

FACULDADE DE ENGENHARIA DA UNIVERSIDADE DO PORTO

Intelligent Football Goals Score/Other Events Video Detection and Analytics

Carlos Miguel Guerra Soeiro



Mestrado em Engenharia Informática e Computação

Supervisor: Prof. Armando Jorge Sousa

July 27, 2025

Intelligent Football Goals Score/Other Events Video Detection and Analytics

Carlos Miguel Guerra Soeiro

Mestrado em Engenharia Informática e Computação

July 27, 2025

Abstract

Football match analysis is an important task that helps both teams and fans to obtain important information. Additionally, the huge amount of broadcasted games, many of which come from low-budget leagues and streaming or broadcasters sources, makes it interesting to have automated event detection in the broadcasted materials. Manual review of the footage would be laborious and expensive. This dissertation addresses the need for inexpensive, automated systems that companies dealing with low-budget game transmissions can use to detect events and analyse football matches.

This dissertation proposes and evaluates a deep learning framework that focuses on the automated detection of kick-off and goal events, likely extensible to other events.

For the temporal classification of the kick-off event, we train a Video Vision Transformer (ViViT) on video clips taken from raw broadcast videos. By using these video clips with data augmentation techniques, the model is able to learn the characteristics of a kick-off event. This model was trained by fine-tuning methods using video samples given by a commercial company.

Regarding goal events, we introduce a novel pipeline that aims to decompose the goal video clips into its core components. By using a combination of the You Only Look Once (YOLO) and Segment Anything Model (SAM) this dissertation aims to precisely identify and segment both the players and the goalposts. Frame subtraction is also used to isolate dynamic objects, leaving the motion of the players and the ball. The purpose is to use both of these processes to create a video mainly containing the moving ball and the goalposts, thus hoping to simplify the representation of the goal event. With these videos we then train another ViViT model to identify the goal event.

The results show the efficacy of ViViT for the detection of temporal events in a football match and the ability to create a simpler internal view of the goal event. This research contributes a practical and effective methodology for automated event detection, paving the way for future work in football analysis from raw broadcast footage.

A data set was prepared that included three full matches of soccer, of which two were women's football. The videos originated from Portuguese channel 11, associated with the Portuguese Football Federation. From those videos, 8 kick-off clips were extracted (positive samples for the kick-off event), and another 8 clips of the ball passing the centre circle were also extracted (non-kick-off samples). These clips were augmented by changing the colors of the shirts of the players, flipping, zooming, tilting and scaling the original videos. The ViViT was fine tuned with these videos and validation was done with videos never used for training. The results present a perfect score in the validation tests. For the goal event 10 clips of the goal event and the kick event (non-goal) were extracted and augmented, as we did with the kick-off event. The results achieved weren't great and a bigger dataset and a more extensive training is needed due to the complexity of the goal event.

Future work will use the above-mentioned simpler internal view to train another ViViT to classify the ball entering the posts event. Hopefully, the event of the ball entering the posts followed by the kick-off will generate a very high probability of the goal event. Similar workings can likely be extended for other events.

Keywords: Computer Vision, Transformers, Event Detection, Football

Resumo

A análise de jogos de futebol é uma tarefa importante que ajuda tanto as equipas como os adeptos a obter informações relevantes. Adicionalmente, a enorme quantidade de jogos transmitidos, muitos dos quais provenientes de ligas de baixo orçamento e de fontes de streaming ou emissoras, torna interessante a existência de deteção automática de eventos nos materiais transmitidos. A revisão manual das filmagens seria laboriosa e dispendiosa. Esta dissertação aborda a necessidade de sistemas automatizados e de baixo custo que empresas que lidam com transmissões de jogos de baixo orçamento possam usar para detetar eventos e analisar jogos de futebol.

Esta dissertação propõe e avalia uma framework de aprendizagem profunda que se foca na deteção automática dos eventos de pontapé de saída e de golo, sendo provavelmente extensível a outros eventos.

Para a classificação temporal do evento de pontapé de saída, treinamos um Video Vision Transformer (ViViT) com segmentos de vídeo retirados de transmissões. Ao usar estes segmentos de vídeo com técnicas de aumento de dados (data augmentation), o modelo é capaz de aprender as características de um evento de pontapé de saída. Este modelo foi treinado através de métodos de fine-tuning usando amostras de vídeo fornecidas por uma empresa comercial.

Relativamente aos eventos de golo, introduzimos um pipeline inovador que visa decompor os segmentos de vídeo de golo nos seus componentes centrais. Através da combinação do You Only Look Once (YOLO) e do Segment Anything Model (SAM), esta dissertação visa identificar e segmentar com precisão tanto os jogadores como os postes da baliza. A subtração de frames é também utilizada para isolar objetos dinâmicos, restando o movimento dos jogadores e da bola. O propósito é usar ambos os processos para criar um vídeo que contenha principalmente a bola em movimento e os postes da baliza, esperando assim simplificar a representação do evento de golo. Com estes vídeos podemos treinar outra ViViT para identificar o evento de golo.

Os resultados mostram a eficácia do ViViT para a deteção de eventos temporais num jogo de futebol e a capacidade de criar uma visão interna simplificada do evento de golo. Esta investigação contribui com uma metodologia prática e eficaz para a deteção automática de eventos, abrindo caminho para trabalhos futuros na análise de futebol a partir de filmagens de transmissão.

Foi preparado um conjunto de dados que incluiu três jogos completos de futebol, dos quais dois eram de futebol feminino. Os vídeos tiveram origem no Canal 11, associado à Federação Portuguesa de Futebol. Desses vídeos, foram extraídos 8 segmentos de pontapé de saída (amostras positivas para o evento de pontapé de saída), e foram também extraídos outros 8 segmentos da bola a passar o círculo central (amostras de não-pontapé de saída). Estes segmentos foram aumentados através da alteração das cores das camisolas dos jogadores, e da inversão, zoom, inclinação e redimensionamento dos vídeos originais. O ViViT foi afinado com estes vídeos e a validação foi feita com vídeos nunca antes usados para treino. Os resultados apresentam uma pontuação perfeita nos testes de validação. Para o evento de golo, foram extraídos e aumentados 10 clips do evento golo e 10 segmentos de um evento de pontapé (não-golo), tal como fizemos para o evento de

pontapé de saída. Os resultados alcançados não foram bons e é necessário um conjunto de dados maior e um treino mais extensivo devido à complexidade do evento de golo.

Trabalhos futuros irão usar a visão interna simplificada acima mencionada para treinar outro ViViT para classificar o evento da bola a entrar na baliza. Espera-se que o evento da bola a entrar na baliza, seguido pelo pontapé de saída, gere uma probabilidade muito alta do evento de golo. Funcionamentos semelhantes poderão provavelmente ser estendidos a outros eventos.

Palavras-chave: Visão computacional, Transformers, Detecção de eventos, Futebol

UN Sustainable Development Goals

The United Nations Sustainable Development Goals (SDGs) provide a global framework to achieve a better and more sustainable future for all. It includes 17 goals to address the world's most pressing challenges, including poverty, inequality, climate change, environmental degradation, peace, and justice.

The specific Sustainable Development Goals mentioned have the following names:

SDG 9 Build resilient infrastructure, promote inclusive and sustainable industrialization, and foster innovation.

SDG 10 Reduce inequality within and among countries

SGD	Target	Contribution	Performance Indicators and Metrics
9	9.5	Advancing research and innovation in automatic event detection to solve a practical problem in the sports industry.	Creation of a novel analysis pipeline combining ViViT, YOLO, SAM, and frame subtraction; Documented methodology and findings.
	9.b	Development of a "low-cost" solution and the use of footage from Portuguese football, providing a blueprint for smaller leagues.	Use of open-source models (YOLO, SAM); A methodology that avoids proprietary hardware.
10	10.2	The creation of a "low-cost" solution, helping in reducing the inequality between high-budget and low-budget sports organizations.	System designed for low-budget contexts; Use of matches from both men's and women's football.

Contents

1	Introduction	1
1.1	Context	1
1.2	Motivation	1
1.3	Objectives	2
1.4	Document Structure	2
2	State of the Art	3
2.1	Computer Vision	3
2.1.1	Object Detection	4
2.1.2	Image Segmentation	5
2.1.3	Event Detection	5
2.2	Artificial Intelligence	5
2.2.1	Machine Learning	5
2.2.2	Deep Learning	6
2.2.3	Supervised Learning	6
2.2.4	Unsupervised Learning	6
2.2.5	Reinforcement Learning	6
2.3	Deep Learning for Video Event Recognition	6
2.3.1	CNN-based Approaches	7
2.3.2	Transformers	8
2.3.3	Vision Transformer (ViT)	8
2.3.4	Video Vision Transformer (ViViT)	8
2.4	Object Detection and Segmentation	9
2.4.1	You Only Look Once (YOLO)	9
2.4.2	Segment Anything Model (SAM)	10
2.5	Deep Learning Frameworks	11
2.5.1	PyTorch	11
2.5.2	TensorFlow	11
2.6	Datasets	11
2.6.1	SoccerNet	11
2.6.2	SoccerDB	12
2.6.3	MultiSports	13
2.6.4	Goal	13
3	Methodology	15
3.1	Intelligent Football Detector (IFD)	15
3.2	Data Acquisition and Preparation	16
3.3	Kick-Off Event Detection with ViViT	17

3.4	Goal Event Isolation and Analysis	19
3.4.1	Player Detection	19
3.4.2	Goalpost Detection	21
3.4.3	Motion Isolation via Frame Subtraction	22
3.4.4	Goal Event Detection with ViViT	22
4	Results and Analysis	23
4.1	Kick-Off Event Classification Results	23
4.2	Player and Goalpost Detection Results	26
4.3	Goal Event Isolation Results	28
4.4	Goal Event Classification Results	30
5	Conclusion and Future Work	32
5.1	Researched contributions	32
5.2	Limitations	32
5.3	Future Work	33
	References	34

List of Figures

2.1	Example of Object Detection. Extracted from [24]	4
2.2	Example of Image Segmentation. Extracted from [19]	5
2.3	Example of Event Detection. Extracted from [8]	5
2.4	Two-stream architecture for video classification. Extracted from [27]	7
2.5	Transformer architecture. Extracted from [32]	8
2.6	Tubelet embedding. Extracted from [1]	9
2.7	YOLO model. Extracted from [25]	10
2.8	Segment Anything Model (SAM) overview.. Extracted from [16]	10
2.9	Example of events annotated by SoccerNet. Extracted from [9]	12
2.10	Example of some annotations in SoccerDB. Extracted from [12]	13
2.11	Example of some annotations in MultiSports. Extracted from [18]	13
2.12	Resources used for the Goal dataset. Extracted from [28]	14
3.1	Pipeline for IFD	15
3.2	Example of a video clip for the kick-off event	16
3.3	Example of a video clip for the goal event	17
3.4	Example of data augmentation for changing color of team equipments	18
3.5	Training Arguments	19
3.6	Example of player detection using YOLO	20
3.7	Full segmentation of a frame taken from a goal clip using SAM	20
3.8	Goalpost dataset example	21
3.9	Detection of the goalposts using the trained YOLO model	21
4.1	Kick-Off ViViT Results	23
4.2	Evaluation Accuracy	24
4.3	Training Loss	24
4.4	Evaluation Loss	25
4.5	Player Detection Results YOLO	26
4.6	Player Detection Results YOLO + SAM	26
4.7	GoalPost Detection Result YOLO11-obb	27
4.8	Frames Subtracted Video	28
4.9	Frames Subtracted Video + Remove Players (YOLO+SAM) + Goalposts (YOLO)	29
4.10	Frames Subtracted Video + Remove Players (YOLO) Bounding Box + Goalposts segmentation (YOLO + SAM). The predicted goal posts are marked white and the ball with heavy motion blur appears with two trails	30
4.11	Goal ViViT Results	30
4.12	Goal Evaluation Accuracy	31
4.13	Goal Evaluation Loss	31

List of Acronyms

FEUP	Faculdade de Engenharia da Universidade do Porto
CNN	Convolutional Neural Networks
ViT	Vision Transformers
ViViT	Video Vision Transformers
YOLO	You Only Look Once
SAM	Segment Anything Model
IFD	Intelligent Football Detector
VAR	Video Assistant Refere

Chapter 1

Introduction

In this chapter a brief introduction is made to outline the context and the objectives of this dissertation.

1.1 Context

This project was developed with the help of Armis Digital Sport ¹, as a dissertation for the Mestrado em Engenharia Informática e Computação at Faculdade de Engenharia da Universidade do Porto (FEUP).

The project presented by Armis was the identification of events in football matches, with the objective of improving the collection of statistics and data from lower leagues. This data can be used to help improve the reach of these leagues for both professionals in the football area and football fans.

The Federação Portuguesa de Futebol (FPF) together with Canal 11 are working with Pixellot ² to install camera systems provided by Pixellot in the stadiums of smaller teams. The objective is to enhance the analysis of football matches, especially in smaller leagues that do not have the resources to do so.

1.2 Motivation

The collection of football matches statistics is a task that requires a lot of effort. A lot of this collection is done by human hands and with the help of expensive technology, such as, very high quality cameras. This isn't viable to do for lower leagues, that don't have the resources to do this for all games.

¹<https://www.armisdigitalsport.com/>

²<https://www.pixellot.tv/>

With the increase of matches from lower leagues that are broadcasted, a low cost, automatic solution that would be able to get statistics of matches would be an ideal fit for these situations.

1.3 Objectives

With this project we aim to create a pipeline capable of detecting goal events in football matches (and, in future, other game events). Because our aim is to implement this project in lower leagues, the goal is to provide a solution that is both cheap and automatic and can be used in several situations.

In order to achieve this we need to detect both the goal event and the kick-off event. The kick-off event is very important because it allows us to verify if a goal is valid or if it was annulled by the referee or the VAR. With this, we can check if an event of kick-off happened after a goal in order to validate it.

1.4 Document Structure

In the first chapter a brief introduction to this dissertation is made. The context of the project is mentioned as well as the objectives of this dissertation.

The second chapter details several existing concepts and projects that relate to this dissertation. Several recent and useful technologies are presented and explored.

The third chapter introduces the pipeline for this project. It also consists of the detailing of the several steps taken to achieve the goal of the pipeline.

Fourth chapter shows the results obtained as well as the analyses of those results.

The fifth chapter is the conclusion. It talks about what has been achieved with this project, the limitations and what can be done in the future.

Chapter 2

State of the Art

2.1 Computer Vision

As humans, it is a simple task to perceive and understand the world around us. We are able to identify objects, their shape, color and translucency without any effort because of our eyes, brain and a lifetime of training. Computer vision is a field of artificial intelligence that enables computers to analyze scenes and achieve a detailed understanding of visual inputs, such as digital images and videos [35]. Ultimately, the aim of this field is to automate visual tasks and perform them faster and better than a human can.

Some of these visual tasks are:

- Image acquisition: Capturing visual data through cameras or other sensors.
- Image processing: Enhancing, manipulating, and filtering the captured data.
- Feature extraction: Identifying key characteristics and patterns within the image.
- Object recognition: Classifying and labeling objects present in the image.
- Scene understanding: Interpreting the overall context and meaning of the visual scene.

Computer vision heavily relies on machine learning algorithms, particularly deep learning, which involves training complex neural networks on vast amounts of visual data. These networks learn to recognize patterns and relationships within images, enabling tasks like object detection, image segmentation, facial recognition and event detection [10].

In the last decade, datasets such as ImageNet helped in the creation of convolutional neural networks (CNN) and deep learning models that laid the foundation for the creation of models with very high success in computer vision tasks [10].

The applications of computer vision are vast and extend into various areas. Some of its applications are:

- Self-driving cars: Recognizing traffic signs, pedestrians, and other obstacles for safe navigation.
- Medical imaging: Analyzing X-rays, MRI scans, and other medical images for diagnosis and treatment planning.
- Security and surveillance: Identifying suspicious activities and objects in real-time.
- Augmented reality and virtual reality: Creating immersive experiences by overlaying digital information on the real world.
- Sports: Tracking the player and the ball and analyzing video footage to identify specific actions like shots, passes or tackles.

2.1.1 Object Detection

Object detection is a field within computer vision that focuses on identifying and locating objects within images or videos [24]. The end result is an image with a bounding box around a determined object and its class, as seen in Figure 2.1.



Figure 2.1: Example of Object Detection. Extracted from [24]

Some of the most common challenges include: objects being partially or fully hidden behind other objects making it difficult to accurately detect and classify them, variations in appearance for objects in the same class and limited training data [4].

2.1.2 Image Segmentation

Image segmentation is the process of dividing a digital image into multiple meaningful segments or regions based on certain characteristics. The goal is to label sections of the image by identifying groups of pixels that are related in some way [19] (see Figure 2.2).



Figure 2.2: Example of Image Segmentation. Extracted from [19]

2.1.3 Event Detection

Event detection refers to the process of identifying and understanding significant occurrences within a specific image or video. It aims to extract meaningful information about these events, such as their type, time and location [8].

In Figure 2.3 we can see an example of a goal with markers describing the state of the game and the different types of events.



Figure 2.3: Example of Event Detection. Extracted from [8]

2.2 Artificial Intelligence

2.2.1 Machine Learning

Machine learning is a subset of artificial intelligence which aims to allow machines to do tasks associated with the human mind, such as learning and solving problems [26]. It is a technique that uses computational methods to create models capable of learning from past experience and make predictions on new data [37]. Some of the applications of machine learning include computer vision, prediction, semantic analysis, natural language processing and information retrieval [26].

2.2.2 Deep Learning

With the advancement of machine learning a new subset was created called deep learning [26]. One of the disadvantages of machine learning was the amount of human effort required for feature design and deep learning fixes that problem by being able to automatically identify the features most useful for a specific task [14]. Deep learning consists of neural networks with several layers that extract features and do transformations [26].

2.2.3 Supervised Learning

Supervised learning is a machine learning technique that receives data that is already labeled and uses it to learn to predict the correct label for new data. There are two types of supervised learning: classification, that given the input data it outputs a category label, and regression, that outputs a continuous value [20].

2.2.4 Unsupervised Learning

Unlike supervised learning, unsupervised learning does not receive data with labels. The goal is to use this data to find patterns or groups with elements that are similar [20].

Clustering is a unsupervised learning technique that aims to group unlabeled data that shares common characteristics. This method uses distance metrics to measure the similarity between data and algorithms to maximize the similarity between data clusters [31].

Another unsupervised learning technique is dimensionality reduction that is used for data compression and data cleaning. The objective of this technique is to reduce the dimension of data while maintaining its dispersion shape [7].

2.2.5 Reinforcement Learning

In reinforcement learning we have an agent that is able to obtain information by interacting with the environment. The agent receives a reward every time it decides on an action in a given state, following a trial and error process. The goal is to learn what actions give the best reward for each state [20].

2.3 Deep Learning for Video Event Recognition

The task of event detection in videos has had major improvements in the last years with the use of deep learning [21]. Good results were obtained early on with Convolutional Neural Networks (CNNs), by translating the results obtained in 2D images into the temporal dimension of videos. Recently, with the introduced transformer models, video event detection has achieved even better

results. By capturing long-range dependencies, they can effectively obtain temporal and spacial information, allowing these models to have state of the art performance. [17]

2.3.1 CNN-based Approaches

At first, CNNs were adapted for video using sequences of frames with no temporal and spatial information between them. In an attempt to improve this basic implementation several projects were developed.

The Two-Stream Convolutional Network was an attempt at obtaining both temporal and spatial information by using two separate CNNs, one used RGB frames to get the spatial data and the other used optical flow for temporal data [27], see figure 2.4.

Another way to learn spatial and temporal information was with the use of 3D Convolutional Networks. At the time, the use of 3-dimensional convolutional networks achieved state of the art results when compared to previous attempts [30] [3].

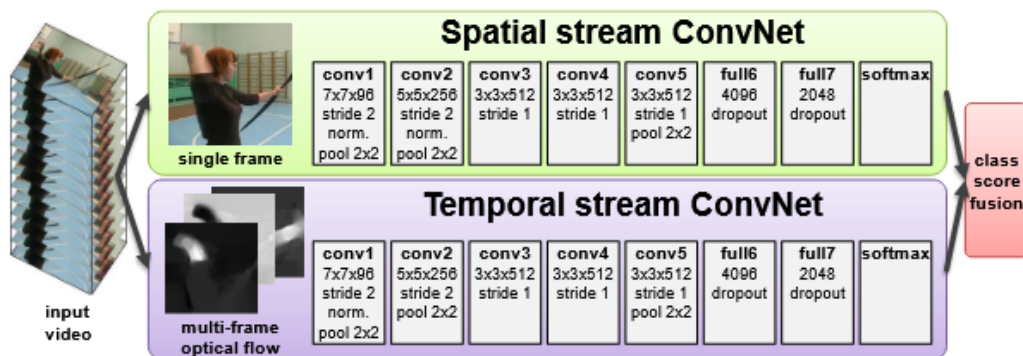


Figure 2.4: Two-stream architecture for video classification. Extracted from [27]

2.3.2 Transformers

The transformer architecture was first developed for natural language processing [32], see figure 2.5. One of the advantages of this architecture is the self-attention mechanism, it allows the model to weigh the importance of different parts of the input sequence in order to know the relation and context between them [33]. It is also capable of working with bigger datasets and is able to be used in several tasks besides natural language processing.

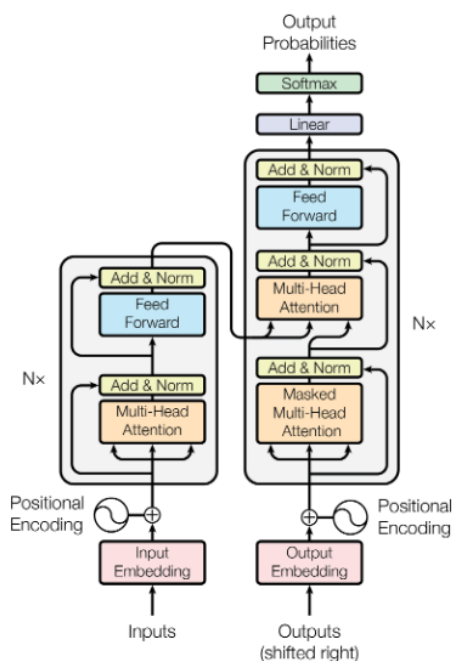


Figure 2.5: Transformer architecture. Extracted from [32]

2.3.3 Vision Transformer (ViT)

Vision Transformer (ViT) is a neural network architecture that adapted the transformer architecture to computer vision. It works by splitting images into patches of tokens in a sequence, similar to the way transformers work for natural language processing. With this, the model is able to obtain long-range dependencies and relationships in the images [6].

2.3.4 Video Vision Transformer (ViViT)

Video Vision Transformer (ViViT) is an architecture that builds on ViT in order to apply the transformer architecture to computer vision tasks in videos. To achieve this, the videos are first processed by turning them into tubelets. These tubelets work as the patches in ViT but now they also contain the time dimension, see figure 2.6. With these spatio-temporal patches and with the transformer architecture, the model is able to learn complex information in both space and time

[1]. This ability to understand long range features is a perfect fit to learn complex events like a kick-off in a football match, that is defined by a sequence of actions over time.

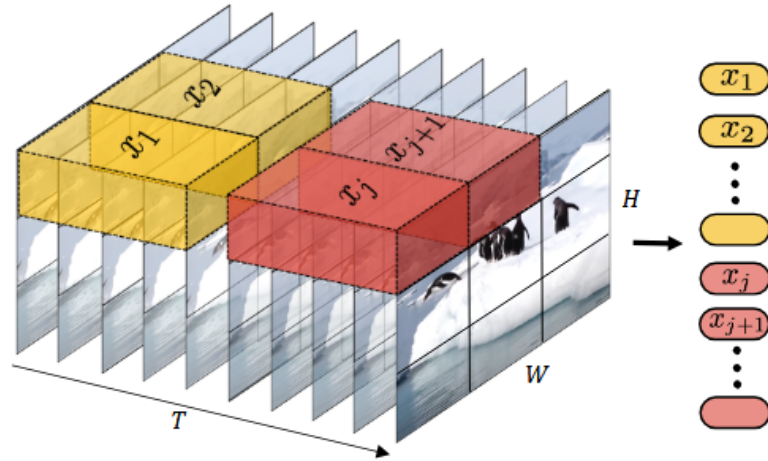


Figure 2.6: Tubelet embedding. Extracted from [1]

2.4 Object Detection and Segmentation

Another important task in the understanding of a scene is the ability to locate an object in an image. Both object detection and segmentation try to achieve this task by either using bounding boxes to delineate objects or finding every pixel in an image that relates to a given object.

2.4.1 You Only Look Once (YOLO)

Unlike other methods, the YOLO models are capable of object detection in real-time. It is able to predict bounding boxes with a single neural network to process the entire image which, allows for a fast prediction. It also has a very high accuracy for most object detection tasks compared to other models. Besides the bounding box prediction it also can classify several objects in a scene. The detection of objects is done by dividing the image into several grids and predicting bounding boxes and classes in each of them [25], see figure 2.7.

This deep learning model is regularly updated with new versions that add new functionalities and achieve better results. Because of its speed and high accuracy it is a very nice fit for detecting objects in football videos.

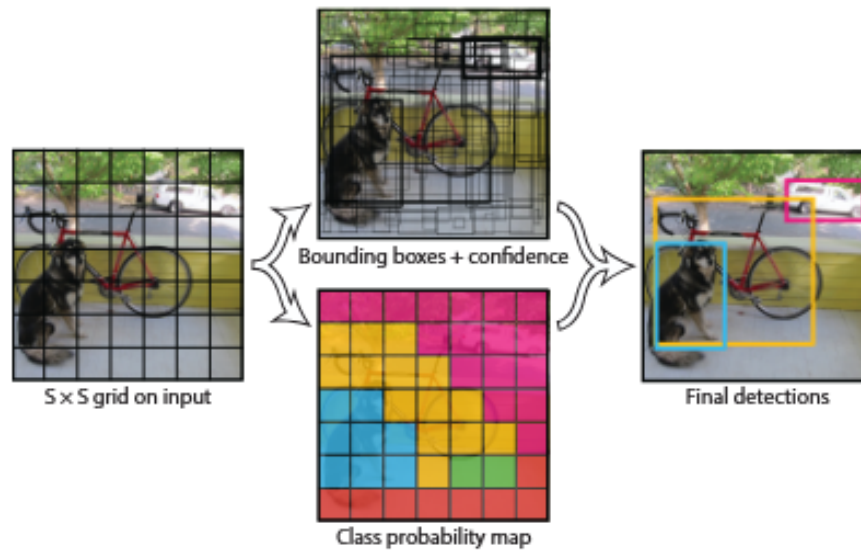


Figure 2.7: YOLO model. Extracted from [25]

2.4.2 Segment Anything Model (SAM)

The SAM model instead of predicting bounding boxes, it predicts segmentation masks. These masks identify all pixels belonging to an object, making it a more precise outline of the object, see figure 2.8. This model is trained on a huge dataset of over on billion masks on 11 million images. The goal is to segment every object in a given image even if not previously trained on it. The user of this model can use points or boxes to specify the objects it wants to identify [16].

For our case, the SAM model can accurately identify objects in football matches like, the players and the goalposts. It also doesn't require training or fine tuning, saving us some time and resources.

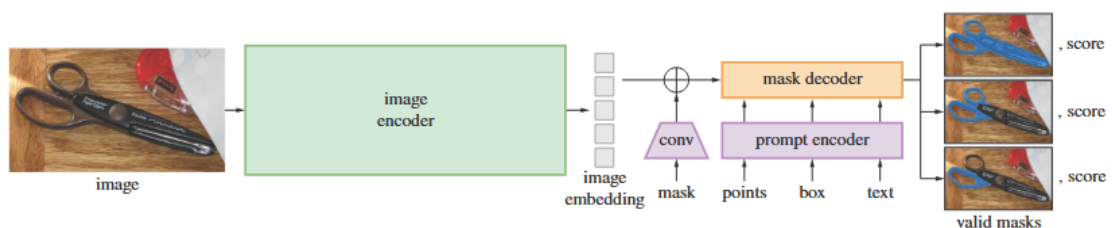


Figure 2.8: Segment Anything Model (SAM) overview.. Extracted from [16]

2.5 Deep Learning Frameworks

2.5.1 PyTorch

PyTorch is a machine learning framework that is built on top of the Python programming language. It is a simple and fast framework that allows for a smooth experience. Some of its best features involve being flexible which let the user customize several models to fit their needs and a dynamic computational graph, which means the user can define and change computational diagrams during runtime [11].

2.5.2 TensorFlow

Tensorflow is an open-source platform for machine learning that uses dataflow graphs. It was developed by the Google Brain team with the goal of researching machine learning and deep neural networks. This framework allows the user to preprocess the data, develop a model and train and test that model. Some of the benefits are its flexibility, scalability, a vast and up to date library and a good community [23] [22].

2.5.2.1 Keras

Keras is the high-level API for Tensorflow. It is a fast and approachable framework for machine learning, focusing on modern deep learning. Also, it allows Tensorflow to be scalable and cross-platform [29].

2.6 Datasets

In the context of computer vision, datasets are a collection of digital images or videos that have been annotated with specific information relevant to the task at hand. These annotations typically describe the content of the data through labels, bounding boxes, or other relevant information, depending on the specific task. Their purpose is to provide the necessary data needed for algorithms to learn and improve and evaluate their performance in specific tasks.

For this dissertation, the relevant datasets will be those that have annotations pertaining to events in a football match such as goals, shots, red or yellow cards and substitutions. Annotations for the position of the players and ball can also be helpful.

2.6.1 SoccerNet

The SoccerNet dataset consists of 500 complete soccer games from six major European leagues spanning three seasons (2014-2017), with a total duration of 764 hours. Its main goal is the training and evaluation of algorithms for action spotting in football matches. This allows the creation of algorithms capable of extract the highlights from a match that can be used to create match summaries, the retrieval of specific events or the analysis of the match, for example, the

goals scored. The annotations for each video consist of temporal labels with one minute resolution of goals, red and yellow cards and substitutions [9] (see Figure 2.9).



Figure 2.9: Example of events annotated by SoccerNet. Extracted from [9]

2.6.2 SoccerDB

SoccerDB is a large-scale dataset dedicated to comprehensive video understanding in the context of football games. It is composed by 171,191 video segments extracted from 346 football games that cover a wide range of game play situations. The annotations contain bounding box for 702,096 objects identified within the frames, such as players, the ball, and goals, 37,709 crucial events labels that are tagged with their time boundaries and 17,115 highlights labels (see Figure 2.10). The rich and diverse annotations make this dataset adequate for many tasks, such as object detection, action recognition, temporal action localization, and highlight detection [12].

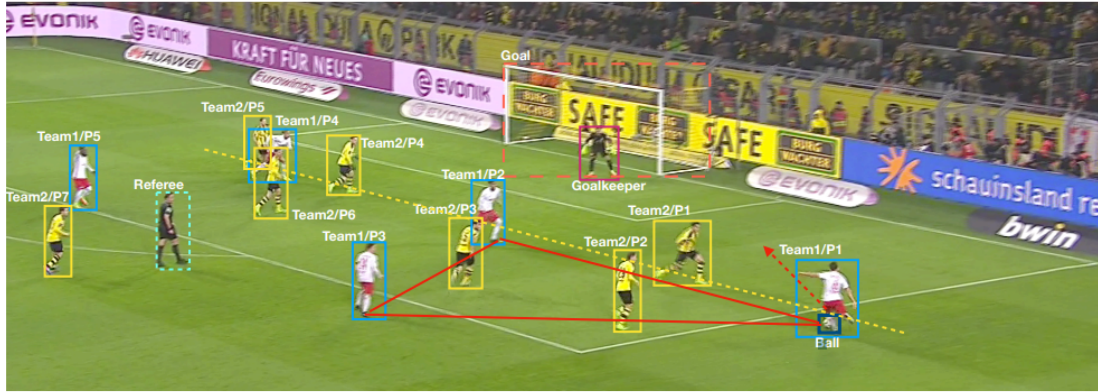


Figure 2.10: Example of some annotations in SoccerDB. Extracted from [12]

When compared to SoccerNet, SoccerDB offers a broader scope with bounding boxes, detailed event labels, and highlight annotations, while SoccerNet focuses more on action spotting for predefined events.

2.6.3 MultiSports

As in the previous datasets, the Multisports dataset also aims to help in the task of action spotting with the slight difference in containing videos for 4 different sports: basketball, football, volleyball, and aerobic gymnastics. Its data consists of 3,200 high-quality video clips showcasing various actions with 37,701 action instances and bounding boxes around the relevant players involved in the action and their corresponding timestamps [18] (see Figure 2.11).



Figure 2.11: Example of some annotations in MultiSports. Extracted from [18]

2.6.4 Goal

The Goal dataset is a bit different than the previous ones because it focuses on football highlights videos with transcribed live commentaries in English. It was created to facilitate the exploration of the interaction between language and visual information in dynamic scenarios like sports broadcasts. Its annotations consist of video of highlights in a football match with their corresponding live commentary transcript (see Figure 2.12). With this data it could be possible to identify specific moments within a football match based on the commentator's description [28].



Figure 2.12: Resources used for the Goal dataset. Extracted from [28]

Chapter 3

Methodology

This chapter presents the Intelligent Football Detector (IFD) that is proposed in this work.

3.1 Intelligent Football Detector (IFD)

Football is a very complex game with many different events. Because of this, we have decided to focus on two events: the goal and the kick-off. The most important event in football is the goal, and that is why we have chosen it. We also chose the kick-off event due to the possibility of a goal being overturned with the use of Video Assistant Referee (VAR). In this way, we can confirm that a goal is valid if it is followed by a kick-off. So, by doing this confirmation with both events occurring together in a short time frame, we are able to know the score of the football game.

The IFD pipeline is represented in 3.1 and is divided into two main parallel pipelines, each related to the goal event and the kick-off event.

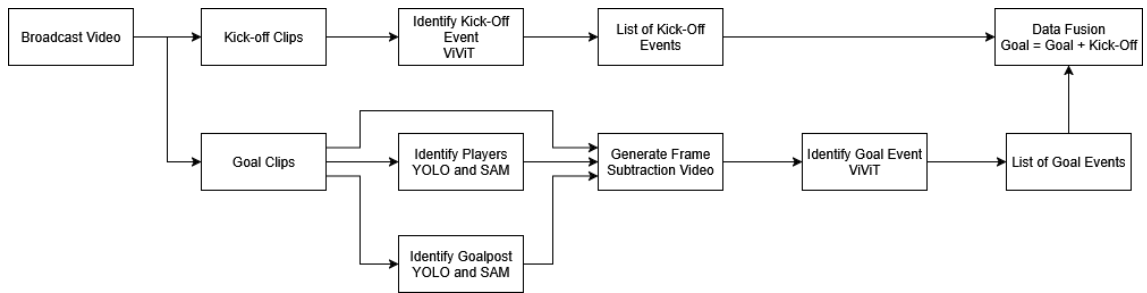


Figure 3.1: Pipeline for IFD

Figure 3.1 is the full proposed pipeline, at the time of writing, the "Identify Goal Event" and the data fusion blocks are future work.

3.2 Data Acquisition and Preparation

At the start of our pipeline is the input data that, in this case, consists of videos of football matches from lower leagues. These broadcast videos are usually less edited and include less repetitions, camera zooms, angles and camera movements than broadcasts from the first football leagues. The broadcast videos we are using have been provided by Armis and include two matches from the top-level Portuguese women's football league, Liga BPI, and one match from the third division of professional men's football in Portugal, Liga 3.

Using these videos we then proceed to cut small video clips that include the events we are interested in. These short video clips don't include any repetition, change of camera or camera zooms. This is done for both the goal event and the kick-off event but also for normal ball passes that occur in the center of the pitch and shots that don't result in goals, see figure 3.2 and figure 3.3 for an example of the video clips. These last two events are also collected in order to help with any false negatives that might happen when trying to detect a kick-off and a goal.



Figure 3.2: Example of a video clip for the kick-off event

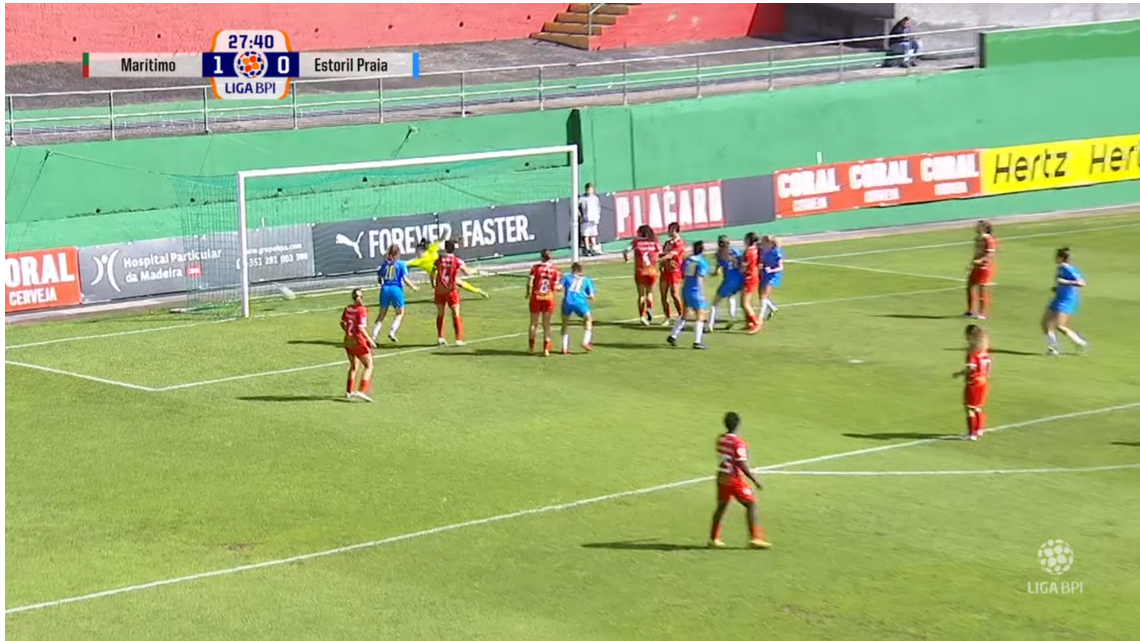


Figure 3.3: Example of a video clip for the goal event

3.3 Kick-Off Event Detection with ViViT

In order to detect the kick-off event, we have decided to use ViViT which is a pure-transformer based model for video understanding that is achieving great results in video classification tasks.

To start using this model we first need to create the dataset that is going to be used to train it. In this case the dataset consists of the video clips of both the kick-off events and the normal ball passes in the center of the field events that we got from the full broadcast video.

Because our dataset is small we used data augmentation to increase the amount of data we have. This is done in order to increase the diversity of the dataset and prevent overfitting. The techniques used are applying zoom, rotations, scaling, mirroring to the videos and change the colors of the team equipments of both teams.

For the changing of the equipment colors a small program was developed. This program opens the video and asks the user to first click the equipment of three different players of the first team and then do the same for the second team. After it gets the average color of each team's equipment it changes the hue of every pixel in the video that has that color into a different color. The final result is a video where the teams have different colors for their equipments, example in figure 3.4.



Figure 3.4: Example of data augmentation for changing color of team equipments

The ViViT model we use is pre-trained in the Kinectics 400 dataset [13] and was taken from the Google organization on Hugging Face, it is called `vivit-b-16x2-kinetics400`. This model corresponds to the proposed ViVit model Spatio-temporal attention.

The ViVit was trained on our dataset for 3 epochs, with learning rate of $5e-05$, evaluate steps of 10 and the optimizer used was AdamW. More information for the training arguments can be found in figure 3.5.


```

training_args = TrainingArguments(
    output_dir=training_output_dir,
    num_train_epochs=3,
    per_device_train_batch_size=2,
    per_device_eval_batch_size=2,
    learning_rate=5e-05,
    weight_decay=0.01,
    logging_dir="./logs",
    logging_steps=10,
    seed=42,
    eval_strategy="steps",
    eval_steps=10,
    warmup_steps=int(0.1 * 20),
    optim="adamw_torch",
    lr_scheduler_type="linear",
    fp16=True,
    report_to="wandb"
)

```

Figure 3.5: Training Arguments

The model performance and evaluation metrics were tracked by WandB [2]. WandB is a machine learning platforms that helps with the tracking of metrics such as, loss, learning rate and accuracy.

3.4 Goal Event Isolation and Analysis

The objective of this pipeline is the detection of the goal event. This event is very complex because of all the variations it can have. Due to all of the complexity of detecting this event we attempted to alter the small video clips of the goal event in order to facilitate the visualization of the ball entering the goalposts. The objective is to remove all the unnecessary components from the video and be left with only the ball and the goalposts.

3.4.1 Player Detection

First thing that is necessary to remove from the video is the players and referees. In order to achieve this there needs to be a way to detect every pixel that corresponds to a player. The project Football player tracking by Bader Dammak on github [5] does exactly that. It uses YOLOv8 and object detection to detect the players, referees and goalkeepers. By using the pre-trained YOLO model from this project we can get the bounding box that surrounds the players for our videos, see figure 3.6.

Now we have the bounding box for each player but, the bounding box contains a lot of pixels that don't belong to the player. To fix this we need to use a segmentation model. The segmentation model we use is the Segment Anything Model (SAM) developed by Meta AI [16]. This state of the art segmentation model is capable of, given an image and a bounding box, segmenting the

objects present in the image (see figure 3.7 or the objects selected using a point or bounding boxes. So, by joining both the YOLO model and the SAM model we can now get every pixel of every player that is in the video. The SAM model used was the VIT-H SAM model.

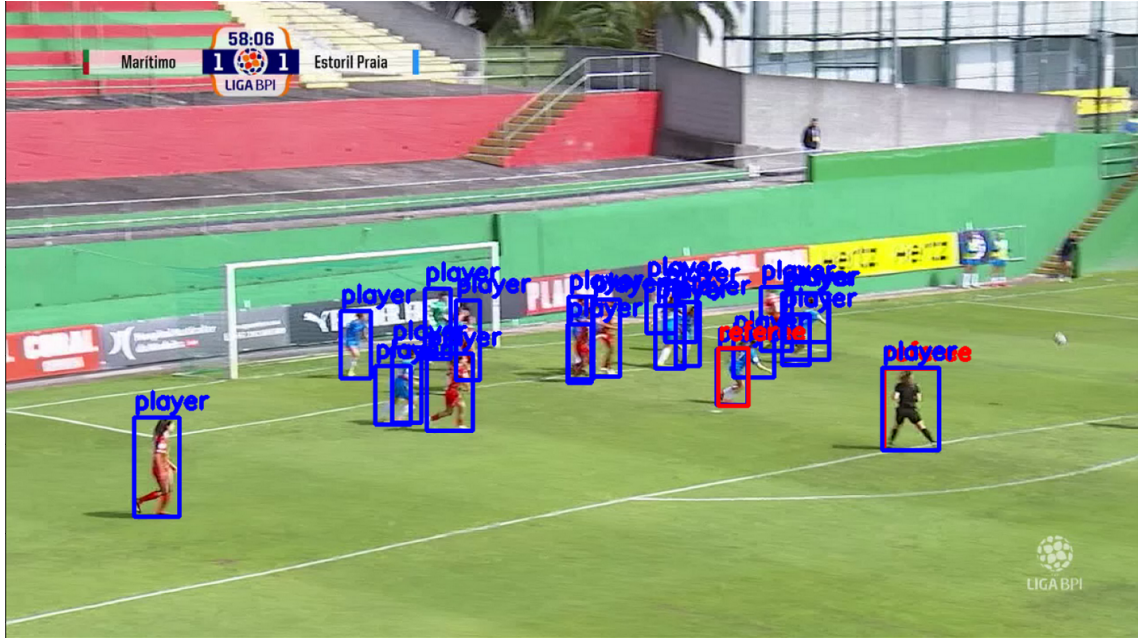


Figure 3.6: Example of player detection using YOLO



Figure 3.7: Full segmentation of a frame taken from a goal clip using SAM

3.4.2 Goalpost Detection

For the detection of the goalposts a YOLO11 [15] model was used. The model used was YOLO11m-obb that is pre-trained on the dataset DOTAv1 [34]. This model is capable of detecting oriented bounding boxes which is helpful for our objective of detecting the goalposts. The dataset used to train this YOLO model was a goalpost dataset by yolo on the platform Roboflow [36], figure 3.8 has an example from the dataset. This dataset consists of several images of goalposts that are annotated with oriented bounding boxes and has 490 images for training, 87 images for evaluation and 12 images for testing. A detection obtained by this YOLO model can be seen in figure 3.9.



Figure 3.8: Goalpost dataset example

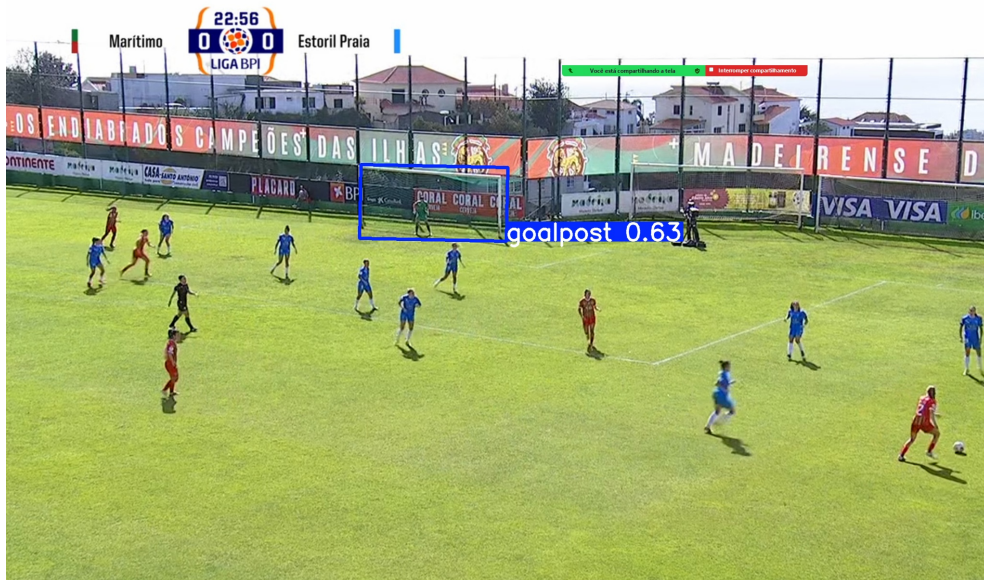


Figure 3.9: Detection of the goalposts using the trained YOLO model

3.4.3 Motion Isolation via Frame Subtraction

Frame subtraction is a technique that allows to remove stationary objects from the video and highlight the movement of moving objects. By using this technique with our small goal clips we can effectively remove most of the objects from the image and get closer to the objective of generating a video with only the ball and the goalposts. To achieve this we only need to subtract each frame of the video with the previous frame and save the new video.

After we apply the frame subtraction to the goal video we notice that besides the ball movement there is also other movements to try and remove. The main movement we want to remove is the players moving. For that, we use the previous YOLO and SAM models to identify the pixels of the players in the original video and then remove those pixels from our subtracted video.

Another task that needs to be done is add the goalposts back to the video because, as they are a stationary object, they don't appear in the subtracted video. Now, using the YOLO model that we trained on the goalpost dataset we can once again identify the goalposts in the original video and then draw them in the subtracted video. Not only do we draw the goalposts but also everything inside of it, this way it makes it easier to detect if the ball has entered inside the goalposts.

After all this, the objective is to have a video of a goal where only the ball and the goalposts are visible so the model can then have better results when trying to identify the goal event. The results can be seen in figure 4.10

3.4.4 Goal Event Detection with ViViT

Finally, with the simplified video we can train the ViViT to detect the goal event. We trained the model with 10 clips of Goal and 10 clips of Kick (kicks that didn't result in goal). Data augmentation was used to increase the data available to make the training of the model more robust. The same process was done as with the training of the ViViT for the kick-off event described in section 3.3.

Chapter 4

Results and Analysis

In this chapter the results are presented and analysed.

4.1 Kick-Off Event Classification Results

The ViViT model trained on our dataset for 3 epochs and achieved a precision, recall and f1-score of 1, as we can see in figure 4.1. In all the 24 video clips of Kick-Off and 8 video clips of NormalPlay the model was able to identify correctly which event it was.

	precision	recall	f1-score	support
KickOff	1.00	1.00	1.00	24
NormalPlay	1.00	1.00	1.00	8
accuracy			1.00	32
macro avg	1.00	1.00	1.00	32
weighted avg	1.00	1.00	1.00	32

Figure 4.1: Kick-Off ViViT Results

Figure 4.2 shows the accuracy of the model over time. We can see that the accuracy increases almost until the end and then decreases for the last steps. This can be a sign of overfitting.

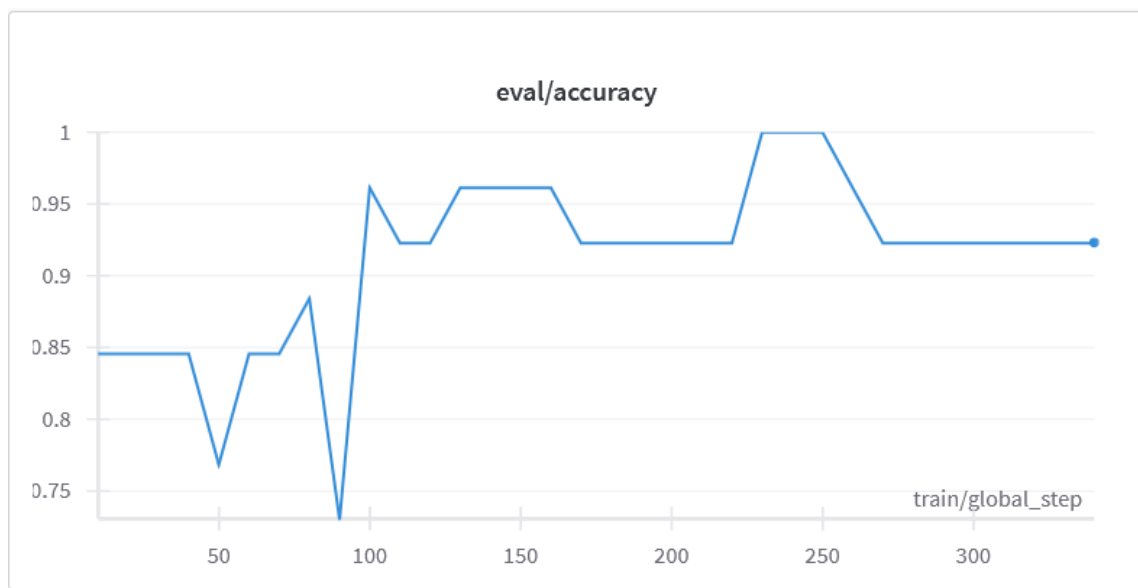


Figure 4.2: Evaluation Accuracy

By looking at the loss graph we can examine that the value of loss has lowered with the increase of the training steps, see figure 4.3. This means the model was fitting to the training data.

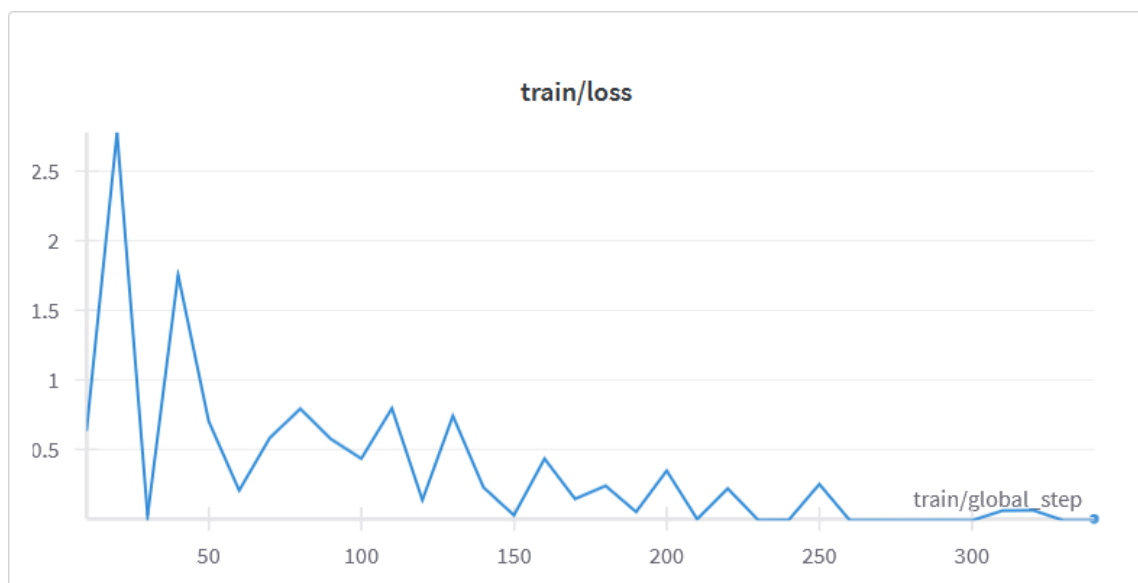


Figure 4.3: Training Loss

The loss value for the evaluation loss graph starts to lower until around step 250 and then starts to increase, see figure 4.4. This curve shows that the model is overfitting to the data which causes a reduction in performance on new data. This is a sign that the dataset is too small and usually happens when the model can't generalize, it fits too closely to the training set.

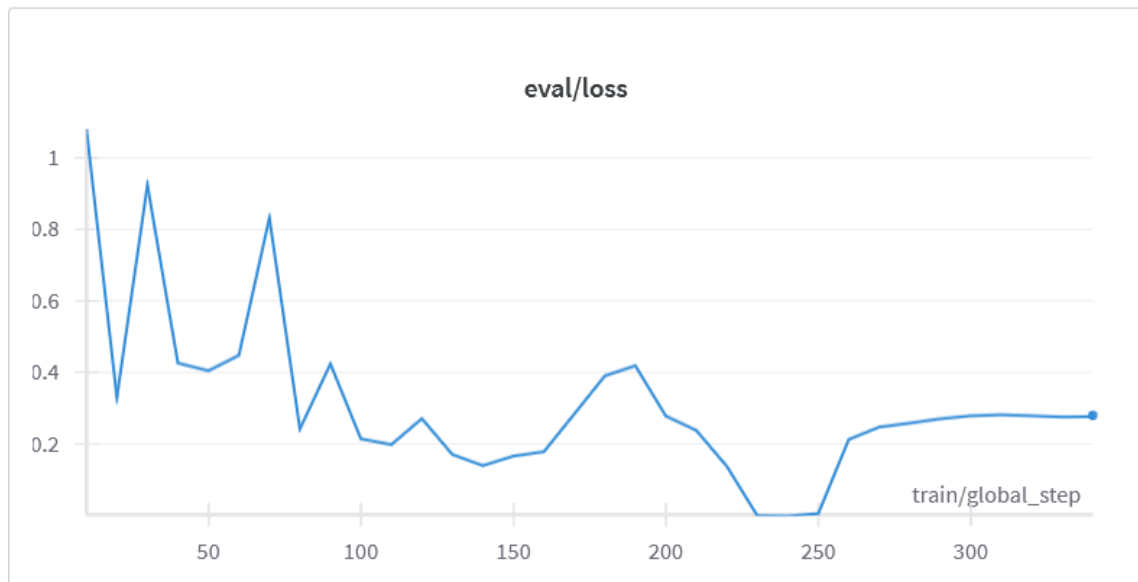


Figure 4.4: Evaluation Loss

4.2 Player and Goalpost Detection Results

For the first step of detecting players using YOLO's object detection, we can see that the model has produced excellent results by detecting every player and referee, see figure 4.5.

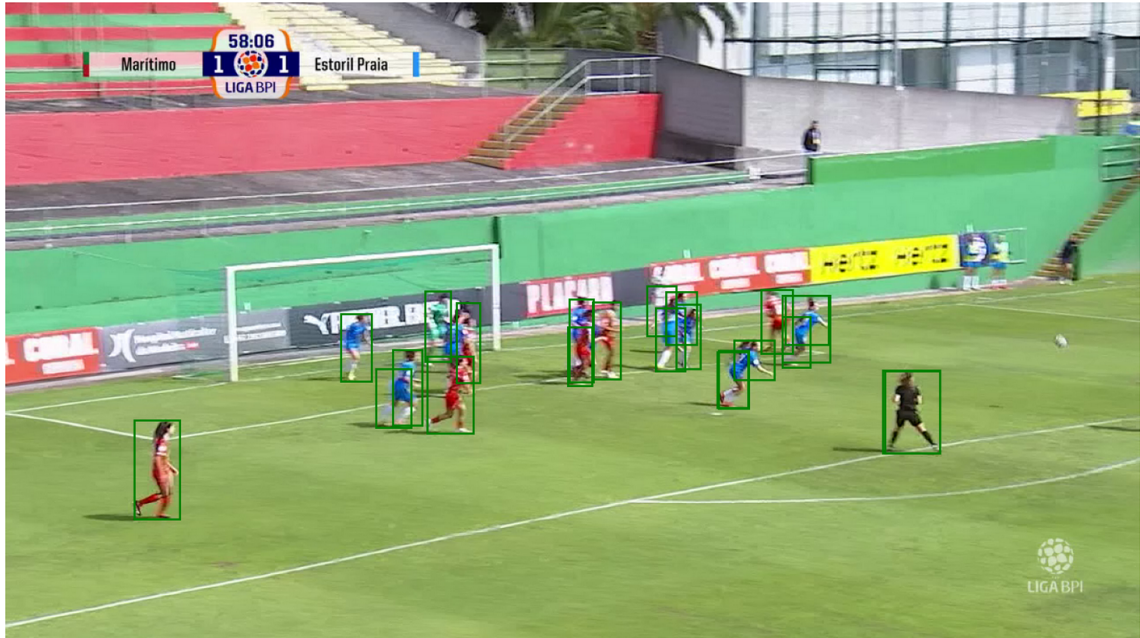


Figure 4.5: Player Detection Results YOLO

In the next step of segmenting the player given the bounding boxes of the players using the SAM model, there is an almost perfect segmentation of the players that we can see in figure 4.6.

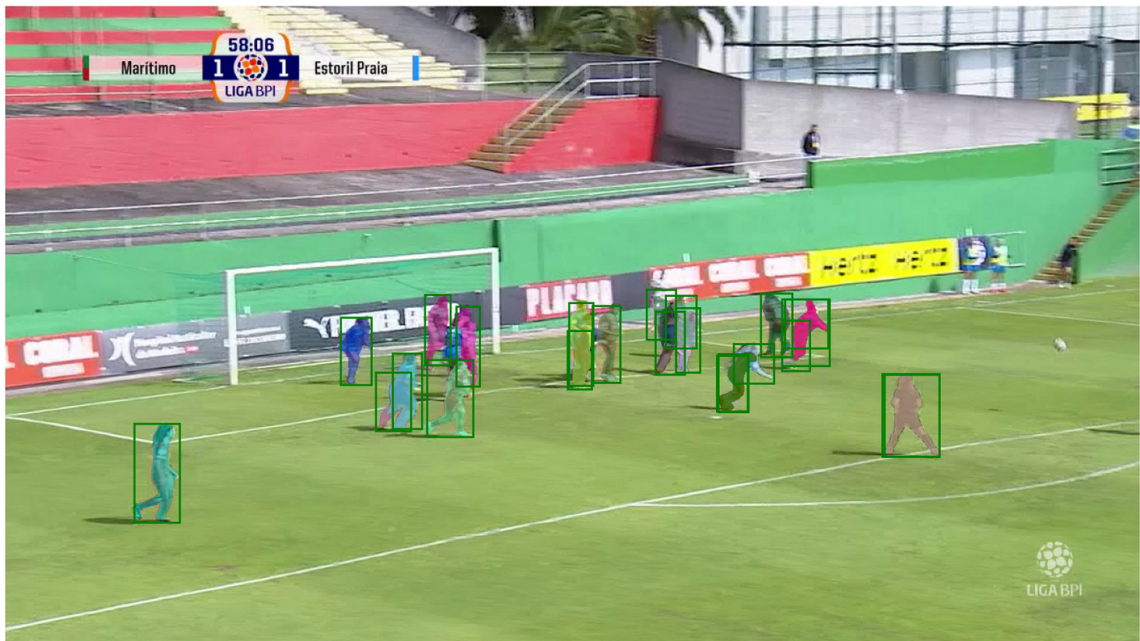


Figure 4.6: Player Detection Results YOLO + SAM

The detection of the goalpost using YOLO11 with oriented bounding boxes didn't achieve the results that we expected. The oriented bounded boxes did not work and we got a normal bounding box that gives us mixed results. It is able to detect the majority of the goalposts but it doesn't fit perfectly on top of the goalposts as we were expecting, see figure 4.7. This is probably a fault with the dataset that isn't perfectly annotated and includes goalposts of other sports.



Figure 4.7: GoalPost Detection Result YOLO11-obb

4.3 Goal Event Isolation Results

Frame subtraction is hindered by the fact that the camera in the moment of the goal is moving towards the left. This gives the whole scene some movement and, as a result the video has a lot more movement than it is ideal. The ball can be seen being shot in the right side of the figure 4.8.



Figure 4.8: Frames Subtracted Video

Using the video with subtracted frames and removing the players while adding the goalposts gives us the result present in the figure 4.9. While not perfect we can see an improvement from the original frames subtracted video. The goalposts are now very visible and the players are more hidden even if not completely gone.

The SAM model is able to achieve amazing results in the segmentation of the players but, in this case, when applied to our frame subtracted video, we can see that some of the motion is still visible. A less rigorous approach might work better for this case.

The goalposts we drew on the video aren't exactly in the right position and occupy a larger area than the real goalposts.

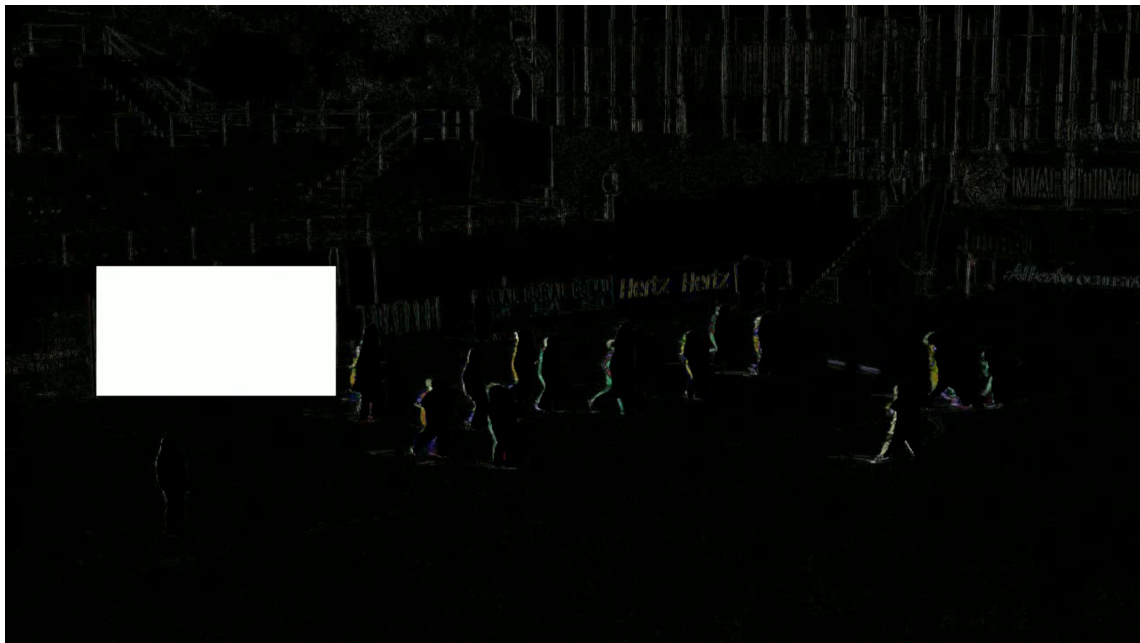


Figure 4.9: Frames Subtracted Video + Remove Players (YOLO+SAM) + Goalposts (YOLO)

In an attempt to improve on the previous results we decided to try a different way. Instead of using SAM to segment the players, we used only the YOLO model and erased everything inside the bounding box. Without the segmentation we now have a rougher detection of the players that is able to get better results in erasing all the movement seen from the players in the subtracted video. This does lead to a problem where when the ball passes through one of the bounding boxes surrounding a player, it also gets erased.

For the goalposts detection, this time, instead of only using the bounding box from the YOLO detection, we added the SAM model to see if it would be able to accurately segment the goalposts present inside the YOLO detection bounding box. In figure 4.10 we can see that the SAM model was very successful in segmenting the goalposts.

This new attempt at isolating the ball movement and the goalposts achieved better results, even if it did introduce some new problems such as the ball being erased if passing near a player.



Figure 4.10: Frames Subtracted Video + Remove Players (YOLO) Bounding Box + Goalposts segmentation (YOLO + SAM). The predicted goal posts are marked white and the ball with heavy motion blur appears with two trails

4.4 Goal Event Classification Results

After the isolation of the goalposts and the movement of the ball we trained the ViViT with 10 clips of Goal and 10 clips of Kick. This time, the results weren't as promising as with thee kick-off events. The precision of the Goal and Kick clips was 0.5, see figure 4.11.

	precision	recall	f1-score
Goal	0.50	0.67	0.57
Kick	0.50	0.33	0.40
KickOff	1.00	0.50	0.67
NormalPlay	0.67	1.00	0.80
accuracy			0.60
macro avg	0.67	0.62	0.61
weighted avg	0.63	0.60	0.58

Figure 4.11: Goal ViViT Results

As we can see in figure 4.12 and figure 4.13, the model is learning but, because the goal is a very varied and complex event, we would need a bigger dataset and a more extensive training to be able to achieve better results when making predictions in new data.

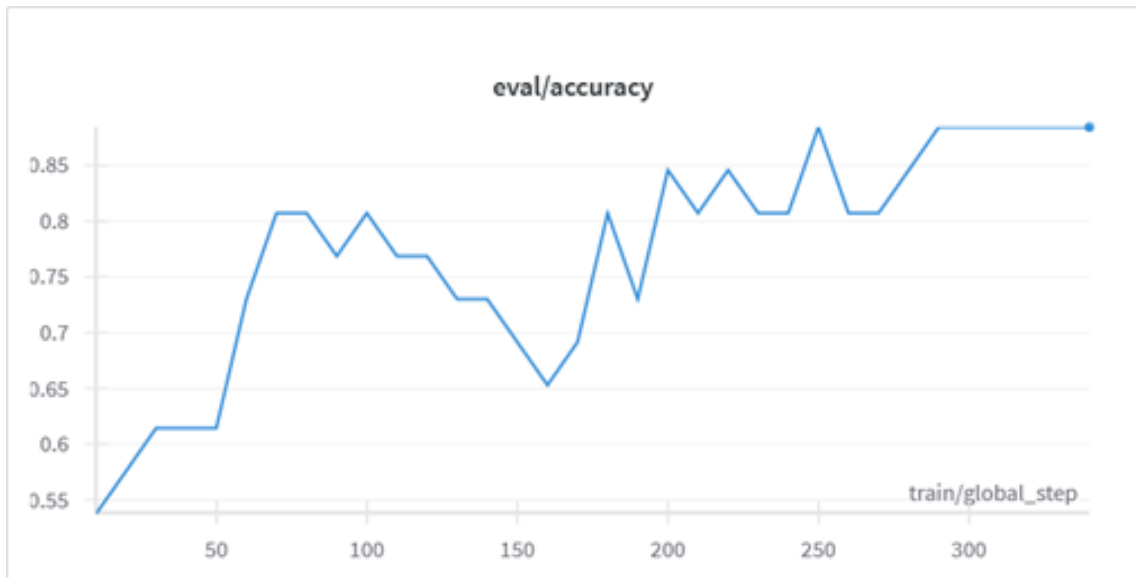


Figure 4.12: Goal Evaluation Accuracy

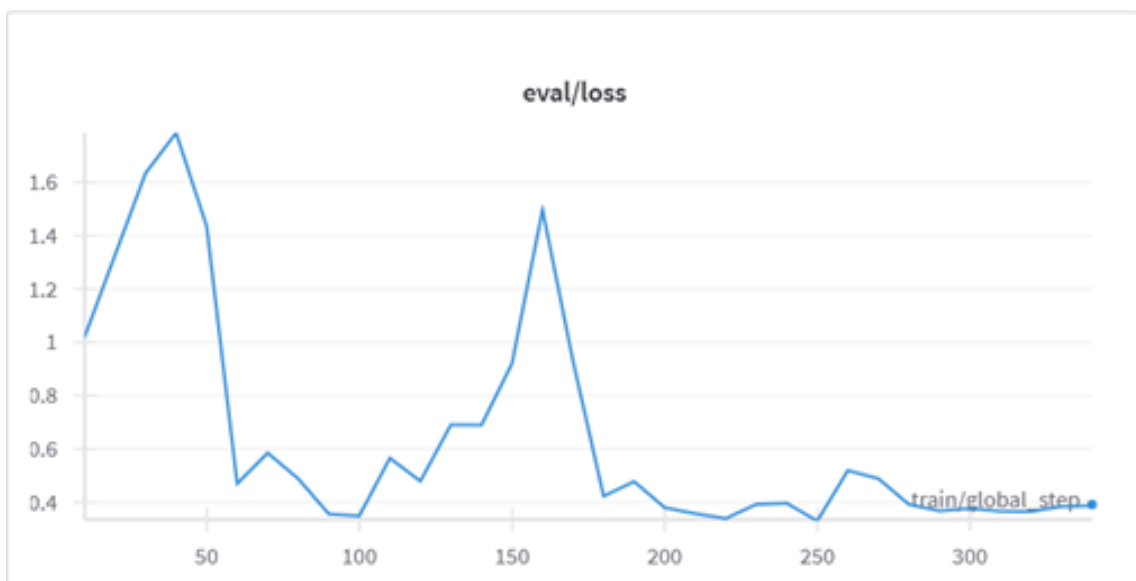


Figure 4.13: Goal Evaluation Loss

Chapter 5

Conclusion and Future Work

In this chapter we reflect on the research contributions, the limitations of the project, and the future work.

5.1 Researched contributions

This dissertation demonstrated the viability of using the ViViT model for football event detection. The results obtained in the kick-off and normal play event detection clearly show that the model is able to distinguish between similar events and identify them correctly in a football setting, on both men's and women's football.

It also introduced a novel way to try and isolate the ball and the goalposts in a football match video. With this attempt at helping the goal event be more clear we can try to make the detection of this event have better results.

The results obtained show the efficacy of ViViT for the detection of temporal events in a football match and the ability to create a simpler internal view of the goal event. This research contributes a practical and effective methodology for automated event detection, paving the way for future work in football analysis from raw broadcast footage.

The full proposed pipeline shown in Figure 3.1 is not yet fully implemented - the data fusion block is future work are underway.

5.2 Limitations

There were several limitations that were felt throughout the development of this project.

One of the main limitations is the machine capabilities. Training large datasets, especially when using videos with high resolutions, requires large computational resources that we don't have.

This restricts us to smaller datasets. Video clips were used instead of full matches to save on resources and make training the model quicker and lighter.

The dataset we used was very small because there is no dataset publicly available that consists of full matches of lower leagues. Most datasets focus on the first league which was not the objective of this project. In our dataset we also got some kick-off and goal events that are not shown in the broadcast which makes it harder to find video clips for them. Especially the kick-off event was harder to obtain because after a goal the broadcast normally goes into replays of the goal and it prevents the kick-off from being shown.

For the video subtraction method there are several limitations, with one of them being the fast camera movements that prevent static objects from being shown in the subtracted video. Another limitation is the angle of the camera sometimes not being ideal for detecting everything.

5.3 Future Work

We had some good results on the kick-off event detection so, for future work we can extend this model to predict other events that will allow for a more detailed summary of the statistics of the match besides the goals. There can also be detected which team scored the goal in order to have the correct final score. With a larger dataset there is also the potential to improve results and introduce a more varied dataset for the different events, allowing the model to be more effective in different scenarios.

Audio can also be used to help in the detection of events. For example, the whistle of the referee can give an additional confirmation of fouls or goals.

Several motion techniques can also be used to help reduce or remove the camera motion. This would be very helpful in achieving better results for the creation of the video with only the ball and the goalposts.

References

- [1] Anurag Arnab et al. *ViViT: A Video Vision Transformer*. 2021. arXiv: 2103.15691 [cs.CV]. URL: <https://arxiv.org/abs/2103.15691>.
- [2] Lukas Biewald. *Experiment Tracking with Weights and Biases*. Software available from wandb.com. 2020. URL: <https://www.wandb.com/>.
- [3] Joao Carreira and Andrew Zisserman. *Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset*. 2018. arXiv: 1705.07750 [cs.CV]. URL: <https://arxiv.org/abs/1705.07750>.
- [4] Gong Cheng et al. “Towards Large-Scale Small Object Detection: Survey and Benchmarks”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.11 (2023), pp. 13467–13488. DOI: 10.1109/TPAMI.2023.3290594.
- [5] Darkmyter. *Football-Players-Tracking*. URL: <https://github.com/Darkmyter/Football-Players-Tracking>.
- [6] Alexey Dosovitskiy et al. *An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale*. 2021. arXiv: 2010.11929 [cs.CV]. URL: <https://arxiv.org/abs/2010.11929>.
- [7] Mohammed Amine El Mrabet, Khalid El Makkaoui, and Ahmed Faize. “Supervised Machine Learning: A Survey”. In: *2021 4th International Conference on Advanced Communication Technologies and Networking (CommNet)*. 2021, pp. 1–10. DOI: 10.1109/CommNet52204.2021.9641998.
- [8] Na Feng et al. “SSET: A dataset for shot segmentation, event detection, player tracking in soccer videos”. In: *Multimedia Tools and Applications* 79.39–40 (Aug. 2020), pp. 28971–28992. DOI: 10.1007/s11042-020-09414-3.
- [9] Silvio Giancola et al. “SoccerNet: A Scalable Dataset for Action Spotting in Soccer Videos”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, June 2018. DOI: 10.1109/cvprw.2018.00223. URL: <http://dx.doi.org/10.1109/CVPRW.2018.00223>.
- [10] IBM. *What is Computer Vision?* Nov. 22, 2023. URL: <https://www.ibm.com/topics/computer-vision>.
- [11] Sagar Imambi, Kolla Bhanu Prakash, and G. R. Kanagachidambaresan. “PyTorch”. In: *Programming with TensorFlow: Solution for Edge Computing Applications*. Ed. by Kolla Bhanu Prakash and G. R. Kanagachidambaresan. Cham: Springer International Publishing, 2021, pp. 87–104. ISBN: 978-3-030-57077-4. DOI: 10.1007/978-3-030-57077-4_10. URL: https://doi.org/10.1007/978-3-030-57077-4_10.

- [12] Yudong Jiang et al. "SoccerDB: A Large-Scale Database for Comprehensive Video Understanding". In: *Proceedings of the 3rd International Workshop on Multimedia Content Analysis in Sports*. MM '20. ACM, Oct. 2020. DOI: [10.1145/3422844.3423051](https://doi.org/10.1145/3422844.3423051). URL: <http://dx.doi.org/10.1145/3422844.3423051>.
- [13] Will Kay et al. *The Kinetics Human Action Video Dataset*. 2017. arXiv: [1705.06950](https://arxiv.org/abs/1705.06950) [cs.CV]. URL: <https://arxiv.org/abs/1705.06950>.
- [14] John D. Kelleher. *Deep learning*. The MIT Press, 2019.
- [15] Rahima Khanam and Muhammad Hussain. *YOLOv11: An Overview of the Key Architectural Enhancements*. 2024. arXiv: [2410.17725](https://arxiv.org/abs/2410.17725) [cs.CV]. URL: <https://arxiv.org/abs/2410.17725>.
- [16] Alexander Kirillov et al. "Segment Anything". In: *arXiv:2304.02643* (2023).
- [17] Qian Li et al. "A Survey on Deep Learning Event Extraction: Approaches and Applications". In: *IEEE Transactions on Neural Networks and Learning Systems* 35.5 (2024), pp. 6301–6321. DOI: [10.1109/TNNLS.2022.3213168](https://doi.org/10.1109/TNNLS.2022.3213168).
- [18] Yixuan Li et al. *MultiSports: A Multi-Person Video Dataset of Spatio-Temporally Localized Sports Actions*. 2021. arXiv: [2105.07404](https://arxiv.org/abs/2105.07404) [cs.CV].
- [19] Shervin Minaee et al. "Image Segmentation Using Deep Learning: A Survey". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44.7 (2022), pp. 3523–3542. DOI: [10.1109/TPAMI.2021.3059968](https://doi.org/10.1109/TPAMI.2021.3059968).
- [20] Eduardo F. Morales and Hugo Jair Escalante. "Chapter 6 - A brief introduction to supervised, unsupervised, and reinforcement learning". In: *Biosignal Processing and Classification Using Computational Learning and Intelligence*. Ed. by Alejandro A. Torres-García et al. Academic Press, 2022, pp. 111–129. ISBN: 978-0-12-820125-1. DOI: <https://doi.org/10.1016/B978-0-12-820125-1.00017-8>. URL: <https://www.sciencedirect.com/science/article/pii/B9780128201251000178>.
- [21] Cherifa Nakkach, Amira Zrelli, and Tahar Ezzeddine. "Deep Learning Algorithms Enabling Event Detection: A Review". In: *Proceedings of the 2nd International Conference on Industry 4.0 and Artificial Intelligence (ICIAI 2021)*. Atlantis Press, 2022, pp. 170–175. ISBN: 978-94-6239-528-2. DOI: [10.2991/aisr.k.220201.030](https://doi.org/10.2991/aisr.k.220201.030). URL: <https://doi.org/10.2991/aisr.k.220201.030>.
- [22] NVIDIA. URL: <https://www.tensorflow.org/learn?hl=pt>.
- [23] NVIDIA. *What is tensorflow?* <https://www.nvidia.com/en-us/glossary/tensorflow/>.
- [24] Rafael Padilla, Sergio L. Netto, and Eduardo A. B. da Silva. "A Survey on Performance Metrics for Object-Detection Algorithms". In: *2020 International Conference on Systems, Signals and Image Processing (IWSSIP)*. 2020, pp. 237–242. DOI: [10.1109/IWSSIP48289.2020.9145130](https://doi.org/10.1109/IWSSIP48289.2020.9145130).
- [25] Joseph Redmon et al. *You Only Look Once: Unified, Real-Time Object Detection*. 2016. arXiv: [1506.02640](https://arxiv.org/abs/1506.02640) [cs.CV]. URL: <https://arxiv.org/abs/1506.02640>.
- [26] Pramila P. Shinde and Seema Shah. "A Review of Machine Learning and Deep Learning Applications". In: *2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA)*. 2018, pp. 1–6. DOI: [10.1109/ICCUBEA.2018.8697857](https://doi.org/10.1109/ICCUBEA.2018.8697857).

- [27] Karen Simonyan and Andrew Zisserman. *Two-Stream Convolutional Networks for Action Recognition in Videos*. 2014. arXiv: 1406.2199 [cs.CV]. URL: <https://arxiv.org/abs/1406.2199>.
- [28] Alessandro Suglia et al. *Going for GOAL: A Resource for Grounded Football Commentaries*. 2022. arXiv: 2211.04534 [cs.CV].
- [29] Tensorflow. URL: <https://www.tensorflow.org/guide/keras>.
- [30] Du Tran et al. *Learning Spatiotemporal Features with 3D Convolutional Networks*. 2015. arXiv: 1412.0767 [cs.CV]. URL: <https://arxiv.org/abs/1412.0767>.
- [31] Dirk Valkenborg et al. “Unsupervised learning”. In: *American Journal of Orthodontics and Dentofacial Orthopedics* 163.6 (June 2023), pp. 877–882. ISSN: 0889-5406. DOI: 10.1016/j.ajodo.2023.04.001. URL: <http://dx.doi.org/10.1016/j.ajodo.2023.04.001>.
- [32] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL]. URL: <https://arxiv.org/abs/1706.03762>.
- [33] Vanna Winland. *What is self-attention?* Mar. 2025. URL: <https://www.ibm.com/think/topics/self-attention>.
- [34] Gui-Song Xia et al. “DOTA: A Large-Scale Dataset for Object Detection in Aerial Images”. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2018.
- [35] Shuyuan Xu et al. “Computer Vision Techniques in Construction: A critical review”. In: *Archives of Computational Methods in Engineering* 28.5 (2020). DOI: 10.1007/s11831-020-09504-3.
- [36] yolo. *goalpost Dataset*. <https://universe.roboflow.com/yolo-b6voi/goalpost-u6e0h>. Open Source Dataset. Nov. 2024. URL: <https://universe.roboflow.com/yolo-b6voi/goalpost-u6e0h>.
- [37] ZHI-HUA ZHOU. *Machine learning*. SPRINGER VERLAG, SINGAPOR, 2022.