Enhancing soccer analysis through computer vision: A study on player detection in broadcast video

Thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science
in
Computer Science and Engineering by Research

by

Chris Andrew Gadde
2018701019
chris.andrew@research.iiit.ac.in

International Institute of Information Technology
(Deemed to be University)
Hyderabad - 500 032, INDIA
December 2023

International Institute of Information Technology
Hyderabad, India

CERTIFICATE

It is certified that the work contained in this thesis, titled **" Enhancing soccer analysis through computer vision: A study on player detection in broadcast video"** by **Chris Andrew Gadde**, has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Adviser: Prof. C.V. Jawahar

To My Mother

Acknowledgments

I express my sincere gratitude to Professor C.V. Jawahar for his invaluable guidance and unwavering support throughout my journey. His mentorship and ability to hold me accountable for my work have been instrumental in successfully completing my master's degree. I am truly grateful for his exemplary teaching and guidance.

My time at CVIT has been an enriching experience, and I consider IIIT-H to be the place where I learned the importance of questioning everything. Being a part of the CVIT lab has provided me with invaluable learning opportunities. I extend my heartfelt appreciation to the friends I have made along the way, including Vamshi, Deepayan, Jerin, Prachi, Riya, Deepak, Ashish, and all others who have shared their time and knowledge with me.

I would also like to express my deep gratitude to the CVIT staff and administrative team, including Shiva, Rohitha, Aradhana, Nimita, Silar, and everyone else who has made my time at the institute easier through their support and assistance.

I am especially grateful for the unwavering support of my friends throughout this journey. I would like to thank Santhoshini, Shyam, Ashutosh, Vishal, Chandu, and Abhishek, who have been my closest friends not just during my college years, but also throughout life.

Lastly, I am indebted to my family for their constant support and encouragement. I express my heartfelt appreciation to my mother, sister, and grandmother, who have always been there for me, providing unwavering support and belief in my abilities.

Abstract

Player detection is a fundamental building block for numerous applications in sports analytics, encompassing player recognition, player tracking, and activity detection. However, the majority of existing research in this domain relies on fixed-camera top-view videos of the field, which inherently simplifies the player detection task. Regrettably, such videos are not readily accessible to the general public, rendering them an unreliable data source for comprehensive player analysis. In contrast, broadcast videos of matches offer a readily available resource. Performing player detection on these videos proves considerably more challenging due to the presence of diverse sources of noise. This study investigates player detection in the context of continuous long-shot broadcast videos, acknowledging the complexities associated with this particular setting.

In the initial phase of our research, we thoroughly examine the distinctions between player detection and person detection while also investigating the multitude of challenges inherent to player detection. We begin by formulating player detection as a domain adaptation problem and analysing the various challenges associated with this approach. Our analysis encompasses an in-depth examination of the overarching challenges encountered in player detection, along with a comprehensive exploration of the unique obstacles posed specifically by broadcast videos and domain adaptation settings.

During the subsequent phase of our research, we worked on the development of an extensively annotated player detection dataset, curated from soccer broadcast videos of the FIFA 2018 World Cup matches. This dataset serves as a robust foundation for evaluating the efficacy of player detection algorithms within the context of broadcast videos. We devise a comprehensive pipeline for generating automatic labels for our dataset, which are then corrected further down the pipeline to facilitate the annotation process. The resultant dataset comprises over 200,000 high-resolution frame images, encompassing more than 2,000,000 annotated bounding boxes extracted from three distinct FIFA 2018 World Cup matches. Notably, our dataset encompasses a diverse set of player positions, orientations, and bounding box sizes, effectively capturing the inherent variability encountered in soccer broadcasts. Additionally, the dataset incorporates numerous instances of challenging noisy data points, elevating its complexity beyond previous datasets in the field.

In the third phase of our research, we present a novel transductive approach to address the player detection challenge, treating it as a domain adaptation problem. We demonstrate the significance of instance-level *domain labels* in achieving effective adaptation, specifically for soccer broadcast videos. To efficiently annotate these domain labels on the bounding box predictions generated by our inductive model, we propose a sophisticated multi-model greedy labelling scheme that leverages visual features. The annotated domain labels are then utilized to train a transductive counterpart of the model, utilizing *reliable* instances derived from the inductive model inferences. This approach proves to be highly advantageous, enabling remarkable performance enhancements for a given match with a minimal number of labelled samples. Our experimental results highlight an average increase of 16 points in mean Average Precision (mAP) for soccer broadcast videos, accomplished by annotating domain labels for approximately 100 samples per video.

In the culminating phase of our research, we demonstrate the practical utilization of robust player detection algorithms in constructing analytical systems to enhance game analysis. Specifically, we develop field heat maps that effectively depict the spatial distribution of players on the field over time. Leveraging bounding box detections derived from our proposed approach, we employ homographic projections to achieve accurate top-view registration of the detected bounding boxes in each frame. These generated heat maps serve as a valuable resource for deriving insightful inferences that directly correlate with significant events transpiring during the match. Furthermore, we present additional potential applications for leveraging reliable detection systems, while also outlining avenues for future enhancements and refinements to our system.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Player detection in broadcast videos

Analysis of sports broadcast videos is an exciting and relatively unexplored area of research in computer vision and media understanding. Tasks such as event detection, activity recognition, player tracking, and team analysis are helpful applications to understand and analyze a game. These downstream tasks require player detection as their primary basis or as supplementary information [13, 33, 17]. Existing player detection approaches utilise input from the camera feed or use a single view of the match for detection [32, 55, 16, 22], which is not readily available. Broadcast videos of the matches are a much more accessible source of unlabelled information. Although broadcast videos are an easily available source of data for analysis, player detection on these videos has major challenges associated with them as opposed to fixed camera videos.

1.1.1 Background

Contemporary player detection systems predominantly rely on prearranging specialized equipment to process live matches, often neglecting the analysis of broadcast videos [3, 49, 34]. Previous works such as [32, 55, 17] adopt non deep-learning based object detection systems, imposing limitations on the camera view of the input video.

More recent approaches, such as [16], employ a student-teacher training paradigm to train a compact network using an enhanced teacher network that compensates for missed detections through a blob detection strategy and human annotations. Notably, the reported results pertain to wide-angle fixed camera views with minimal occurrences of false positives, yielding satisfactory performance for fine-tuned detectors.

In a similar approach, [22] leverage a Feature Pyramid Network and integrate lower-level features with higher spatial and higher-level features possessing a wider receptive field. Their compact net-

work exhibits commendable performance on the ISSIA dataset [7] and the Soccer Player Detection dataset [28]. These supervised approaches necessitate annotated training data to be readily available, which is expensive. None of these works deal with broadcast videos, and most require either human-annotated labels or fixed camera views. Our work focuses on performing player detection on broadcast videos, without the expensive task of annotating bounding box data for training.

1.1.2 Challenges

Models trained on extensive object detection datasets, such as MS Coco [26], commonly incorporate classes for "person" which can be leveraged for player detection. However, the unique characteristics of broadcast videos present inherent challenges not encountered in these large-scale object detection datasets. These include, but are not limited to:

- **Motion Blur:** Motion blur emerges as a significant source of noise in broadcast videos, stemming from the high-speed movements of players and camera motion.
- **Pose Variations:** Players frequently engage in dynamic actions like diving, jumping, or falling, resulting in a wide range of pose variations. These variations are not adequately represented in large-scale object detection datasets, thereby causing models trained on such data to struggle in these scenarios.
- **Player Truncation:** Players often move in and out of the frame, either vertically or horizontally, resulting in partial visibility. Large-scale object detection models encounter difficulties when confronted with only partial views of players.
- **Player Occlusion:** Instances of players jumping over one another or overlapping within the video frame further exacerbate detection challenges, constituting a prevalent hurdle in soccer videos as a whole.

These inherent challenges in broadcast videos necessitate tailored approaches for player detection, distinct from the methodologies employed with large-scale object detection models.

1.2 Our approach

The challenges outlined in the preceding section may be addressed with a sufficient amount of training data, albeit with the expensive requirement of annotation. However, our experiments reveal that models trained on videos from a specific match tend to exhibit sub-optimal performance on other matches. This observation underscores the necessity for approaches capable of learning from match-specific instances to attain desired outcomes. Annotating training data, albeit for fine-tuning, for each match becomes impractical. Our focus in this work lies on self-supervised and weakly supervised techniques that harness knowledge from specific matches. These approaches facilitate tailored tuning for

each match, without the added cost of annotation. The resultant performance gains on training models on specific matches outweigh the associated computational costs and our focus lies in optimising such approaches.

To facilitate player detection in the context of broadcast videos, we adopt a transductive approach, leveraging reliable instances from a given match to refine the model further. By formulating player detection as a domain adaptation problem, we address the prominent issue of *domain-noise* prevalent in broadcast videos, as we will discuss in Chapter 2. Our findings emphasize the significance of instance-level *domain labels* for achieving optimal performance using a transductive approach with broadcast video data.

Furthermore, we observe a lack of publicly available datasets encompassing continuous long-shot broadcast videos. To evaluate the efficacy of our approach, we curate and release a comprehensive dataset comprising fully annotated player detections sourced from FIFA 2018 World Cup matches. This dataset not only serves as a valuable benchmark for assessing our methodology but also provides supplementary information for various downstream tasks within the realm of sports analysis.

1.3 Contribution

This thesis makes the following major contributions:

- We propose a novel transductive approach to player detection that yields significant performance enhancements, even with a limited number of domain-labeled samples.
- We introduce a methodology for collectively assigning instance-level *domain labels*, which is essential for tackling the *domain-noise* prevalent in target data, such as broadcast videos.
- We curate and release a comprehensive dataset comprising fully annotated soccer broadcast videos, facilitating the evaluation of player detection techniques.
- We showcase an application of reliable detection systems for analysis of soccer matches using automatically generated field heat maps.

1.4 Thesis Outline

The organization of this thesis is as follows:

- **Chapter 2:** We delve into the distinctions between player and person detection. Additionally, we formulate player detection as a domain adaptation problem, highlighting the significant *domain shift* encountered when transitioning from person to player detection. We underscore the necessity

of instance-level adaptation to fine-tune person detection models for optimal player detection performance.

- **Chapter 3:** We present the creation of a fully annotated player detection dataset derived from soccer broadcast videos. We outline the pipeline used for generating automatic annotations through a pre-trained detection model, subsequently refining these annotations via player tracking and manual correction, resulting in a high-quality dataset.

- **Chapter 4:** We propose a novel multi-model greedy clustering approach to collectively assign instance-level *domain labels* to an initial set of proposed bounding boxes. We leverage *reliable* instances from these proposals to update model parameters of a transductive copy of the detection model, yielding a significant boost in player detection performance. On average, annotating approximately 100 samples per video with *domain labels* results in a remarkable 16-point improvement in mAP.

- **Chapter 5:** We demonstrate a compelling application of a reliable detection system by tracking player movements throughout the match. To illustrate this, we generate field heat maps using bounding box detections from our transductive model. These heat maps serve as a valuable resource for extracting meaningful insights that directly correlate with significant match events.

Chapter 2

Player detection

In this chapter, we aim to comprehensively model the player detection problem as a domain adaptation challenge. We begin by modelling the player detection problem as a domain adaptation problem, focusing specifically on the difference between *image level* and *instance level* adaptation, emphasizing the importance of *instance-level* learning for effective domain adaptation.

We also investigate some of the challenges encountered in soccer broadcast videos, distinct from general-purpose person detection. Our focus is primarily on the impact of camera views and jersey colors on the performance of person detectors. While these factors are implicitly addressed when more data is introduced for learning, we underscore the significant *domain shift* between these two problem domains, emphasizing the need for models to adapt and acquire domain-specific knowledge.

We further conduct a series of experiments that highlight the prevalent *false positive problem* in person detectors, shedding light on the limitations of utilizing pre-trained person detectors for accurate player detection. These false positives serve as *domain-noise* inhibiting proper *instance level* adaptation during the learning process. This poses a challenge in transductive learning, particularly the presence of this *domain-noise* during the inductive phase, which hinders the proper learning of transductive models.

By establishing the theoretical framework, we provide a new perspective to player detection as a domain adaptation challenge, building upon existing models that are widely used for person detection. We also outline the specific challenges that must be addressed in order to successfully employ transductive approaches for this task.

2.1 Player detection as domain adaptation

In this section, we delve into the formulation of player detection as a domain adaptation problem. Our focus lies on differentiating between a *person* and a *player*, with the latter being defined as the 22 individuals comprising the two teams in a soccer match.

We define player detection as the task of learning the joint distribution $P(C, B, I)$, where C denotes the class label, B represents the bounding box, and I signifies the input image. As shown in [5], the joint distribution can effectively be decomposed as in Equation 2.1:

$$P(C, B, I) = P(C|B, I)\, P(B, I) \tag{2.1}$$

While we distinguish between the *person* and *player* classes, it is important to note that one is a subset of the other. Consequently, we posit that a person detector can serve as a player detector, assuming that $P(C|B, I)$ remains consistent across domains. The *domain shift* in this case arises due to the detector represented by $P(B, I)$, which can be further decomposed as:

$$P(B, I) = P(B|I)\, P(I) \tag{2.2}$$

To address the domain shift, we therefore, need a joint consideration of image-level and instance-level domain adaptation, aiming to ensure consistency between $P(I)$ and $P(B|I)$ across the source and target domains. Although sufficient image-level adaptation is relatively easy to achieve, instance-level adaptation is much more challenging due to the presence of *domain noise* as we will discuss in the upcoming sections.

2.2 Domain Shift

We begin by highlighting the considerable domain shift observed between player detection in broadcast videos and the broader domain of person detection, resulting in the need to adapt person detectors into player detectors. We conduct a comprehensive analysis focusing on two key factors: camera views and jersey colours.

2.2.1 Test dataset

To conduct this experiment, we carefully curate a test set comprising 400 images extracted from four distinct matches of FIFA 2018 (100 images per match, 25 per camera view). In our evaluation, we considered four primary camera views, namely: Top view Zoomed-in, Top view Zoomed-out, Ground view Zoomed-in, and Ground view Zoomed-out. These camera views align with the foundational framework established in prior research [46]. Samples from the test set highlighting the different camera views can be seen in Fig. 2.1. Additionally, these matches were intentionally selected to encompass a wide range of jersey colors, thereby providing a representative sample. The selection of jersey colors for our test set was informed by a thorough examination conducted in accordance with the study presented in [50], which identified the five most frequently observed colors worn by teams.

(a) Top view Zoomed-in

(b) Top view Zoomed-out

(c) Ground view Zoomed-in

(d) Ground view Zoomed-out

Figure 2.1: The four different camera views generally found in soccer broadcast videos, namely: Top Zoomed-in, Top Zoomed-out, Ground Zoomed-in and Ground Zoomed-out. Models that rely on anchor boxes, such as YOLOv3, need to be tuned using anchors that cover these views.

2.2.2 Investigation

In order to assess the impact of both jersey colour and camera views on a detector, we employ a single YOLOv3 detection model [40] that has been pre-trained on the MS COCO dataset [26]. The model is utilized to detect instances of the *person* class and mean average precision (mAP) is calculated with an intersection over union (IoU) threshold of 0.5, following standard evaluation protocols [37]. The results of this experiment are presented in Table 2.1.

Table 2.1: Detection results for different camera views and jersey colours

	Top-in	**Top-out**	**Ground-in**	**Ground-out**
Blue	0.555	0.562	0.436	0.483
Red	0.53	0.566	0.424	0.464
White	0.53	0.566	0.44	0.472
Green	0.467	0.457	0.388	0.448
Yellow	0.467	0.457	0.388	0.448

The results clearly indicate that both camera views and jersey colours significantly impact the performance of detectors. A notable finding is that players wearing green or yellow jerseys pose challenges for detection models. These colours make it difficult to distinguish the players from the background,

Table 2.2: Results using pre-trained person detectors FasterRCNN, YOLOv3 with SPP(YOLOv3-SPP), and RetinaNet

	FR vs. CR			**FR vs. BE**			**EN vs. CR**		
	P	*R*	*mAP*	*P*	*R*	*mAP*	*P*	*R*	*mAP*
FasterRCNN	0.46	0.80	0.38	0.63	0.84	0.54	0.38	0.79	0.38
YOLOv3-SPP	0.42	0.83	0.59	0.49	0.85	0.65	0.34	0.82	0.54
RetinaNet	0.50	0.79	0.41	0.62	0.83	0.52	0.40	0.77	0.40

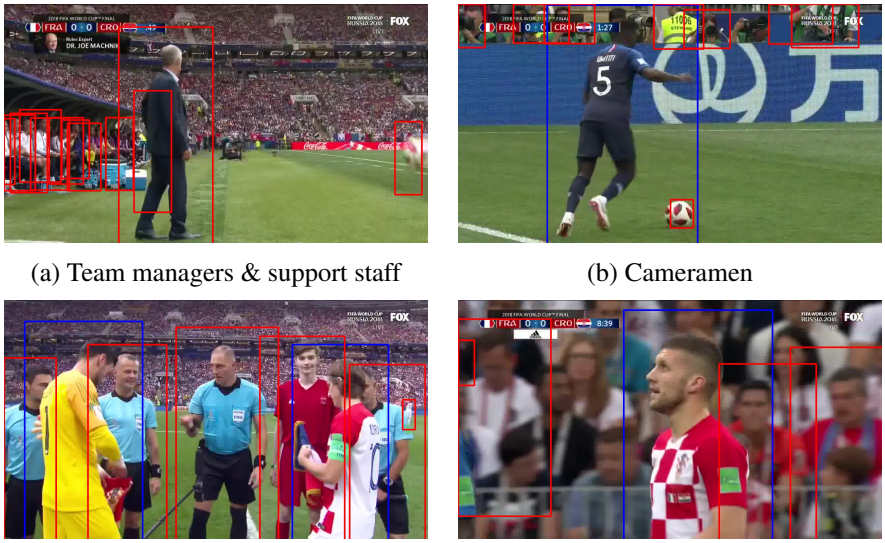
typically consisting of grass, leading to decreased detection accuracy.

Contrary to intuition, zoomed-in views are observed to be generally more challenging than zoomed-out views. Although zoomed-in views offer a clearer visual of the players, they often capture moments of high interest in a broadcast match, such as goals, fouls, or free kicks. These instances involve players in more complex positions, thereby increasing the difficulty of detection. Furthermore, ground views exhibit notably poorer performance compared to top views. Despite confident player detection in ground views, they suffer from significant occurrences of false positives. This phenomenon will be further explored in the subsequent section.

In summary, changing camera views presents a considerable challenge for detectors, making broadcast videos inherently more demanding than scenarios with fixed camera solutions. Moreover, the suitability of person detectors is compromised when dealing with colours that are difficult to differentiate from the background. This observation underscores the necessity of training models with features that remain robust in such cases. Both these issues highlight the domain shift between player detection and person detection, and we will explore ways to overcome this domain shift in further chapters.

2.3 The False Positive Problem

The *false positive problem* refers to the large number of false positives observed when using person detectors for player detection in broadcast videos. To thoroughly investigate this issue, we conduct a comprehensive evaluation using three widely recognized detectors: Faster R-CNN [41], YOLOv3 with SPP (YOLOv3-SPP) [40, 14], and RetinaNet [25]. These detectors are pretrained on large-scale datasets such as MS Coco [26]. We assess the performance of these detectors in terms of Precision (*P*), Recall (*R*), and mean average precision (*mAP*) with an IoU threshold of 0.5, adhering to standard evaluation practices [37]. For player detection, we consider the detection of the *person* class as the *player* class. The results of this experiment are presented in Table 2.2. To ensure a comprehensive evaluation, we test these models on three distinct soccer broadcast matches from our proposed dataset (details of which will be discussed in Chapter 3).



(a) Team managers & support staff (b) Cameramen

(c) Referees (d) Audience members

Figure 2.2: Domain noise present in soccer broadcast videos. Detections are correct for the *person* class; however, there are many false positives for the *player* class. This noise prevents the use of transductive or self-supervised approaches for training without being removed.

The performance of these detectors as player detectors is significantly inadequate. On average, the mean average precision (mAP) values are approximately 35% lower compared to the mAP values achieved on datasets like MS Coco [26] when detecting the *person* class. While challenging conditions in soccer broadcast videos could contribute to this disparity, qualitative analysis reveals a notable presence of false positives during the evaluation process. These false positives include referees, cameramen, audience members, support staff, and managers, rather than solely players, as depicted in Fig. 2.2.

Fine-tuning the models on a small subset of labeled data from each match might be considered to address this issue. However, in subsequent chapters, we demonstrate that such methods do not yield satisfactory performance improvements and involve expensive annotations. Instead, we propose the incorporation of player-specific *identification* features into the pipeline to enhance the detector's robustness. Additionally, we will later present a more efficient annotation pipeline that requires labeling only domain labels on a few samples per match, resulting in significant performance boosts.

In the context of domain adaptation for player detection as discussed in Section 2.1, these false positives serve as *domain noise*, preventing the proper learning of the instance level distribution $P(B|I)$. In conventional settings, image-level and instance-level adaptation are performed using samples labeled as belonging to either the source or target domain. Several domain adaptation approaches leverage the notion of "domain labels," particularly those utilizing domain classifiers [5, 35, 10]. In most cases, these domain labels are annotated at the image level and are assumed to hold true for all instances within the image. However, such an assumption does not hold in the case of soccer broadcast videos, as we have

already shown. Without removing invalid instances or *domain noise*, we observe that sufficient image-level adaptation can be achieved, while instance-level adaptation remains lacking. This holds true for any real-world dataset containing instances from multiple domains.

In Chapter 4, we will propose an efficient framework that assigns a domain label to each bounding box instance using a greedy multi-model collective labeling scheme based on identification features. Subsequently, we employ these instance-level labels to perform domain adaptation by training a transductive model exclusively on *valid* instances from the target domain data. This is achieved by eliminating the *domain noise* from the initial predictions generated by our inductive model.

Chapter 3

Dataset

In this chapter, we present an overview of existing player detection datasets in the literature and discuss their suitability for evaluating player detection methods in continuous long-shot broadcast videos. However, due to the limited availability of appropriate datasets, we have created a comprehensive player detection dataset specifically for evaluating our method. Our dataset was collected from the broadcast videos of three FIFA 2018 World Cup matches. To ensure accurate annotations, we employed a semi-automatic annotation approach that involves using a person detector in conjunction with tracking to generate initial annotations, which are then manually corrected by human annotators.

3.1 Background

The literature lacks player detection datasets specifically designed for continuous long-shot broadcast videos. Among the few notable datasets is the ISSIA soccer dataset [7], which contains 18,000 frames annotated with player and referee bounding boxes. However, this dataset only includes a single match captured from three fixed wide-angle camera views, lacking the variability of shots at different zoom levels, camera movements, and shot transitions commonly encountered in soccer broadcast videos.

Another dataset proposed in [28] consists of 2,019 annotated images recorded by three broadcast pan-tilt-zoom (PTZ) cameras. While this dataset incorporates camera movements, it remains limited to fixed views and does not provide a sufficient number of images to cover the diverse scenarios typically observed in broadcast matches.

The SoccerDB dataset [20], derived from the SoccerNet dataset [12], includes 346 clips annotated for player detection using an automated labeling scheme. Although the videos are sourced from broadcast matches, the annotations are done at the clip level, with each clip containing only a few hundred frames and lacking significant shot transitions.

Given the absence of publicly available datasets containing continuous long-shot broadcast videos, we have created and released our own dataset as part of this work to address this gap and facilitate the evaluation of player detection methods.

Table 3.1: Details and statistics of the proposed dataset.

Match	FR vs. CR	FR vs. BE	EN vs. CR
Date	15.07.2018	10.07.2018	11.07.2018
Broadcaster	Fox Sports	Fox Sports	Fox Sports
Resolution	1280×720	1280×720	1280×720
Length(Frames)	95,176	95,944	74,505
Annotated Frames	86,954	89,268	56,096
Total bounding boxes	747,876	950,802	416,818
Avg. bounding boxes	8.6	10.65	7.43
Bounding boxes automatically annotated	838,555	1,055,735	1,009,835
Bounding boxes deleted during correction (FP)	90,679	104,933	593,017

3.2 Data

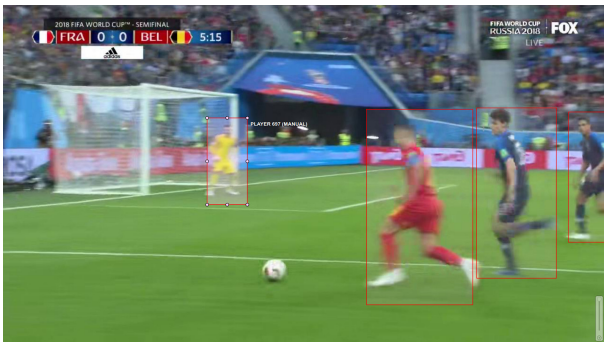
For our dataset, we have carefully selected videos from three distinct broadcast matches of the FIFA 2018 World Cup. These videos represent continuous, unedited live broadcasts of the matches as they were telecast to viewers. Each video in our dataset is annotated with bounding boxes at a per-frame level, ensuring detailed and accurate annotations throughout.

To provide a diverse set of samples, we have included videos that showcase at least four different camera views: top-zoomed-in, top-zoomed-out, bottom-zoomed-in, and bottom-zoomed-out. Moreover, our dataset captures instances of smooth transitions between these camera views, adding further complexity to the dataset. The matches in the dataset involve four different teams, namely France, Croatia, Belgium, and England. This deliberate selection ensures variations in player appearances and jersey colors, reflecting real-world scenarios.

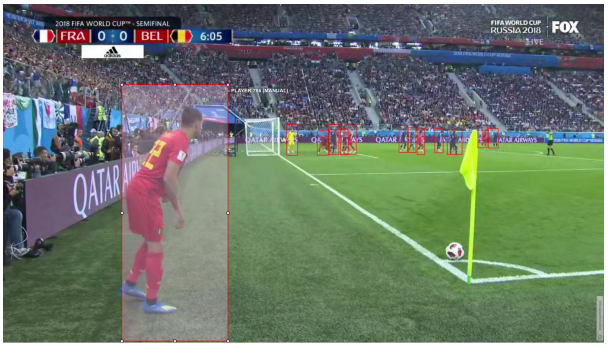
In terms of scale, our dataset comprises a remarkable 265,625 frame images, featuring an impressive 2,115,496 annotated bounding boxes. As far as we are aware, this dataset stands as the largest of its kind, providing an extensive resource for player detection research. We have ensured a wide coverage of player positions and orientations, encompassing a substantial variability in bounding box sizes. Additionally, the dataset contains instances of player occlusion and players being partially outside the frame, introducing challenges commonly encountered in player detection tasks.

Table 3.1 presents further detailed information about the dataset at the match level, providing a comprehensive overview of its characteristics. To visually showcase the challenging scenarios captured in our dataset, we present a selection of sample images in Fig. 3.1.

(a) Top camera view

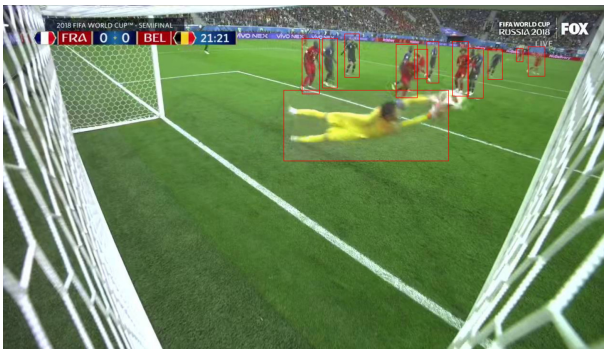


(b) Bottom moving camera view

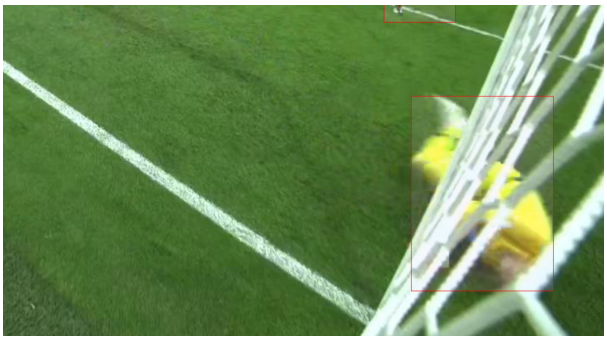


(c) Bottom view with multiple scale bounding boxes

(d) Player pose(kicking)



(e) Player pose(diving)



(f) Partial occlusion with net

Figure 3.1: Annotated sample images from the dataset, showing different views and scenarios some of which are challenging to perform detection on.

3.3 Annotation Pipeline

For the creation of this dataset we follow a semi-automatic pipeline as shown in Fig 3.2. The dataset was annotated with the help of a pre-trained YOLOv3 [40] detector along with the DeepSORT [54] tracking algorithm, followed by manual corrections by humans.

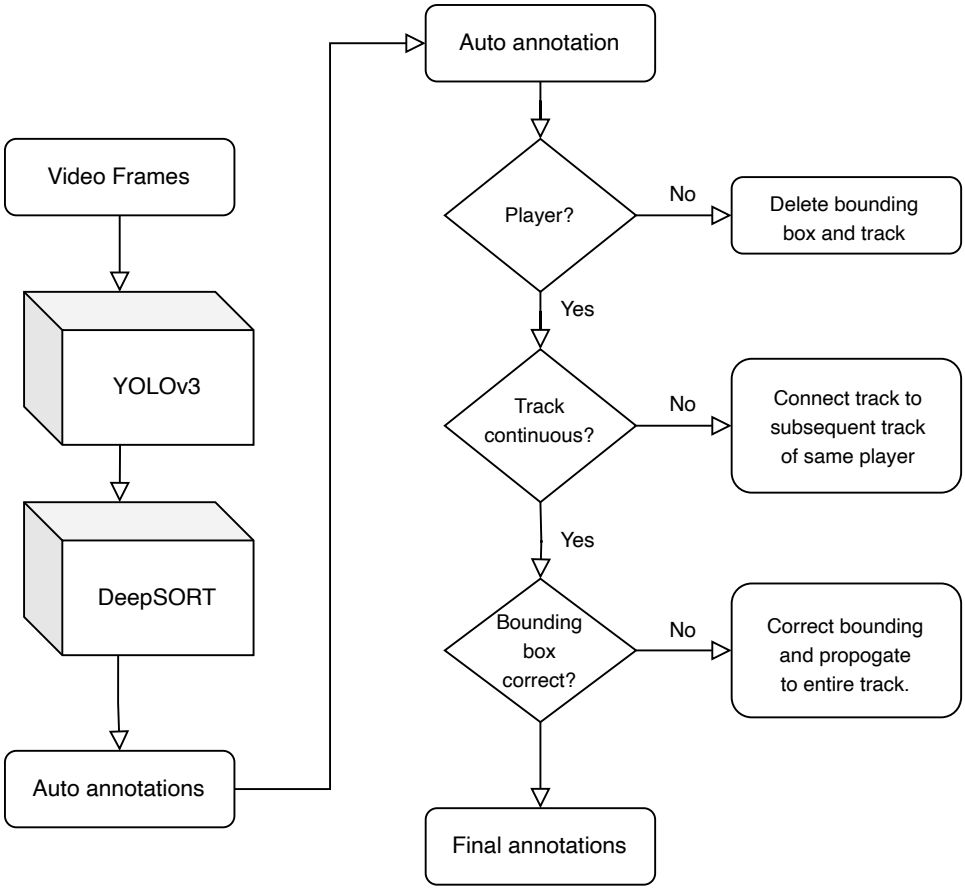


Figure 3.2: Semi-automatic annotation pipeline

3.3.1 Automatic annotations

In the initial stage of our pipeline, we employ the YOLOv3 Detection model to perform detection on each frame of our video data. Specifically, we focus on detecting the *person* class, using a class confidence threshold of 0.3 and an NMS threshold of 0.7. This approach allows us to extract a substantial number of bounding boxes while minimizing the significant overlap between them. These bounding boxes serve as the basis for our subsequent player-tracking process.

Tracking plays a crucial role in our pipeline as it provides a preliminary level of filtering for the proposed bounding boxes. It helps identify and eliminate outlier boxes that cannot be associated with any

meaningful tracks. Moreover, tracking utilizes contextual information from bounding boxes in previous frames to refine and correct detections. By maximizing the number of bounding boxes detected initially and applying filtering through tracking, we can improve the overall accuracy of our results. For tracking, we employ the DeepSORT [54] tracking algorithm. DeepSORT builds on top of SORT [2] which has demonstrated its effectiveness in various multi-object tracking tasks [47, 52, 8], including person and player tracking [31, 48].

The algorithm builds upon the widely used SORT algorithm, which employs the Hungarian algorithm [23] with Kalman filtering [21], for associating bounding boxes across neighboring frames. It introduces an additional layer of utilizing a deep learning model-based embedding to enhance the association of neighboring bounding boxes. Additionally, two novel metrics that integrate motion and appearance information are incorporated.

To incorporate motion information, a metric using the Mahalanobis distance between predicted Kalman states and newly arrived measurements is introduced. This metric is represented by the following equation:

$$d^{(1)}(i,j) = (d_j - y_i)^T S_i^{-1} (d_j - y_i)$$

Here, d_j represents the newly measured bounding box, y_i represents the Kalman state for track i, and S_i is the projection to the measurement space.

Similarly, to incorporate appearance information into the SORT algorithm, an appearance descriptor is computed for each bounding box. In our work, we utilize a person re-identification model [56] to extract identity-related information and incorporate it into these descriptors. Further details regarding the re-identification model and its training are provided in Section 4.2. The appearance information is integrated using the following metric:

$$d^{(2)}(i,j) = \min\{1 - r_j^T r_k^{(i)} \mid r_k^{(i)} \in R_i\}$$

where r_j represents the appearance descriptor of the newly detected bounding box, and R_i is the gallery of descriptors of the last 100 bounding boxes added to track i.

Both the motion metric and the appearance metric are thresholded to determine whether a bounding box should be added to a track. Tracks that do not receive any new additions for ten consecutive frames are terminated. Furthermore, to be considered a valid track, a track must exist for at least 3 consecutive frames. Two temporally consecutive bounding boxes are considered to belong to the same track if their Intersection over Union (IoU) is greater than 0.7.

3.3.2 Human correction

Following the completion of our initial detection and tracking pipeline, human annotators are employed to rectify the automatic annotations and address any missed detections. The correction methodology adopted by our annotators is as follows:

- For each detection, a thorough examination is conducted to determine whether it corresponds to a player or an outlier category, such as referees, cameramen, audience members, support staff, managers, or others. Non-player detections are promptly removed, along with their associated bounding boxes and tracks.
- Every track is meticulously reviewed to ensure its continuity. In cases where a single track has been inadvertently fragmented into multiple segments, efforts are made to reconnect these segments if they pertain to the same player within the same shot.
- The shape of each bounding box in a track is carefully adjusted in keyframes to provide a more accurate fit around the player. To enhance efficiency, interpolation techniques are employed to estimate the bounding boxes in intermediate frames, minimizing the number of required corrections.

Upon completion of the automatic annotation correction, a second round of annotations is performed for each video to incorporate any missed detections and tracks that were not captured by our initial pipeline. It is important to note that while tracking information is annotated in our dataset, we make no claim of complete accuracy in tracking, as our primary focus was to ensure precise detection. To facilitate the annotation process, we utilize the CVAT annotation tool [44], and all annotations are stored in the CVAT annotation format.

Chapter 4

Transductive Weakly Supervised player detection

The transductive setting means using the unlabelled test data to improve generalisation accuracy [9]. Unlike traditional inductive learning, which focuses on generalizing from labeled training examples to unseen data, transductive learning aims to make predictions specifically for unlabeled data points in the test set. As demonstrated in Section 2.2, there exists a notable domain shift between the *person* and *player* classes. In this study, we employ transductive learning to train a dedicated *person* detection model for player detection. However, as discussed in Section 2.3, the issue of false positives hampers the efficacy of self-supervised approaches when learning from broadcast videos.

To address this challenge, we start by identifying reliable predictions from the initial set of predictions generated by a pre-trained person detector (inductive model). We adopt a greedy multi-model clustering approach to collectively label and aggregate similar bounding boxes, thereby significantly reducing the annotation workload. The predictions identified as *reliable* from the inductive phase are utilized to update the parameters of a copy of the initial person detector (transductive model), resulting in improved predictions. An overview of our approach is depicted in Figure 4.1. In the subsequent sections, we will provide detailed explanations of each individual component and present relevant results.

4.1 Background

Research on transductive settings for object detection is a relatively nascent area of study, with the majority of previous research focusing on transductive approaches in tasks such as image recognition. Work in this domain can be observed in [9], where the authors propose a novel heterogeneous multi-view hypergraph label propagation method for zero-shot learning in the transductive embedding space. While these methods may be suitable for classification problems like image recognition, they encounter challenges in object detection tasks where *domain noise*, as demonstrated in Section 2.3, exists.

In the context of domain adaptation for object detection, one of the pioneering works is presented in [5], where the authors utilize domain classifiers and consistency regularization to enhance object detection performance on the KITTI dataset [11]. Another work by [43] introduces distillation loss

along with soft labels to address single-category detection tasks such as face and pedestrian detection. Soft labels are generated with the assistance of a tracking algorithm. However, both domain classifiers and soft labeling approaches assume image-level labels that are propagated to all instances within the image. These approaches, however, fall short in soccer broadcast videos where not all instances in an image are deemed *valid*.

To the best of our knowledge, only one study has employed a transductive approach for object detection, as demonstrated in [38]. The authors propose a zero-shot learning paradigm using fixed and dynamic pseudo-labels to train a transductive model that yields improved performance in the target domain. Additionally, they incorporate semantic information using word vectors for label generation. However, such supplementary sources of information to enhance performance may not be available for broadcast videos.

The existing research primarily revolves around domain labels existing at the image level, which are assumed to be identical at the instance level, or the incorporation of additional information specific to the target domain to facilitate the learning process. However, these assumptions do not hold true for soccer broadcasts. In our work, we aim to address this challenge by efficiently assigning domain labels at the instance level and training solely on broadcast video data to enhance detection performance.

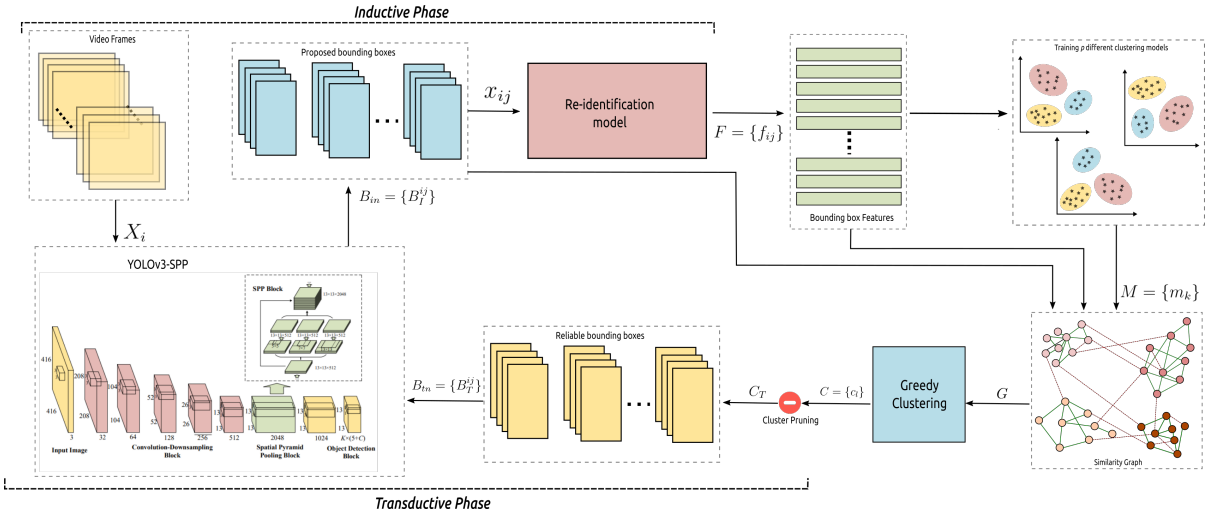


Figure 4.1: An overview of the detection pipeline proposed. The **inductive** phase includes using a pre-trained detection model to obtain initial bounding box proposals and a re-identification model to obtain visual features. The **second** stage includes clustering the obtained bounding boxes and labelling them as *reliable* or *unreliable* using a multi-model greedy clustering approach. The **transductive** phase includes using *reliable* bounding boxes to fine-tune the detector parameters to perform better detection.

4.2 Inductive phase

At the start of our pipeline lies the **inductive step**, where we employ a pretrained person detector to generate an initial set of predictions. Specifically, we utilize a YOLOv3 [40] detector with spatial pyramid pooling (SPP) [14] that has been pretrained on the MS COCO dataset [26] as the inductive model. Let B_I^{ij} denote the initial predictions obtained from the inductive model, where i represents the frame identifier and j corresponds to the jth prediction for that frame. The formulation for obtaining these initial predictions is given by Eq. 4.1.

$$B_I^{ij} = \text{NMS}(f_I(X_i)) \tag{4.1}$$

Here, f_I denotes the forward pass of the model, and NMS represents the Non-Maximal Suppression function [39] used to extract the bounding boxes from the output of the model's forward pass.

Concurrently, we extract visual features for each proposed bounding box generated by our model. These visual features play a crucial role in collectively assigning domain labels to the inductive predictions and identifying predictions that are deemed *reliable*. For this purpose, we employ a person re-identification model, to obtain visual features suitable for our instance-level domain labeling scheme. The architecture of the utilized model is depicted in Figure 4.2. It consists of a wide residual network [56] with one convolutional layer and four residual blocks.

We choose a re-identification model over a more general image recognition model due to the presence of various individuals such as audience members and support staff wearing team jerseys in the broadcast matches. The descriptors produced by standard image classification networks were not sufficiently distinct to differentiate them from the players. To obtain the visual features, we pre-train the re-identification model using the MARS dataset [57], following the methodology outlined in [54]. The visual features, which have a dimensionality of 512, are computed from the output of the average pool layer in the model.

Let x_{ij} represent the cropped region from the image corresponding to bounding box B_I^{ij}. We pass this region through the re-identification model to obtain the visual features, as depicted in Eq. 4.2, where $g(x)$ denotes the output from the average pool layer of the re-identification model.

$$f_{ij} = \frac{g(x_{ij})}{||g(x_{ij})||_2} \tag{4.2}$$

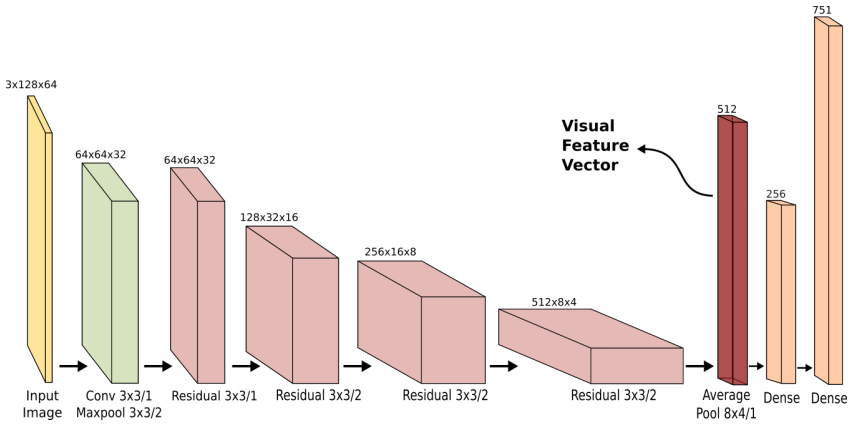


Figure 4.2: Re-identification model architecture used. The model is a wide residual network [56] consisting of 4 residual blocks.

4.3 Identifying Reliable Predictions

To efficiently assign domain labels to the bounding boxes obtained in the inductive step, we employ a clustering-based approach. We create a ***similarity graph*** among the bounding boxes obtained during the inductive phase and utilize a greedy cluster deletion technique to determine *reliable* and *unreliable* clusters. By propagating the label assigned to a representative sample within each cluster, we extend the label to all bounding boxes within that cluster.

4.3.1 Similarity Graph

To facilitate clustering, we define a similarity graph, G, based on the bounding boxes obtained in the inductive phase. In this graph, the bounding boxes serve as nodes, while the edges are generated using a multi-model similarity metric.

To construct the graph, we train p distinct unsupervised clustering models using a random subset of visual features $S = f_{ij}$ extracted from the video data. We leverage a combination of the k-means algorithm and Gaussian Mixture Models [42] with varying numbers of clusters. These models are used in establishing positive and negative edges between the nodes. The similarity metrics employed for edge creation are as follows:

- For two bounding boxes x and x', $S(x, x')$ represents the number of clustering models that assign the same clusters to both bounding boxes based on their corresponding features.
- Similarly, for two bounding boxes x and x', $\hat{S}(x, x')$ denotes the number of models that assign different clusters to the two bounding boxes based on their corresponding features.

A positive edge is established between two samples if $S(x, x') \geq t_p$, where t_p is a pre-defined threshold. Conversely, a negative edge is formed if $\hat{S}(x, x') \geq t_p$. By selecting t_p relative to p, such that the two cases are mutually exclusive, we ensure that two bounding boxes can only have a positive edge, a negative edge, or no edge between them.

4.3.2 Greedy Cluster Deletion and Labeling

We present a clustering approach to group the bounding boxes based on the similarity graph introduced in the previous section. Our method draws inspiration from the Lambda Correlation Clustering (LambdaCC) algorithm proposed in [51], where the problem of *cluster deletion* involves identifying the minimum number of edges in a graph that need to be removed to convert it into a disjoint set of cliques (clusters). Our approach offers a greedy approximation to optimize the LambdaCC [51] objective function, defined as follows:

$$\min \sum_{(i,j)\in E^+} (1-\lambda)x_{ij} + \sum_{(i,j)\in E^-} \lambda(1-x_{ij}) \tag{4.3}$$

Here, E^+ and E^- represent the positive and negative edges in the similarity graph, respectively, and x_{ij} denotes the binary distances for the edges. A high value of lambda applies a significant penalty if negative edges exist within a cluster, thereby ensuring that the clusters are internally dense and externally sparse.

We define the sets $F = f_{ij}$, $B = B_I^{ij}$, E^+, and E^- as the visual features, bounding boxes, positive edges, and negative edges in the similarity graph, respectively. Our greedy approximation of LambdaCC [51] for *cluster deletion* is outlined in Algorithm 1. The algorithm starts by randomly selecting a bounding box from the inductive set and expanding a cluster around it by including bounding boxes with positive edges connected to the sampled box.

We ensure that no negative edges exist within the cluster, thereby fostering homogeneity and internal density. In the best-case scenario, where we set $t_p = p$, the algorithm exhibits linear time complexity since negative edges are disregarded. However, this trade-off results in a larger number of bounding boxes ending up in single-element clusters, leaving fewer *reliable* samples for learning in the transductive phase. We also introduce a threshold on the cluster size, denoted as t_s, to eliminate small clusters primarily containing outliers, which may be inherently *unreliable* predictions.

Algorithm 1: Greedy approximation of cluster deletion approach

Output: A set of clusters $C = \{c_l\}$ containing visually similar bounding boxes
Input: $B = \{B_I^{ij}\}$, $F = \{f_{ij}\}$, E^+, E^-, t_s
$C \leftarrow \{\}$ (Initialise an empty set C);
while $|B| \neq 0$ **do**
    Randomly sample bounding box b and the corresponding feature f from B and F;
    $c_l \leftarrow \{b\}$ (Initialise c_l with sample b);
    for $x \in B - \{b\}$ **do**
        if $(x, b) \in E^+$ *and* $\forall x_j \in c_l | (b, x_j) \notin E^-$ **then**
            $B \leftarrow B - \{x\}$ (Remove x from B);
            $c_l \leftarrow c_l \cup \{x\}$ (Add x to c_l);
        end
    end
    if $|c_l| \geq t_s$ **then**
        $C \leftarrow C \cup c_l$ (Add c_l to C);
    end
end
return *Set of clustered bounding boxes* $C = \{c_l\}$

4.3.2.1 Representative Sample

We select a representative sample from each cluster using the optimization approach described in Eq. 4.4:

$$r_i = \min_{x \in c_i} \left(\sum_{x_j \in c_i} \frac{x_j - x}{|c_i|} \right)^2 \tag{4.4}$$

In this equation, r_i represents the representative sample for cluster c_i, and x_j denotes the visual feature used for clustering the bounding boxes within c_i. Subsequently, we label the obtained representative bounding box corresponding to the visual feature r_i as either *reliable* or *unreliable*, propagating the label across the entire cluster. We collect all the bounding boxes from the set of *reliable* clusters, C_T, and apply a threshold based on their prediction confidence, denoted as t_c. Bounding boxes meeting the threshold are added to the set B_{tn}, which will serve as the training data in the transductive phase.

4.4 Transductive Phase

In the transductive step, we create a duplicate of our inductive model while keeping the backbone, convolution-downsampling block, and SPP block frozen. We then fine-tune the remaining layers using the *reliable* samples identified in the previous stage, B_{tn}. Our experiments have shown that re-training both the detection layers and the convolution upsampling layers of YOLOv3 with SPP leads to improved performance compared to training only the detection layers. This approach facilitates image-level adaptation of the model, as discussed in Section 2.1, while re-training the detection layers contributes to instance-level adaptation. The backbone of the model has been pre-trained on the ImageNet dataset [6]. Further sections will provide detailed information on the training procedure of the transductive model and its evaluation for player detection.

4.5 Implementation Details

In this section, we provide the implementation and training details of our approach, as well as the different hyperparameters we use and their effects.

4.5.1 Identifying Reliable Predictions

To construct the similarity graph mentioned in our approach, we use four clustering models, including two K-Means models and two Gaussian Mixture Models [42], with cluster sizes of 10 and 20. We observe that the generated similarity graph does not vary significantly with the specific choice of clustering models, as long as an adequate number of models and sufficiently large cluster sizes are used.

For the *cluster deletion* phase, we set the threshold parameters as $t_p = 4$ and $t_c = 0.8$. Setting $t_p = p$ allows the cluster deletion algorithm to run in linear time, which is advantageous considering the large number of bounding boxes proposed by our initial inductive model (approximately 500,000). However, using this setting results in the creation of numerous outlier clusters. To address this, we apply a threshold on the cluster size to remove small clusters that mainly contain outliers. Despite pruning, we did not encounter a lack of training data for the transductive model due to the abundance of proposed bounding boxes from our inductive set. We label representative samples with domain labels and retain only the *reliable* predictions for the transductive stage.

4.5.2 Training Transductive Model

The training of the transductive model involves resizing the input images to a size of 416x416 with padding to maintain the aspect ratio of the bounding boxes. We use batch sizes of 32, as we have observed that larger batch sizes can lead to poor generalization and rapid overfitting of the model.

The model is trained using a combination of multi-part mean square error (MSE) and cross-entropy loss, as described in [39], represented by Equation 4.5.

$$L = L_c(y) + \lambda_n L_n(C) + L_o(C) + \lambda_l L_l(S) \tag{4.5}$$

In Equation 4.5, L_c represents the classification loss for the class label y, L_n represents the object confidence loss when no object is present in the bounding box (scaled by a factor of λ_n), L_o represents the object confidence loss when an object is present in the bounding box, and L_l represents the localization loss on the bounding box coordinates S (scaled by a factor of λ_l).

To address the imbalance between boxes containing objects and those without objects, we set $\lambda_n = 0.5$. Additionally, we have found that setting $\lambda_l = 8$ helps improve localization, resulting in more accurate bounding box predictions.

The model is trained for 200 epochs, starting with an initial learning rate of 0.0001, which is then decayed by a factor of 0.1 at the 100th and 150th epoch. Training is performed on a per-video basis, and the trained models are subsequently tested in the evaluation phase. While our method requires training models for each video, the performance improvements obtained are significant compared to models trained on multiple different videos, such as SoccerDB [20]. These improvements are achieved solely through the use of domain labels, without introducing any new bounding box information to the model apart from the initial predictions made by the inductive model.

Table 4.1: Comparative results on our proposed dataset using: pre-trained general purpose detectors; supervised approaches for player detection from SoccerDB [20] and FootAndBall [22]; self supervised approach mentioned in [43] without domain noise removal. Bottom row represent number of samples annotated in the labelling stage.

	FR vs. CR			**FR vs. BE**			**EN vs. CR**		
	P	*R*	*mAP*	*P*	*R*	*mAP*	*P*	*R*	*mAP*
FasterRCNN [41]	0.46	0.80	0.38	0.63	0.84	0.54	0.38	0.79	0.38
YOLOv3-SPP [40, 14]	0.42	0.83	0.59	0.49	0.85	0.65	0.34	0.82	0.54
RetinaNet [25]	0.50	0.79	0.41	0.62	0.83	0.52	0.40	0.77	0.40
SoccerDB [20]	0.59	0.74	0.45	0.63	0.75	0.48	0.49	0.73	0.40
FootandBall [22]	0.75	0.62	0.47	0.74	0.67	0.50	0.66	0.53	0.38
Self-Supervised [43]	0.31	**0.87**	0.35	0.40	**0.88**	0.51	0.24	**0.86**	0.29
Ours	0.76	0.85	**0.79**	0.89	0.79	**0.76**	0.77	0.78	**0.72**
(annotated samples)	105 samples			55 samples			64 samples		

4.6 Results

Our detection results are compared with several established approaches for player detection:

- Pre-trained supervised person detectors, primarily trained on large-scale object detection datasets such as MS COCO [26].
- Fine-tuned supervised approaches specifically trained on soccer data.
- Self-supervised approaches commonly used for domain adaptation.

We employ precision, recall, and mean Average Precision (mAP) with Intersection over Union (IoU) of 0.5 as the evaluation metrics, following the standard practice [37]. Additionally, we assess the performance of our approach in removing unreliable predictions compared to other simpler methods. Furthermore, we conduct additional experiments to fine-tune a detector in a supervised manner to gauge the effectiveness of our weakly-supervised approach in comparison to supervised fine-tuning.

4.6.1 Baselines

For the evaluation of pre-trained supervised person detectors, we employ three widely recognized detectors: FasterRCNN [41], YOLOv3 [40] with SPP [14] (YOLOv3-SPP), and RetinaNet [25]. These detectors are utilized as baselines, focusing on the *person* class, to highlight the challenges posed by *false positives* in player detection. This emphasizes the existence of a *domain shift* between the two objectives, as previously demonstrated, and underscores the need for methods to mitigate this shift. The pre-trained networks used in this evaluation are trained on the MS COCO dataset [26]. Detailed results can be found in Table 4.1.

4.6.2 Transductive model

We evaluate our transductive model, in a similar manner as discussed previously on our proposed dataset. Our model demonstrates excellent detection performance, with only around 100 samples per video annotated with *domain labels*. To provide a comprehensive evaluation, we directly compare our results with two recent works: SoccerDB [20] and FootAndBall [22], both of which utilize supervised models trained on labeled datasets for player detection. The comparative results can be found in Table 4.1. The models trained in both these approaches are used to perform player detection on our proposed dataset.

Notably, our method surpasses the performance of models trained in a supervised setting by a considerable margin. This improvement can be attributed to our model being specifically trained on the target data of the individual videos, allowing for more accurate and tailored detection capabilities.

4.6.3 Comparison with self-supervised approaches

Furthermore, we conduct a comparison with a recent self-supervised approach that lacks the utilization of instance-level *domain labels* during the fine-tuning of model parameters, as depicted in Table 4.1. For this purpose, we train a YOLOv3-SPP detector using the self-training approach outlined in [43]. In this approach, we employ DeepSORT [54] tracking to generate refined bounding boxes and soft labels, utilizing distillation loss for training, as described in [43].

Once again, we observe that our transductive model surpasses the performance of self-supervised approaches that do not perform any *domain-noise* removal. Since no instance-level *domain labels* were available to isolate the bounding boxes specific to the target domain, the model's performance does not exhibit improvement compared to the baseline. Instead, the model attempts to learn the *domain-noise* present in the data. This is evident from the increased recall of the trained model, accompanied by a decrease in precision, indicating that the model generates more predictions overall, including a higher number of false positives.

4.6.4 Qualitative results

We present compelling qualitative results in Figure 4.3, showcasing a comparison between the results generated by SoccerDB [20] (left) and our model (right).

In the first row, we demonstrate the ability of our model to address the *false positive problem* by avoiding the detection of staff members wearing team jerseys. This scenario poses a significant challenge since team staff often wear jerseys similar to players, making it difficult to distinguish between them.

Figure 4.3: Qualitative comparison of detection results. *Left*: SoccerDB [20]. *Right*: Our proposed method.

In the second row, we observe that our model correctly identifies the absence of players in frames where no detections are made. This highlights the capability of the transductive model to effectively differentiate between the *person* class and *players*, demonstrating its focused learning approach.

In the third row, we demonstrate the model's ability to recognize players during the initial line-up, despite the absence of these detections in the initial predictions made by the inductive model. This indicates that the transductive model has learned distinct features unique to each player, resulting in improved predictions.

These qualitative results emphasize the effectiveness of our transductive model in addressing specific challenges related to player detection, including the mitigation of false positives, accurate identification of players, and improved detection performance compared to existing approaches.

4.6.5 Comparison with supervised fine-tuning

We conduct a comprehensive evaluation of our detection performance compared to supervised fine-tuning approaches. To assess the effectiveness of our method, we adopt a random sampling strategy, where we select a percentage of ground-truth annotations and utilize them to fine-tune a YOLOv3-SPP model. The detection and upsampling convolution layers are trained with this sampled data, and the model is subsequently tested on the remaining data. The evaluation results are presented in Table 4.2.

Our results demonstrate that our approach achieves comparable performance to models trained using ground-truth data, despite incorporating only instance-level *domain labels* on approximately 100 samples. Although our method does not surpass the mAP scores of approaches utilizing annotated bounding box labels, we observe that our performance remains on par with such methods. It is worth noting that our approach consistently exhibits higher precision, indicating the enhanced accuracy of the generated bounding box predictions.

These findings highlight the effectiveness of our weakly supervised approach, as it yields competitive detection performance while leveraging limited supervision in the form of instance-level *domain labels*. Moreover, the consistently higher precision indicates the improved accuracy of our bounding box predictions compared to the fine-tuning approaches using ground-truth data.

Table 4.2: Supervised fine-tuning results on YOLOv3[40] with spatial pyramid pooling[14]. The first column shows the percentage of labelled data used for training.

Training Data	FR vs. CR			FR vs. BE			EN vs. CR		
	P	R	mAP	P	R	mAP	P	R	mAP
10% frames with labels	0.69	0.91	0.85	0.77	0.94	0.89	0.65	0.91	0.83
15% frames with labels	0.71	0.91	0.86	0.79	0.94	0.90	0.67	0.92	0.84
20% frames with labels	0.73	0.91	0.86	0.81	0.94	0.90	0.69	0.92	0.85
25% frames with labels	0.74	0.91	0.86	0.82	0.94	0.91	0.70	0.92	0.85
30% frames with labels	0.74	0.91	0.87	0.82	0.94	0.91	0.71	0.92	0.86
Domain labelled frames	**0.76**	0.85	0.79	**0.89**	0.79	0.76	**0.77**	0.78	0.72

4.6.6 Clustering baselines

Table 4.3: Comparative results of K-Means, Gaussian Mixture Model(GMM) and our multi-model greedy(MMG) clustering for identifying *reliable* predictions

Video	K-Means		GMM		MMG	
	TPR	FPR	TPR	FPR	TPR	FPR
FR vs. CR	0.97	0.36	0.97	0.45	0.56	**0.82**
FR vs. BE	0.99	0.36	0.96	0.44	0.75	**0.72**
EN vs. CR	0.78	0.53	0.79	0.66	0.74	**0.74**

To evaluate the effectiveness of our clustering approach, we train models using different clustering algorithms and assess their performance in identifying reliable predictions during the cluster pruning stage. For this purpose, we utilize the visual features obtained from the re-identification model, denoted earlier as f_{ij}. We train clustering models using K-Means and Gaussian Mixture Models [42] with varying numbers of clusters, namely 20, 40, and 80. Further identifying reliable predictions using the approach discussed in Section 4.3. The performance of the best model for each clustering method is compared to our proposed multi-model greedy (MMG) clustering.

Since our approach determines the number of clusters based on the similarity graph, we evaluate different numbers of clusters for K-Means and GMM, reporting the best performance for each. As mentioned earlier, we set $t_p = p$ in the similarity graph to optimize efficiency. We found that setting $t_p < p$ resulted in larger clusters, with multiple samples competing as candidates for the representative sample of the cluster, sometimes even having different domain labels.

To evaluate the clustering results, we introduce two metrics:

- False Positive Removal Ratio (FPR): The ratio of false positives removed to the total false positives after pruning using a given clustering method.

- True Positive Retention Ratio (TPR): The ratio of true positives retained to the total true positives after pruning.

Both metrics aim to assess the ability to successfully identify reliable predictions. The higher the values of FPR and TPR, the more desirable the clustering method is in effectively removing false positives while retaining a significant number of true positives for training.

In Table 4.3, we compare the results of K-Means, Gaussian Mixture Models (GMM), and our proposed multi-model greedy (MMG) clustering. We observe that MMG consistently achieves a higher rate of false positive removal (i.e., *domain noise*) across all videos, while still retaining a substantial number of true positives that are valuable for training purposes.

Additionally, we conducted experiments using simpler features from image classification networks such as ResNet-18 [15]. However, we found that these features posed challenges in distinguishing between players and audience members or staff wearing team jerseys. The generated visual features were too similar, making it difficult to differentiate true positives from false positives. Furthermore, these features yielded lower values for both TPR and FPR in the task of identifying reliable samples compared to the features obtained from the re-identification model.

4.6.7 Improving image-level adaptation

Table 4.4: Training only the detection layers(YOLO) versus training both the convolution upsampling and detection layers(YOLO+Up) to achieve better image level adaptation.

Method	FR vs. CR			FR vs. BE			EN vs. CR		
	P	R	mAP	P	R	mAP	P	R	mAP
YOLO	0.64	0.79	0.67	0.69	0.80	0.69	0.66	0.72	0.60
YOLO+Up	0.76	0.85	**0.79**	0.89	0.79	**0.76**	0.77	0.78	**0.72**

To investigate the impact of training different layers in the YOLOv3-SPP model, we conducted experiments to evaluate image-level adaptation. Typically, when fine-tuning YOLOv3, only the final detection layers are retrained using the target domain data. However, we observed that jointly training both the detection and upsampling convolution layers of our transductive model resulted in significantly improved feature maps at the detection layer. This, in turn, increased the number of detections made by the model during the Non-Maximal Suppression (NMS) stage [40] and ultimately enhanced the overall detector performance.

Comparisons between different training configurations are presented in Table 4.4, highlighting the performance differences achieved by training specific layers in the YOLOv3-SPP model.

Chapter 5

Game Analytics

This chapter delves into a compelling application of precise player detection systems: the generation of field heat maps for analyzing spatial player distribution during a match. These heat maps provide valuable insights into various aspects of the game, including player positioning and the effectiveness of team offense and defense strategies. Prior to the advent of automated detection systems, performing such analyses manually was a labor-intensive task. Leveraging player detection systems streamlines the process, enabling efficient post-match reviews and the enhancement of team performance. While player detection systems have a multitude of applications, this chapter specifically concentrates on this particular application, showcasing its implementation across three distinct matches from our proposed dataset.

5.1 Background

Many downstream game analysis tasks such as detection, activity recognition, player tracking, and team analysis rely on player detection as the foundational component to better understand and analyze a game.

In [27], the authors propose an imitation learning method using recurrent neural networks to learn individual player behaviors and perform rollouts of player movements on previously unseen play sequences. As the foundation of their models, they incorporate temporal player detection and tracking information. The data used in their study consists of manually annotated player positions and speed. However, annotating such data for new players and matches can be costly. In this context, reliable detection systems can be employed to model such information for new players.

In [29], the authors develop a method to accurately estimate the likelihood of a shot being taken during a game. Their analysis focuses on the spatiotemporal patterns within a ten-second window leading up to a shot, involving nearly 10,000 shots. Their findings highlight the importance of strategic features such as defender proximity, player interactions, speed of play, and shot location in determining the

likelihood of a shot and the team's goal-scoring potential. The authors utilize proprietary tracking and detection information from [36] to conduct their analysis. However, relying solely on such proprietary data sources may not always be feasible. Fortunately, reliable player detection systems can automatically provide the majority of these features, offering an alternative and accessible solution.

In [1], the authors explore the use of the Qualitative Trajectory Calculus [45] (QTC), a spatiotemporal qualitative calculus that describes the relative movement between objects, for spatial movement pattern recognition of players in soccer. Their approach incorporates features such as player speed, distance covered, and position to identify meaningful patterns in the game. Once again, these features can be obtained automatically using reliable player detection systems.

While we have discussed only a few examples of such applications, it is important to note that numerous studies in the literature leverage player positions, speed, relative movement, and distance covered for a wide range of analyses. This highlights the crucial need for reliable and accurate player detection systems that can provide this information, particularly in cases where prior player tracking has not been conducted or is unavailable.

5.2 Field heat maps

An intriguing application of a reliable player detection system is the tracking of player positions throughout a match. While individual instances of player location may not provide substantial insights into overall statistics, a valuable way to visualize player movements during the game is by generating field heat maps depicting the distribution of player positions [30]. In this study, we focus on the first 20 minutes of three different matches from the FIFA 2018 dataset to demonstrate this approach.

To generate the heat maps, we utilize our transductive models to detect player-bounding boxes by sampling frames from the video. For each frame, we apply top-view registration techniques, similar to the method described in [19], to establish a homography matrix. This matrix is then used to map the player bounding box positions onto a field map (Fig 5.1). By aggregating this information temporally throughout the video, we create comprehensive heat maps.

While our expertise may not encompass a profound analysis of the intricacies of a soccer match, we are able to establish correlations between the features present in these heat maps and specific events that occurred during the match. It is important to note that in-depth analysis of the match itself is beyond the scope of our work, and we defer such detailed examination to domain experts with a deeper understanding of the game.

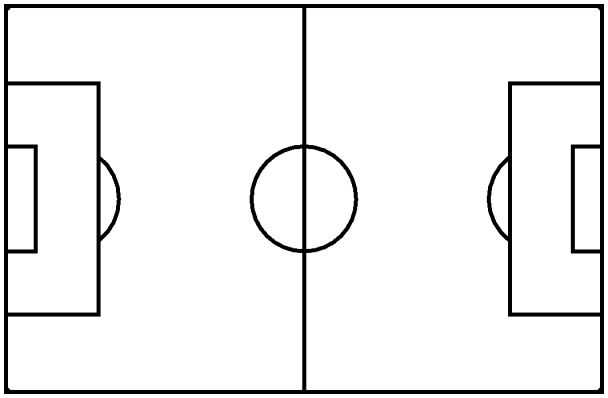


Figure 5.1: Field map template used for homography estimation

5.2.1 Player Detection

We begin by employing the transductive models for each match to perform player detection. Frames are sampled at every second for the video, at each frame we perform player detection and save both the frame image and the detected bounding boxes. We set the IoU threshold at 0.5, the class confidence threshold at 0.1, and the NMS [39] threshold at 0.5 respectively for detection.

5.2.2 Homography Estimation

For each sampled frame, homography estimation is performed to warp the frame image onto a field map template. This is accomplished through a two-stage pipeline. In the first stage, a fine-tuned ResNet-18 [15] architecture modified to predict the 8 homography parameters, as shown in Eq. 5.1, is employed.

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & 1 \end{bmatrix} \tag{5.1}$$

The initial estimation network F_θ provides an initial estimate of the homography using the input image, denoted as I, as follows:

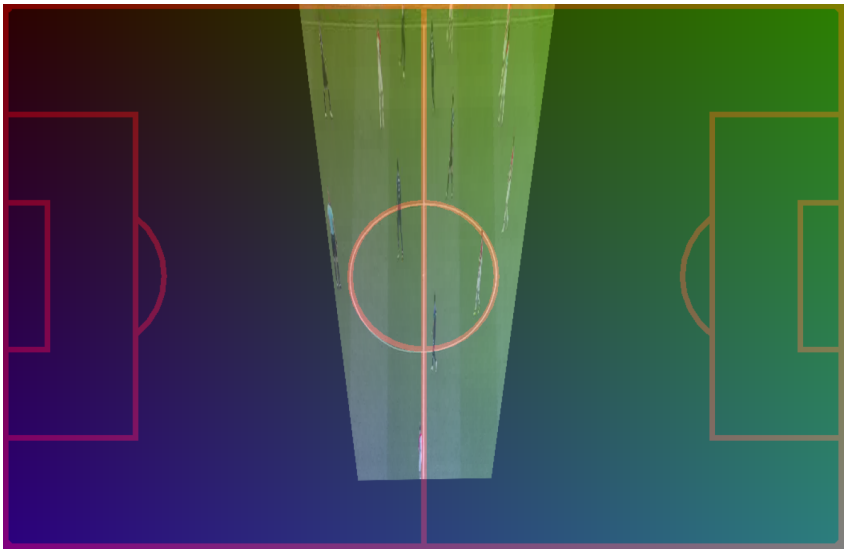
$$H^{(0)} = F_\theta(I) \tag{5.2}$$

The field template T is then warped [18] using the estimated homography $H^{(0)}$ and concatenated with the input image, resulting in a new image denoted as $\hat{I}$. This augmented image is then fed into a second network G_ϕ, which is also a modified ResNet-18 [15] architecture, to estimate the error in the warping.

$$E^{(i)} = G_\phi(\hat{I}) \tag{5.3}$$

(a) Video frame from FR vs. CR match

(b) Warped field template on frame using final homography matrix generated



(c) Frame warped onto field template using inverse of final homography matrix

Figure 5.2: Various stages of mapping player bounding boxes onto a field template

The predicted error $E^{(i)}$ is obtained from $G_\phi(\hat{I})$. The gradient of the error with respect to $H^{(0)}$ is calculated, and the homography is optimized using this gradient and the Adam optimization algorithm to obtain a new estimate $H^{(1)}$. This process of estimating the warping error and recomputing the homography is iteratively repeated m times to obtain the final homography estimate H.

Pre-trained models for both the initial homography estimation and the error prediction model are utilized from [19]. An example of the warped image obtained from the final estimated homography for one of the frames is shown in Figure 5.2.

5.2.3 Heat map generation

Once the final homography is estimated, the inverse homography H^{-1} is employed to map the original image coordinates to the field map template image, as depicted in Figure 5.2. Subsequently, the bounding box detections for frame I are also mapped onto the template field map using H^{-1}.

A spatial histogram is generated over the field map by iterating through all the sampled images from the video and collecting the spatial coordinates where a player bounding box is warped onto the field map. This spatial histogram is then converted into a heat map by applying window-based smoothing over the histogram, achieved through convolutions with both a mean kernel and a Gaussian kernel.

5.3 Analysis

The generated heat maps for three different matches are depicted in Figures 5.3, 5.4, and 5.5. Several noteworthy correlations can be observed between these heat maps and the various events that occur during the respective match segments.

In the heat map for the France vs. Croatia match (Figure 5.3), brighter regions are evident towards the left goal post (bottom of the image), which corresponds to the Croatia side. Interestingly, Croatia conceded a goal in the 18th minute. Throughout the analyzed segment, a higher concentration of players, including both defenders and attackers from the opposing team, can be observed on the Croatia side.

The heat map for the France vs. Belgium match (Figure 5.4) exhibits a relatively uniform distribution of players across the field. This distribution aligns with the match dynamics, as both teams had a comparable possession of the ball, and no goals were scored during the analyzed segment. Notably, due to the absence of noteworthy events, the camera view remained top-zoomed-out, resulting in brighter regions, as more players appeared in each frame, and a more dispersed player distribution across a larger area.

In the England vs. Croatia match heat map (Figure 5.5), a small region of heightened activity is visible on the Croatia side (bottom of the image), where England scored a goal at the 5th minute. The subsequent segment remained relatively uneventful, with players engaged in midfield ball possession battles. This is reflected by the uniform yet sparse player distribution in the central region of the heat map.

These match-specific heat maps, generated using reliable detection models, serve as valuable tools for game analysis. Heat maps represent just one example of the analytical capabilities afforded by dependable player detection systems. As previously mentioned, our analysis focuses solely on correlating these heat maps with match events, while detailed investigations fall beyond the scope of our work.

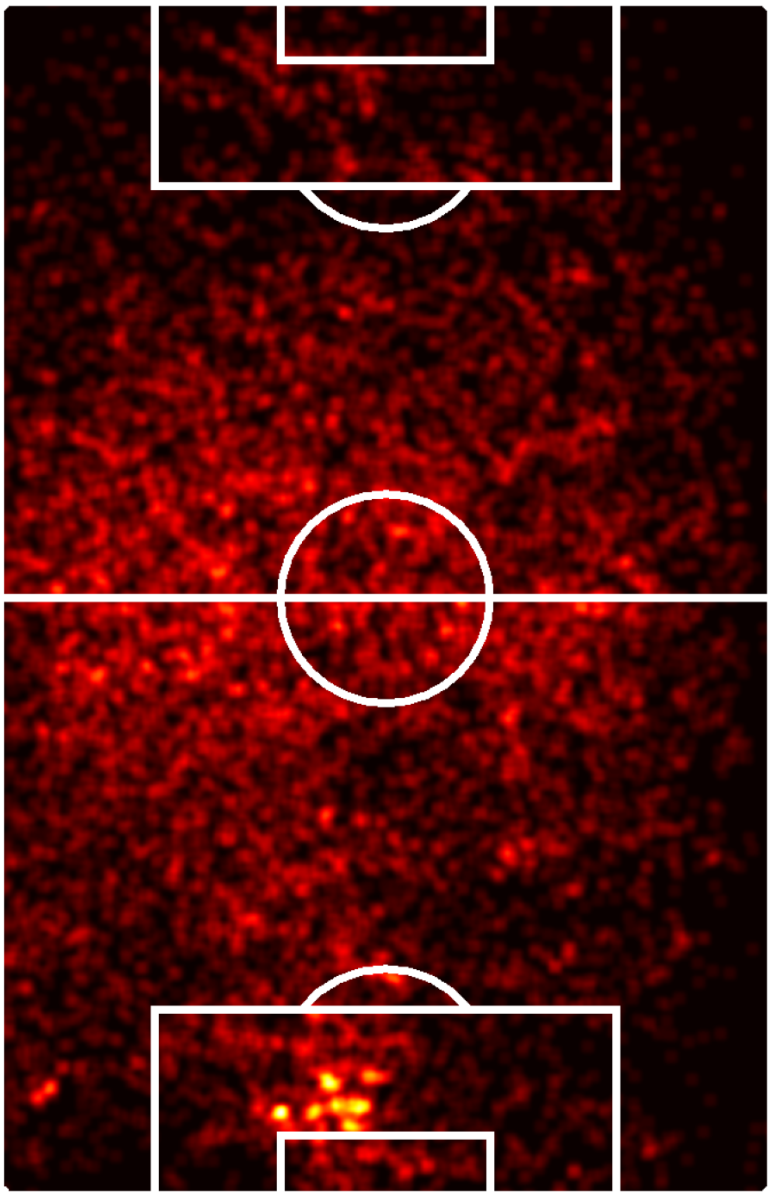


Figure 5.3: Heat-map of player locations for first 20 minutes of France vs. Croatia. Brighter regions contain more players.

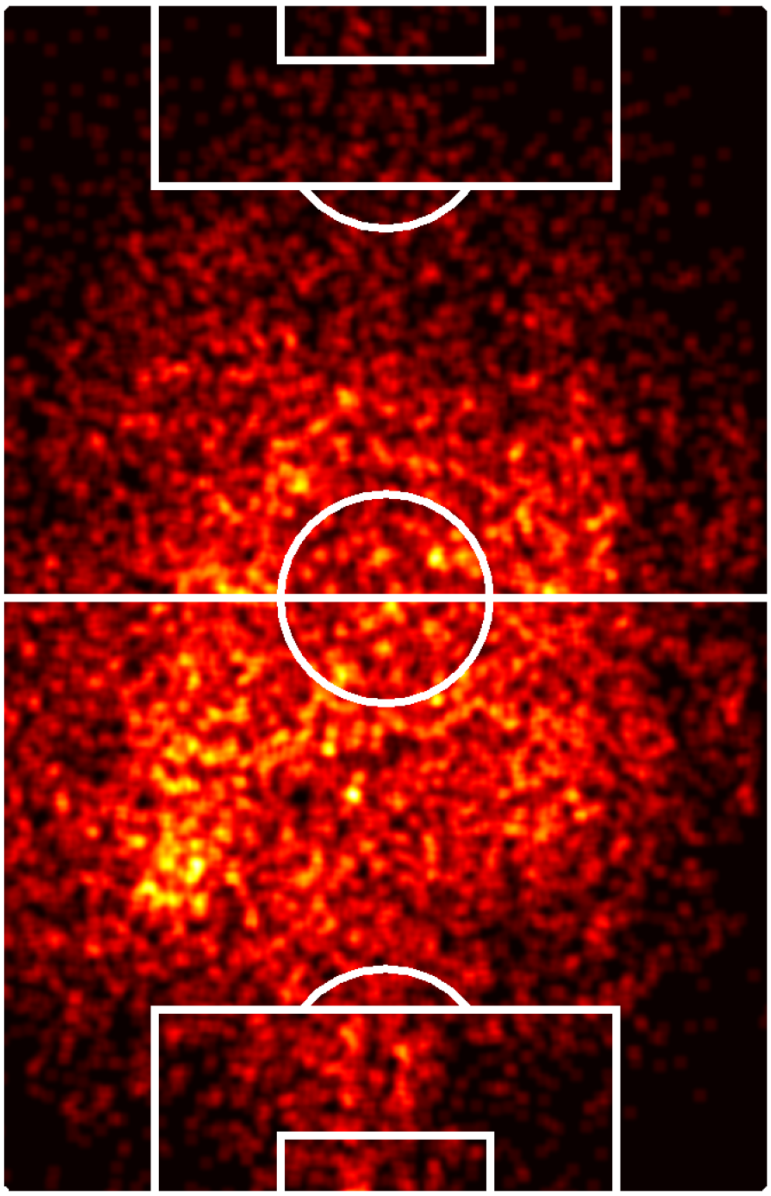


Figure 5.4: Heat-map of player locations for first 20 minutes of France vs. Belgium. Brighter regions contain more players.

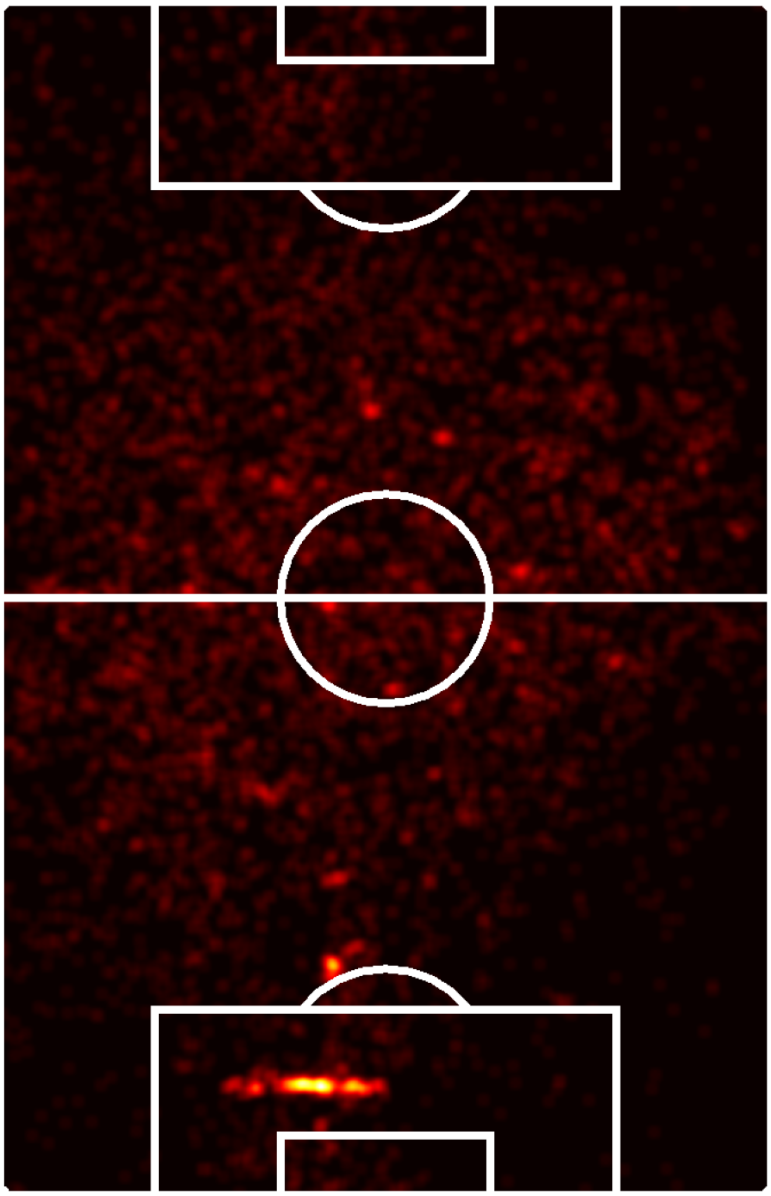


Figure 5.5: Heat-map of player locations for first 20 minutes of England vs. Croatia. Brighter regions contain more players.

Chapter 6

Conclusions and Future Directions

6.1 Summary

Our work focuses on the analysis of player detection in unconstrained soccer broadcast videos using a transductive approach that eliminates the need for annotated ground-truth data.

We frame player detection as a domain adaptation problem, highlighting the significant *domain shift* between person and player detection tasks, as well as the prevalent *domain noise* issues in soccer broadcast videos. These challenges hinder the performance of unsupervised and self-supervised learning methods for player detection.

To address these challenges, we introduce a novel player detection dataset created through a semi-automatic annotation pipeline. This dataset comprises videos from three different soccer broadcast matches from the FIFA 2018 World Cup, with over 2 million annotated bounding boxes in more than 200,000 images. It stands as the largest dataset of its kind in the existing literature.

We propose a novel transductive pipeline for learning player detection from soccer broadcast videos. By annotating domain labels for only a few samples per video, we achieve significant performance improvements. Our approach includes a unique clustering method for collective instance-level annotation of *domain labels*, effectively mitigating *domain noise* and the *false positive problem*. Our model outperforms other supervised and self-supervised methods in player detection, even with a limited number of samples annotated with domain labels.

Using our trained transductive model, we demonstrate a practical application by generating heat maps that track player positions across the field, allowing for insightful match analysis. The combination of accurate detection models with additional match-related information enables various diverse applications.

6.2 Future directions

It would be intriguing to assess the performance of modern detection architectures, such as [4] and [58], within our transductive pipeline. Our proposed approach is adaptable to any detection model, and leveraging the advancements of modern detection architectures could significantly enhance the performance of our method.

While we currently utilize only visual information from broadcast videos, audio commentary presents another valuable source of information for game analysis. Exploring how a reliable detection pipeline could incorporate audio or other supplementary information to enhance analysis would be of great interest.

Integrating a person identification model, such as [53], with our detection pipeline would enable player-specific analysis. This would allow us to generate heat maps specifically for individual players, providing insights into their strengths and weaknesses.

Event detection and action recognition are also important applications in broadcast video analysis. By combining ball detection methods, such as [24], with reliable detection pipelines, we can identify patterns that facilitate the detection of significant events like fouls, free kicks, and goals.

Applying the same pipeline to other sports broadcast videos, where domain noise is prevalent, such as basketball, would be of great interest. The *false positive problem* is even more pronounced in basketball, and we believe that our proposed approach would greatly benefit detection systems in such sports.

Related Publications

Conference:

1. **Chris Andrew Gadde**, C. V. Jawahar. **Transductive Weakly-Supervised Player Detection using Soccer Broadcast Videos** *In The IEEE Winter Conference on Applications of Computer Vision (WACV), 2022*

Bibliography

[1] J. Beernaerts, B. De Baets, M. Lenoir, and N. Van de Weghe. Spatial movement pattern recognition in soccer based on relative player movements. *PloS one*, 15(1):e0227746, 2020.

[2] A. Bewley, Z. Ge, L. Ott, F. Ramos, and B. Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[3] C. Bialik. The people tracking every touch, pass and tackle in the world cup. *Fivethirtyeight. com*, 2014.

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[5] Y. Chen, W. Li, C. Sakaridis, D. Dai, and L. Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[7] T. D'Orazio, M. Leo, N. Mosca, P. Spagnolo, and P. L. Mazzeo. A semi-automatic system for ground truth generation of soccer video sequences. In *2009 Sixth IEEE International Conference on Advanced Video and Signal Based Surveillance*, pages 559–564. IEEE, 2009.

[8] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, and Q. Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018.

[9] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong. Transductive multi-view zero-shot learning. *IEEE transactions on pattern analysis and machine intelligence*, 37(11):2332–2345, 2015.

[10] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[11] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun. Vision meets robotics: The kitti dataset. *The International Journal of Robotics Research*, 32(11):1231–1237, 2013.

[12] S. Giancola, M. Amine, T. Dghaily, and B. Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1711–1721, 2018.

[13] S. Giancola and B. Ghanem. Temporally-aware feature pooling for action spotting in soccer broadcasts. *arXiv preprint arXiv:2104.06779*, 2021.

[14] K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(9):1904–1916, 2015.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[16] S. Hurault, C. Ballester, and G. Haro. Self-supervised small soccer player detection and tracking. In *Proceedings of the 3rd International Workshop on Multimedia Content Analysis in Sports*, pages 9–18, 2020.

[17] D. Israni and H. Mewada. Feature descriptor based identity retention and tracking of players under intense occlusion in soccer videos. *International Journal of Intelligent Engineering and Systems*, 11(4):31–41, 2018.

[18] M. Jaderberg, K. Simonyan, A. Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

[19] W. Jiang, J. C. G. Higuera, B. Angles, W. Sun, M. Javan, and K. M. Yi. Optimizing through learned errors for accurate sports field registration. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 201–210, 2020.

[20] Y. Jiang, K. Cui, L. Chen, C. Wang, and C. Xu. Soccerdb: A large-scale database for comprehensive video understanding. In *Proceedings of the 3rd International Workshop on Multimedia Content Analysis in Sports*, MMSports '20, page 1–8, New York, NY, USA, 2020. Association for Computing Machinery.

[21] R. E. Kalman. A new approach to linear filtering and prediction problems. 1960.

[22] J. Komorowski, G. Kurzejamski, and G. Sarwas. Footandball: Integrated player and ball detector. *arXiv preprint arXiv:1912.05445*, 2019.

[23] H. W. Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[24] A. Kukleva, M. A. Khan, H. Farazi, and S. Behnke. Utilizing temporal information in deep convolutional network for efficient soccer ball detection and tracking. In *RoboCup 2019: Robot World Cup XXIII 23*, pages 112–125. Springer, 2019.

[25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[26] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *Computer Vision – ECCV 2014*. Springer International Publishing, 2014.

[27] P. Lindström, L. Jacobsson, N. Carlsson, and P. Lambrix. Predicting player trajectories in shot situations in soccer. In *Machine Learning and Data Mining for Sports Analytics: 7th International Workshop, MLSA*

2020, Co-located with ECML/PKDD 2020, Ghent, Belgium, September 14–18, 2020, Proceedings 7, pages 62–75. Springer, 2020.

[28] K. Lu, J. Chen, J. J. Little, and H. He. Light cascaded convolutional neural networks for accurate player detection. *arXiv preprint arXiv:1709.10230*, 2017.

[29] P. Lucey, A. Bialkowski, M. Monfort, P. Carr, and I. Matthews. quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data. 2015.

[30] V. Machado, R. Leite, F. Moura, S. Cunha, F. Sadlo, and J. L. Comba. Visual soccer match analysis using spatiotemporal positions of players. *Computers & Graphics*, 68:84–95, 2017.

[31] A. Maglo, A. Orcesi, and Q.-C. Pham. Efficient tracking of team sport players with few game-specific annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3461–3471, 2022.

[32] M. Manafifard, H. Ebadi, and H. A. Moghaddam. Multi-player detection in soccer broadcast videos using a blob-guided particle swarm optimization method. *Multimedia Tools and Applications*, 76(10):12251–12280, 2017.

[33] P. L. Mazzeo, P. Spagnolo, M. Leo, and T. D'Orazio. Visual players detection and tracking in soccer matches. In *2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance*, pages 326–333. IEEE, 2008.

[34] K. McSurley and G. Rybarczyk. An introduction to fieldf/x. 2011.

[35] K. Osumi, T. Yamashita, and H. Fujiyoshi. Domain adaptation using a gradient reversal layer with instance weighting. In *2019 16th International Conference on Machine Vision Applications (MVA)*, pages 1–5, 2019.

[36] Prozone Sports. Prozone, n.d. Accessed on February 27, 2015.

[37] S. Rahman, S. Khan, and N. Barnes. Polarity loss for zero-shot object detection. *arXiv preprint arXiv:1811.08982*, 2018.

[38] S. Rahman, S. Khan, and N. Barnes. Transductive learning for zero-shot object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6082–6091, 2019.

[39] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[40] J. Redmon and A. Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[41] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv preprint arXiv:1506.01497*, 2015.

[42] D. A. Reynolds. Gaussian mixture models. *Encyclopedia of biometrics*, 741:659–663, 2009.

[43] A. RoyChowdhury, P. Chakrabarty, A. Singh, S. Jin, H. Jiang, L. Cao, and E. Learned-Miller. Automatic adaptation of object detectors to new domains using self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 780–790, 2019.

[44] B. Sekachev, N. Manovich, M. Zhiltsov, A. Zhavoronkov, D. Kalinin, B. Hoff, TOsmanov, D. Kruchinin, A. Zankevich, DmitriySidnev, M. Markelov, Johannes222, M. Chenuet, a andre, telenachos, A. Melnikov,

J. Kim, L. Ilouz, N. Glazov, Priya4607, R. Tehrani, S. Jeong, V. Skubriev, S. Yonekura, vugia truong, zliang7, lizhming, and T. Truong. opencv/cvat: v1.1.0, Aug. 2020.

[45] L. Sha, P. Lucey, Y. Yue, P. Carr, C. Rohlf, and I. Matthews. Chalkboarding: A new spatiotemporal query paradigm for sports play retrieval. In *Proceedings of the 21st International Conference on Intelligent User Interfaces*, pages 336–347, 2016.

[46] R. A. Sharma, V. Gandhi, V. Chari, and C. Jawahar. Automatic analysis of broadcast football videos using contextual priors. *Signal, Image and Video Processing*, 11:171–178, 2017.

[47] B. Shuai, A. Berneshawi, X. Li, D. Modolo, and J. Tighe. Siammot: Siamese multi-object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12372–12382, 2021.

[48] D. Stadler and J. Beyerer. Modelling ambiguous assignments for multi-person tracking in crowds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 133–142, 2022.

[49] StatsPerform. Sports data - sports ai, technology, data feeds. Accessed: 2021-04-07.

[50] T. Tan. Color psychology in football: The effect of shirt color on a team's performance in the dutch eredivisie. *Erasmus University Rotterdam*, 2008.

[51] N. Veldt, D. F. Gleich, and A. Wirth. A correlation clustering framework for community detection. In *Proceedings of the 2018 World Wide Web Conference*, WWW '18, page 439–448, Republic and Canton of Geneva, CHE, 2018. International World Wide Web Conferences Steering Committee.

[52] X. Weng, J. Wang, D. Held, and K. Kitani. 3d multi-object tracking: A baseline and new evaluation metrics. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10359–10366. IEEE, 2020.

[53] M. Wieczorek, B. Rychalska, and J. Dabrowski. On the unreasonable effectiveness of centroids in image retrieval. In *Neural Information Processing: 28th International Conference, ICONIP 2021, Sanur, Bali, Indonesia, December 8–12, 2021, Proceedings, Part IV 28*, pages 212–223. Springer, 2021.

[54] N. Wojke, A. Bewley, and D. Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[55] Y. Yang and D. Li. Robust player detection and tracking in broadcast soccer video based on enhanced particle filter. *Journal of Visual Communication and Image Representation*, 46:81–94, 2017.

[56] S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

[57] L. Zheng, Z. Bie, Y. Sun, J. Wang, C. Su, S. Wang, and Q. Tian. Mars: A video benchmark for large-scale person re-identification. In *European Conference on Computer Vision*, pages 868–884. Springer, 2016.

[58] Y. Zhong, J. Yang, P. Zhang, C. Li, N. Codella, L. H. Li, L. Zhou, X. Dai, L. Yuan, Y. Li, et al. Regionclip: Region-based language-image pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16793–16803, 2022.