frame level predictions. Consecutive frame thresholds, majority voting, and vote locking mechanisms provided temporal stability without sacrificing interpretability or scalability. This approach demonstrates that temporal consistency can be indeed achieved through structured post processing.

Finally, the experiments reveal the broader challenge of domain shift in visual recognition systems and sports analytics (Torralba & Efros, 2011). Broadcast production styles, camera angles, lighting conditions, and referee uniforms vary across games and competitions. Even within the same league, distributional differences affected model behavior. This reinforces the necessity of cross game validation and iterative dataset expansion when developing models intended for real world deployment. Taken together, these methodological observations suggest that successful gesture recognition in broadcast environments depends not only on model capacity but on thoughtful task decomposition, realistic data, and system level design decisions. The following section outlines the primary limitations of the current study.

### 7.3 Limitations

Despite the promising results of the proposed referee-centric pipeline, several limitations must be acknowledged. First, the overall dataset size remains relatively constrained, particularly for multiclass signal classification. While the detection and binary models benefited from several thousand training instances after augmentation, Model 3 operated on a substantially smaller set of manually filtered signal images. Even after grouping signals into four broader semantic categories, class imbalance and limited variability restricted the model’s capacity to learn fine grained distinctions. Data limitation likely contributed to confusion between visually similar arm configurations and reduced generalization to new viewing conditions.

Second, the dataset is restricted to a single competition (EuroLeague) and a limited number of recorded games. Although cross game validation was performed, the visual distribution remains relatively homogeneous in terms of referee uniforms, broadcast production style, and arena environment. As a result, the system’s robustness to other leagues, camera setups, or lighting conditions cannot be fully established. Broader evaluation across multiple competitions would be necessary to confirm generalization.

Third, annotation quality represent a limitation. Manual labeling of signal versus no-signal frames, as well as multiclass gesture categorization, required human judgment. Borderline cases, such as transitional arm movements or incomplete signals, introduce ambiguity. While careful review

reduced labeling noise, some degree of subjectivity is unavoidable in gesture recognition tasks, especially when using static frames rather than temporally contextualized sequences.

Fourth, the system relies exclusively on static images. Although temporal aggregation was implemented at the pipeline level, the models themselves do not explicitly learn motion dynamics. Certain referee signals are partially defined by movement trajectories rather than static arm positions. Consequently, the absence of explicit motion modeling may limit recognition of signals that are ambiguous in single frames but clear when observed over time.

Fifth, real time deployment constraints were not benchmarked. While the system was designed with computational efficiency in mind and successfully operated on an RTX 2070 SUPER GPU, a comprehensive evaluation of latency and stability in a live broadcast scenario was beyond the scope of this thesis.

Finally, the modular design, while advantageous for interpretability and debugging, introduces cumulative error propagation. Imperfect detection can affect binary filtering, and binary misclassification can reduce the available data for multiclass prediction. Although confidence thresholds and temporal voting mitigated these effects, complete elimination of upstream error influence is not possible within a staged architecture.

These limitations do not undermine the core contribution of the thesis but rather outline the boundaries within which the proposed system has been validated. The next section outlines potential directions for addressing these constraints and extending the work in future research.

## 7.4 Future Work

The findings and limitations identified in this study open several avenues for future research. A primary direction concerns dataset expansion and diversification. Collecting and annotating referee signals from additional leagues, competitions, and broadcast environments would significantly improve robustness to domain shift. Variations in referee uniforms, camera placement, lighting conditions, and production style should be explicitly incorporated into future datasets. A larger dataset would also enable finer multiclass classification without requiring semantic grouping of signal categories.

Future work could explore dedicated referee focused recording setups. Broadcast footage is optimized for viewer experience rather than analytical precision. Camera zoom, dynamic framing, and occlusions introduce variability that complicates gesture recognition. A controlled camera positioned to consistently capture referees could dramatically improve signal visibility.

Moreover, temporal modeling could be incorporated directly into the learning architecture. While this thesis employed frame level classification combined with temporal aggregation logic, sequence based models may better capture motion dynamics in certain signals. Architectures such as

3D convolutional networks (Tran, Bourdev, Fergus, Torresani, & Paluri), recurrent neural networks, or transformer based video models could learn motion trajectories and temporal dependencies directly from short video clips.

Finally, the modular nature of the proposed architecture allows integration into broader sports analytics platforms. Referee signal recognition could be combined with event detection, scoreboard parsing, or player tracking to produce structured, machine readable game logs. Such integration would extend the impact of the current work beyond isolated gesture recognition toward automated game understanding.

Collectively, these directions demonstrate that the proposed pipeline constitutes a step rather than a final solution. With expanded data, richer temporal modeling, and system level optimization, referee signal recognition can evolve into a robust component of intelligent sports analysis systems.

## 7.5 Final Conclusion

This thesis investigated the feasibility of automatic basketball referee signal recognition from broadcast video using a structured, multi stage computer vision pipeline. It proposed a referee centric architecture consisting of detection, binary filtering, and multiclass signal classification, complemented by temporal aggregation mechanisms. By isolating the referee prior to classification, the proposed pipeline enforced a more semantically aligned learning objective. It further highlighted the importance of dataset realism, careful augmentation strategies, and representative negative samples in binary classification tasks. It also showed that, under constrained data conditions, grouping semantically related gestures can meaningfully enhance stability and practical usability. Although limitations remain, the results confirm that referee signal recognition is achievable within a structured and computationally efficient framework.

In conclusion, this thesis contributes not only an implemented and evaluated referee signal recognition pipeline, but also methodological insights into dataset bias, architectural decomposition, and the role of temporal reasoning in gesture recognition. The findings emphasize that careful task formulation and system design are as critical as model selection in achieving reliable performance in real world computer vision applications.

## References

- Aggarwal, J., & Ryoo, M. (2011). Human activity analysis: A review. *ACM Computing Surveys*, 43(3), 16:1–16:43. doi:10.1145/1922649.1922653
- Bedico, S., Lope, E., Lope, E., Lunjas, E., Lustre, A., & Tolentino, R. (2020). Gesture recognition of basketball referee violation signal by applying dynamic time warping algorithm using a wearable device. *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, (pp. 249-254). doi:10.1109/ICCMC48092.2020.ICCMC-00047
- Bradski, G. (2000). The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.
- Cao, Z., Hidalgo Martinez, G., Simon, T., Wei, S., & Sheikh, Y. (2017). Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 1302-1310). doi:10.1109/CVPR.2017.143
- FIBA- International Basketball Federation. (2024). *Official Basketball Rules 2024*. FIBA. Retrieved from <https://assets.fiba.basketball/image/upload/documents-corporate-fiba-official-rules-2024-v10a.pdf>
- Geirhos, R., Jacobsen, J., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2, 665–673. doi:10.1038/s42256-020-00257-z
- He, H., & Garcia, E. (2009). Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), 1263–1284. doi:10.1109/TKDE.2008.239
- Jocher, G., & Ultralytics. (2023). YOLOv8. *Computer software*. GitHub. Retrieved from <https://github.com/ultralytics/ultralytics>
- Moeslund, T., & Granum, E. (2001). A Survey of Computer Vision-Based Human Motion Capture. *Computer Vision and Image Understanding*, 81(3), 231–268. doi:10.1006/cviu.2000.0897
- Poppe, R. (2010). A survey on vision-based human action recognition. *Image and Vision Computing*, 28(6), 976–990. doi:10.1016/j.imavis.2009.11.014
- Roboflow, U. (2024). Basketball Referee Signals 2 Dataset. *Dataset*. Roboflow. Retrieved from <https://universe.roboflow.com/tfg-iz8nx/basketball-referee-signals-2/dataset/1>
- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437. doi:10.1016/j.ipm.2009.03.002
- Torralba, A., & Efros, A. (2011). Unbiased look at dataset bias. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, (pp. 1521–1528). doi:10.1109/CVPR.2011.5995347
- Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (n.d.).

- Wang, B., HoKun, Y., & Zhou, C. (2024). A real-time basketball referee gesture recognition system based on multi-scale spatio-temporal features. *Applied Mathematics and Nonlinear Sciences*, 9(1), 1–23. doi:10.2478/amns-2024-0050
- Yeh, C., Pan, T., & Hu, M. (2017). A Sensor-Based Official Basketball Referee Signals Recognition System Using Deep Belief Networks. In L. Amsaleg, G. Guðmundsson, C. Gurrin, B. Jónsson, & S. Satoh (Eds.), *MultiMedia Modeling (MMM 2017)* (Vol. 10132, pp. 565–575). Springer. doi:10.1007/978-3-319-51811-4\_46
- Žemgulys, J., Raudonis, V., Maskeliūnas, R., & Damaševičius, R. (2018). Recognition of basketball referee signals from videos using Histogram of Oriented Gradients (HOG) and Support Vector Machine (SVM). *Procedia Computer Science*, 130, 953–960. doi:10.1016/j.procs.2018.04.095
- Žemgulys, J., Raudonis, v., Maskeliūnas, R., & Damaševičius, R. (2020). Recognition of basketball referee signals from real-time video. *Journal of Ambient Intelligence and Humanized Computing*, 11, 979–991. doi:10.1007/s12652-019-01209-1