VARS: Video Assistant Referee System for Automated Soccer Decision Making from Multiple Views

Jan Held*,1 Anthony Cioppa*,1,2 Silvio Giancola*,2
Abdullah Hamdi[2] Bernard Ghanem[2] Marc Van Droogenbroeck[1]
[1] University of Liège [2] KAUST

Abstract

The Video Assistant Referee (VAR) has revolutionized association football, enabling referees to review incidents on the pitch, make informed decisions, and ensure fairness. However, due to the lack of referees in many countries and the high cost of the VAR infrastructure, only professional leagues can benefit from it. In this paper, we propose a Video Assistant Referee System (VARS) *that can automate soccer decision-making. VARS leverages the latest findings in multi-view video analysis, to provide real-time feedback to the referee, and help them make informed decisions that can impact the outcome of a game. To validate VARS, we introduce* SoccerNet-MVFoul*, a novel video dataset of soccer fouls from multiple camera views, annotated with extensive foul descriptions by a professional soccer referee, and we benchmark our VARS to automatically recognize the characteristics of these fouls. We believe that VARS has the potential to revolutionize soccer refereeing and take the game to new heights of fairness and accuracy across all levels of professional and amateur federations.*

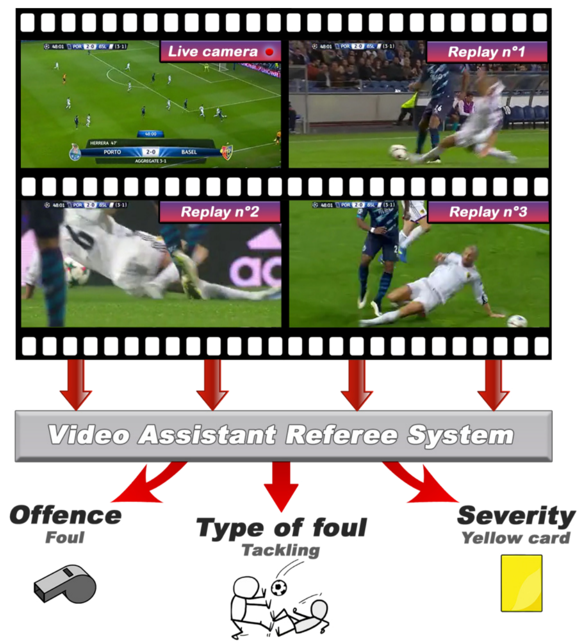

Figure 1. **Video Assistant Referee System (VARS).** We propose an automated VARS for automatically classifying whether an action is a foul, determining the type of foul (*e.g*., 'Tackling', 'Pushing', etc.), and the appropriate punishment the player should receive for the foul (*i.e*., 'No card', 'Yellow card', or 'Red card') from a multi-view camera setup.

1. Introduction

Over the past decades, the technology used by referees in soccer has undergone a drastic evolution. Before the beginning of this century, referees and their assistants only relied on their own judgment, and the communication between them was based on eye contact and body language. The French and Scottish refereeing trios were the first ones to be linked through wireless mini-earphones during league matches, facilitating communication among them [56]. Nowadays, wireless headsets are essential pieces of equipment for referees worldwide, made mandatory for high-level competitions. Another important breakthrough in professional soccer was the introduction of goal-line technology, which uses a combination of cameras and sensors to determine whether the entire ball has crossed the goal line or not. This technology aims to prevent controversial goals such as the famous "ghost goal" scored by Geoff Hursts in the 1966 World Cup final against Germany, where the ball may not have fully crossed the line and led to England receiving the world champion title [17]. Successively, the International Football Association Board (IFAB) approved the introduction of extra referees, namely the video assistant referees (VAR), to prevent game-changing errors. More recently, artificial intelligence systems appeared for the first time during the World Cup 2022 in Qatar. Semi-

(*) Denotes equal contributions
Contacts: jan.held@student.uliege.be, silvio.giancola@kaust.edu.sa, anthony.cioppa@uliege.be. Data/code available at www.soccer-net.org.

automated offside technology now supports the VAR to help referees make faster, more accurate, and more reproducible offside decisions [18]. This new system relies on 12 well-calibrated cameras to track the ball and the player's body pose, sending an automated offside alert to the video assistant referee inside the video operation room. This shows that soccer is moving towards more assistance or even automated systems to help referees make better decisions.

However, despite its intention to improve the accuracy of referee decisions, VAR has become a source of frustration and anger for many football fans around the world. Since we have a different video assistant referee for each game, we do not always have consistent decisions. Sometimes the VAR predicts different outcomes for similar situations in different games and leagues. Moreover, the implementation of the VAR technology and infrastructure requires a substantial financial investment, limiting its accessibility to only the top-tier leagues and clubs. As a result, semi-professional or amateur leagues are unable to benefit from the VAR due to financial constraints. Additionally, the shortage of referees worldwide makes it impossible in staffing additional referees as Video Assistant Referees, except for the professional leagues.

In this work, we propose a first step towards a fully automated "Video Assistant Referee System" (VARS) which could support or replace the current VAR. We attempt to automatically predict all fouls and suggest appropriate sanctions to the players. In case the on-field referee makes a significant mistake, our VARS could intervene to suggest a revision. It is intended that, just like regular VAR, our VARS serves as a support system for the referee, but the final decision remains in the hands of the on-field referee. To achieve this objective, we rely on multi-view uncalibrated camera video streams, which are already leveraged to edit broadcast games. Specifically, we release a new dataset comprising 3,901 actions with multi-view clips of 5 seconds around the action, annotated by a professional referee. We focus our analysis on the classification of foul types and evaluate their severity to identify the sanction for the player. Practically, our VARS analyses the different streams and combines the information from the multiple cameras. We show that using a multi-view system largely improves the performance compared to a single view and that we reach good performance on our video recognition tasks.

Contributions. We summarize our contribution as follows: **(i)** We publicly release *SoccerNet-MVFouls*, a new multi-view video dataset containing video clips captured by multiple cameras, annotated with 10 properties. **(ii)** We propose *VARS*, a new multi-camera video recognition system for classifying fouls and their severity. **(iii)** We propose a thorough study of using multiple views and how different types of camera views can influence the performance of VARS on two new video recognition tasks.

2. Related work

Sports understanding. As a research topic, sports video understanding has increased in popularity thanks to its challenging and fine-grained nature [41, 54]. Nowadays, most state-of-the-art automatic methods are based on deep learning and have shown impressive performance on tasks such as player detection and tracking [7, 39, 58], tactics analysis [53], pass feasibility [2] and prediction in soccer [25], talent scouting [11], or player re-identification in occluded scenarios [50]. Video classification started as a key area of research in this field [65], with approaches proposed to recognize specific actions [32, 45] or distinguish between different game phases [8]. With the growing interest in temporal activity localization [3], the task of action spotting [4, 6, 10, 26, 48, 49, 67] has gained interest as it provides precise localization of specific actions within a soccer game.

The progress in those tasks was made possible thanks to the availability of large-scale datasets [28,43,46,57,66]. Giancola *et al*. [20] introduced the SoccerNet dataset, which has grown to be the most extensive collection of data and annotations for video understanding in soccer, including benchmarks for 10 different tasks, ranging from broadcast understanding [12], field understanding [5] and player understanding [9]. The SoccerNet team also organizes yearly competitions on these different tasks to foster research in the field [21]. The dataset presented in this paper extends SoccerNet by proposing a novel multi-view video collection including foul annotations for video recognition tasks.

Video understanding. For a long time, video understanding lagged behind image understanding due to the lack of large-scale video datasets such as ImageNet or CIFAR-100 [13, 34] in the video domain. However, the release of large video understanding datasets such as UCF101 [51], ActivityNet [3], YouTube-8M [1], and Kinetics [31] has led to a surge in popularity and interest in the field. Video understanding tasks include video classification [16, 30, 42], action recognition [47, 61], video captioning [19, 33, 63], and video generation [36].

The interest in developing video classification models that capture spatio-temporal information has significantly grown. Temporal Segment Network (TSN) [62] aggregates features across multiple temporal video segments to improve recognition performance. Tran *et al*. [55] proposed a new spatio-temporal convolutional block R(2+1)D and analyze its effect on action recognition models. Recently, the Multiscale Vision Transformer (MViT) [15, 37] came as a way to combine the strengths of both convolutional neural networks (CNNs) and transformers for video classification, capturing both spatial and temporal attentions. In this work, we train different video representations to learn per-clip features that we aggregate from multiple views to identify the different properties of the fouls.

Dataset	Context	Task	Videos/Images	View	Type
Kinetics 400 [31]	Human actions	Classification	650,000	Single-view	Videos
NTU RGB+D 120 [38]	Human actions	Classification	114,480	Multi-view	Videos
Northwestern-UCLA Multiview [60]	Human actions	Classification	1,493	Multi-view	Videos
UWA3D Multiview II [44]	Human actions	Classification	900	Multi-view	Videos
SoccerNet-v2 (Actions) [21]	Soccer	Classification	110,458	Single-View	Videos
SoccerNet-v3 (Re-id.) [5]	Soccer	Re-identifcation	33,986	Multi-View	Images
SoccerNet-MVFouls (Ours)	Soccer (Fouls)	Classification	8,923	Multi-view	Videos

Table 1. **Video action understanding datasets**. Comparative overview of relevant datasets for multi/single-view action recognition in videos/images. Our dataset is the only one providing multi-view videos for classification in sports, with 10 annotated properties per action.

Multi-view understanding. Su *et al.* [52] introduces the idea of training image encoders to recognize 3D objects from multiple views, benefiting from the mature 2D computer vision. Most effort focused on informative aggregation between views, introducing cross-view confidence [29], group convolutions to learn rotation-equivariant representations [14], graph convolutions to learn view aggregation [64]. Alternatively, MVTN [23] predicted the viewpoints from a differentiable 3D renderer. In the video domain, synthetic views (*e.g.* 3D motion or optical flow) are created for single-stream videos as a way to obtain better representation learned in a self-supervised fashion [35, 59]. In this work, we leverage a simple multi-view pipeline for video understanding, trained in a fully supervised fashion, that incorporates multiple replay streams from soccer broadcast videos.

3. SoccerNet-MVFouls dataset

In this section, we introduce our novel multi-view foul classification soccer dataset, called *SoccerNet-MVFouls*.

Table 1 presents an overview of our dataset and compares it with other datasets that propose action recognition using either single or multiple views. Our dataset is the only one for multi-view video action recognition in sports, and the first dataset to focus specifically on referee's decisions.

SoccerNet-MVFouls gathers 3,901 actions extracted from 500 soccer games from six main European leagues, covering three seasons from 2014 to 2017, extracted from the SoccerNet dataset [12, 20]. Each action is composed of at least two videos depicting the live action and at least one replay. The actions are annotated with 10 different properties describing the characteristics of the foul from a referee's perspective (*e.g.* the severity of the foul, the type of foul, *etc.*). To ensure high-quality annotations, all these properties were manually annotated by a professional soccer referee with 6 years of experience and more than 300 official games. The referee watched the videos from all available views at any speed to accurately characterize the foul.

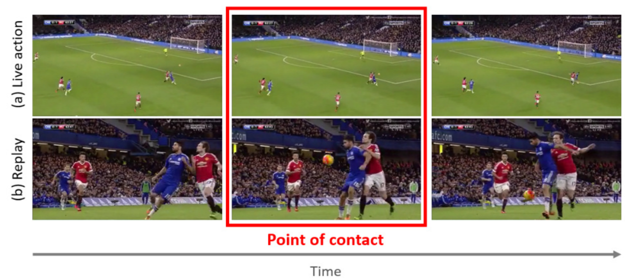

Figure 2. **Example of a multi-view sequence from our dataset.** Each foul has at least (a) one live-action clip (usually taken from the main camera) and (b) one synchronized replay clip (usually a close-up view). We annotate the exact frame where the point of contact happens (red box). The ground-truth properties for this example are: "Offence", "Challenge", "No card", "With contact", "Upper body", "Use of shoulder", "Ball is not played", "Tried to play the ball", "No handball", and "No handball offence".

3.1. Dataset collection

The dataset was collected in three steps: (i) we extracted the relevant action clips from soccer broadcast videos, (ii) we temporally aligned the clips related to the same action, and (iii) we annotated several foul properties.

Clip extraction. As a starting point, we used the SoccerNet-v2 dataset [12], which contains timestamp annotations of fouls for 500 full broadcast games. Furthermore, the SoccerNet dataset also provides annotations of the replays of some of the fouls, allowing us to retrieve, for the same action, different viewpoints. Since our goal is to design a multi-view video assistant referee system, we only keep actions for which we have access to at least two different points of view. In most cases, the extracted clips should cover sufficient information to determine all the foul properties. Also, to prevent bias towards the on-field referee's decision, the 5 second clips should not contain the decision of the referee (*e.g.* if the player is given a yellow card). Therefore, we extracted 5 second clips per action, starting 3 seconds before and ending 2 seconds after the timestamp annotation. In the following, "live action clip" will refer to the clips taken from the main camera, while "replay clips" will denote all the replay clips typically taken from closer shots. Figure 2 shows an example of such extracted clip.

Clips alignment. We build a Multi-View Foul Annotator tool with a similar interface to a VAR room to ensure the quality of the annotations performed by our referee. At first, the referee is presented with all available clips of an action simultaneously on a grid layout. Our annotator tool enables users to modify the annotated point of contact (see Figure 2) for each clip individually and adjust the speed and offset of the clips to align them temporally, taking into account the fact that replays are frequently broadcasted at a slower speed. The referee may browse simultaneously the synchronized videos either at regular speed or frame by frame to accurately understand and describe the properties of the action. More information and an example of annotation using our annotator may be found in the supplementary material.

Property annotations. The SoccerNet-v2 dataset provides annotations for fouls and yellow/red cards given by the actual game referee. However, the on-field referee has only his own point of view to characterize the foul. Judging foul play incidents from the referee's position at playing time leads to an average error rate of 14% [40]. Our referee annotator has no time pressure and access to multiple perspectives, which results in more accurate decisions compared to the on-field referee who has to take a quick decision and only has a single view. To ensure a high-quality dataset and avoid any bias, our professional soccer referee manually annotated all properties without seeing the on-field referee's decision.

We defined several properties for each action that are necessary for the referee to take the final decision. These properties include (i) if the clip contains a foul (*i.e.* an action which breaks/violates the *Laws of the Game* [27]), (ii) the class of the foul, (iii) the severity of the foul, (iv) if the player plays the ball, (v) if the player tries to play the ball, (vi) if any player touches the ball with his hand or arm, deliberately or not, (vii) and whether it is an offence according to the *Laws of the Game* [27]), (viii) if there is contact between two players, (ix) the action foul relates to the upper or underbody, and, finally, (x) we further discriminate for the upper body between arms and shoulders. We have special labels corresponding to grey areas for the property (i), we use the label "Between" when both "Foul" and "No foul" decisions are equally valid and there is no obvious decision. For property (iii), we use the labels "Borderline No card/Yellow card" and "Borderline Yellow card/Red card" to indicate a grey area when either "No card" or "Yellow card', (resp. "Yellow card" or "Red card"), would be the correct decision.

Fouls		Offence		Offence Severity	
Class	**Prob.**	**Class**	**Prob.**	**Class**	**Prob.**
St. tackling	43.6	Offence	85.8	No card	55.3
Tackling	15.6	No offence	10.7	Yellow card	26.6
Challenge	13.0	Between	3.4	NC/YC	15.2
Holding	12.5			YC/RD	1.7
Elbowing	5.9			Red card	1.1
High leg	3.5				
Pushing	2.9				
Dive	0.9				

Table 2. **Distribution of classes in our SoccerNet-MVFouls dataset.** The distribution of the classes for the "Offence", "Severity" and "Type of foul" properties is highly imbalanced. The distribution for the other properties is shown in supplementary. "St. tackling" stands for standing tackling.

Foul Class	Succ. Rate	Severity Distribution		
		No Card	🟨 Card	🟥 Card
Standing Tackling	**0.94**	0.79	0.18	0.02
Tackling	0.87	0.37	**0.58**	**0.04**
High Leg	0.87	0.31	**0.63**	**0.06**
Holding	0.90	0.60	0.40	0.00
Pushing	0.84	**0.99**	0.01	0.00
Elbowing	**0.93**	0.43	**0.53**	0.03
Challenge	0.75	**0.94**	0.05	0.01
Dive	/	0.00	1.00	0.00

Table 3. **Referees success rate and severity per foul class.** Referees are successful in most classes but struggle with "Challenge". Some classes are less likely to return a card, *e.g.* "Tackling" or "High leg". The success rate for "Dive" is unknown, as we cannot know if a referee whistle for the foul or the dive.

3.2. Dataset statistics

Number of views. On average, we have 2.29 clips per foul action, around 75% of them have two viewpoints (live and replay), 20% have a second replay, and around 5% have a third replay video. No foul has more than four views.

Properties distribution. Table 2 shows the distribution of the properties "Offence", "Severity" and "Type of foul'. We can see that in all three cases, the distribution is highly unbalanced towards "No card", "Offence" and "Standing tackling", respectively. This analysis follows our intuition of soccer, where yellow and red cards are usually rarer than simple free-kicks given after a foul.

Success rate of the referees. As we only have extracted fouls for which the on-field referee has given a foul in the game, we can analyze the success rate of the referees by analyzing the property "No offence". From the 3901 fouls given by the referees in the games, our referee annotated 368 fouls as "No offence", leading to an error rate of 10.7%. "Standing tackling" and "Elbowing" are the most well classified with 94% success rate, as shown in Table 3. For the

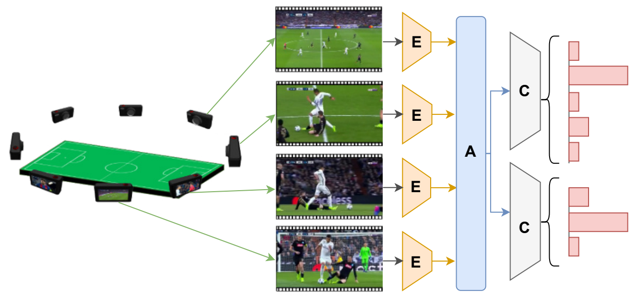

Figure 3. **VARS: Video Assistant Referee System.** From multi-view video clips input, our system encodes per-view video features (**E**), aggregates the view features (**A**), and classifies different properties of the foul action (**C**).

remaining action classes, the referees have a similar success rate of approximately 87%, except for the foul class "Challenge", where the referees have an error rate of 25%. Our analysis is aligned with the finding of Mallo *et al.* [40].

Severity for different foul classes. The distribution of the severity among different foul classes can provide insight into how often certain types of fouls result in a card. The results are presented in Table 3. "Tackling", "High leg", and "Elbowing" are three types of fouls that very often result in a yellow card, as they represent fouls that are dangerous for opponents. Contrarily, some classes like "Pushing" or "Challenge" are very unlikely to get a yellow or red card.

4. Methodology

Our VAR system is a multi-view video architecture, that automatically identifies different properties for an action. We illustrate our proposed VARS in Figure 3.

4.1. Classification tasks

We formally define two tasks for our dataset.

Task 1: Fine-grained foul classification. Given multiple clips of the same foul instance, the objective is to classify the foul into one of 8 fine-grained foul classes: "Standing tackling", "Tackling", "High leg", "Pushing", "Holding", "Elbowing', "Challenge", "Dive/Simulation".

Task 2: Offence severity classification. Given multiple clips of the same foul instance, the objective is to classify whether the foul constitutes an offence, as well as the severity of the foul. We have defined four classes: "No offence", "Offence + No card", "Offence + Yellow card", and "Offence + Red card". We put aside clips labeled "Between" as well as the clips annotated as "Borderline". Therefore, for this particular task, we use a subset of our SoccerNet-MVFoul dataset.

4.2. Video Assistant Referee System (VARS)

We propose a novel Video Assistant Referee System (VARS) for the task of video recognition from multiple camera views. The pipeline of the VARS is presented in Figure 3. Our VARS takes multiple video clips denoted by $\mathbf{v} = \{v_i\}_1^n$ as input, showing the same action from n different views. A video v_i is fed into a video encoder $\mathbf{E}$ with parameters θ_E to extract a vector f_i containing the spatio-temporal features for that specific view:

$$f_i = \mathbf{E}_{\theta_\mathbf{E}}(v_i) \ . \tag{1}$$

We aggregate the feature vectors through a function $\mathbf{A}$ that outputs a single multi-view representation $\mathbf{R}$ following:

$$\mathbf{R} = \mathbf{A}(\{f_i\}_{i=1}^n) \ , \tag{2}$$

with $\mathbf{A}$ being a max or mean aggregation function. For the single-task classifier, we input the pooled features through a classification head $\mathbf{C}$ with parameters θ_C. VARS predicts the final class from the maximum probability score of the classification head, as given by:

$$\mathbf{VARS} = \arg\max \mathbf{C}_{\theta_\mathbf{C}}(\mathbf{R}) \ . \tag{3}$$

We train our model to minimize the following loss:

$$\mathcal{L} = \mathbf{L}(\mathbf{C}_{\theta_\mathbf{C}}\ (\mathbf{A}(\{\mathbf{E}_{\theta_\mathbf{E}}(v_i)\}_{i=1}^n)), y) \ , \tag{4}$$

with $\mathbf{L}$ being the cross entropy loss function, and y the ground truth associated to $\{v_i\}_{i=1}^n$. For the offence severity classification, VARS has to understand the game of soccer in order to correctly classify fouls into "No card", "Yellow card", and "Red card". In fact, bringing contextual information about the type of foul inside the network is essential to determine the severity of the offence. As the foul and offence classifiers share common features, we train a model to perform both tasks simultaneously. Our multi-task VARS learns to leverage these shared features to improve its predictions for both tasks. For the multi-task classifier, we define two heads, $\mathbf{C}^{\mathbf{foul}}$ and $\mathbf{C}^{\mathbf{off}}$, respectively for the tasks of fine-grained foul classification and offence severity classification. From the probability vector of each task, the VARS will take the maximum as the final prediction:

$$\mathbf{VARS}^\mathbf{t} = \arg\max \mathbf{C}^\mathbf{t}_{\theta_{\mathbf{C}^\mathbf{t}}}(\mathbf{R}) \quad \forall \mathbf{t} \in \{\text{foul}, \text{off}\} \ . \tag{5}$$

We train our model by minimizing both tasks loss with:

$$\alpha_{\text{foul}} \mathcal{L}^{\text{foul}} + \alpha_{\text{off}} \mathcal{L}^{\text{off}} \ . \tag{6}$$

By choosing different values for α, we can assign more or less importance to tasks. This scaling is necessary when the losses have significantly different magnitudes. In the case of our two tasks, the losses have a similar order of magnitude, so we typically select $\alpha_{foul} = \alpha_{off} = 1$.

Video encoder E. We considered different encoders to extract features from the video clips: **(i)** ResNet [24] may be used on videos by running the network on each frame independently and then using a max or mean pooling operation on the features across the frames to obtain a single feature vector that represents the entire video. While this approach works well for extracting spatial features, it does not capture temporal dynamics. **(ii)** R(2+1)D [55] extends the 2D CNN architecture with an additional temporal convolutional layer that operates on a sequence of frames to capture the temporal dynamics of the video. The advantage compared to ResNet is that it both captures spatial and temporal features directly. **(iii)** MViT [15, 37] integrates a multiscale feature representation with a transformer-based architecture to capture both spatial and temporal information from video clips. The feature encoders are typically pre-trained on ImageNet [13] (ResNet) or Kinetics [31] (R2+1D and MViT).

Multi-view aggregator A. To combine the extracted features from multiple views, we introduce two different pooling strategies [22], in particular: **(i)** Mean pooling takes the average value for each feature, and **(ii)** Max pooling which takes the maximum value per feature.

Classification heads C. Our classification heads consist of two dense layers with softmax activation. The output is a probability vector with dimensions that match the number of classes in the classification problems.

5. Experiments

5.1. Experimental setup

Training details. For both classification tasks, we leverage clips of 16 frames, spanning temporally for 1 second, with a spatial dimension of 224×398 pixels. Specifically, the clips contain 8 frames before the foul and 8 frames after the foul. The encoders **E** are pre-trained as detailed in the methodology, and the classifier **C** is trained from scratch, while both are trained in an end-to-end fashion. We use a cross-entropy loss, optimized with Adam with an exponential decreasing learning rate starting at 10^{-4} and a batch size of 8. The model starts overfitting after 10 epochs, and it takes around 9 hours to train on a single Nvidia V100 GPU.

Evaluation metrics. We report the classification accuracy, which is defined as the ratio of actions correctly classified with respect to the total number of actions. We also provide the top-2 accuracy (where a sample is considered well classified if the class appears in the top two highest confidence predictions) to get more insight into the model's performance. As our dataset is unbalanced, we also provide the balanced accuracy, which is defined as follows:

$$\text{Balanced Accuracy (BA)} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{P_i}\ , \qquad (7)$$

Feature Extractor	Pooling	Acc. @1	Acc. @2	BA
ResNet [24]	Mean	0.31	0.56	0.28
ResNet [24]	Max	0.32	0.60	0.28
R(2+1)D [55]	Mean	0.31	0.55	0.34
R(2+1)D [55]	Max	0.32	0.56	0.33
MViT [15, 37]	Mean	0.40	0.65	**0.45**
MViT [15, 37]	Max	**0.47**	**0.69**	0.43

Table 4. **Main results for the multi-view video foul classification.** We compare three feature encoders and two pooling methods. The best performance is obtained with MViT and a max pooling between the views. BA indicates the balanced accuracy after normalizing by the frequency of that class.

with N the number of classes, TP (True Positives) is the number of times where the model correctly predicted the class i and P_i (Positives) is the number of ground-truth samples for that class in the dataset.

5.2. Main Results

Task 1: Fine-grained foul classification. Our results may be found in Table 4. By extracting spatio-temporal features with MViT, we achieve significant improvements in performance compared to ResNet and R(2+1)D. This indicates that using a more advanced feature encoder can significantly enhance the model's ability to identify and classify the type of foul. The influence of the pooling method on the performance is however not significant, although max pooling shows slightly better results. In general, max pooling might be better when not all views are equally informative. Taking the max values helps identify the most important features for the most informative views while ignoring less useful information. In contrast, mean-pooling takes into account the information from all views, including those with a poor perspective. Overall, the best performance is obtained by using MViT as video encoder and max pooling.

Task 2: Offence severity classification. For the offence severity classification, we study the same feature encoders and pooling techniques. The top part of Table 5 shows the results obtained by our single-task classifier. Regardless of the used feature extractor or pooling technique, the model has more difficulties in classifying the actions. These difficulties are mainly due to two factors. First, the dataset exclusively consists of actions that were awarded a free kick by the on-field referee. As a result, the "No offence" actions are visually similar to a foul, and not to clear "No offence" actions. The model often struggles to differentiate these actions from actual fouls, which can be further seen in the Supplementary Material. Secondly, the visual appearance of an offence with no card, yellow card, or red card can vary greatly. In Figures 4a and 4b, we compare two frames of two different foul classes that have a little visual similarity. However, in both cases, the defender acted with disregard

Feature Extractor	Pooling	Task	Acc.	BA
ResNet [24]	Mean	Single	0.25	0.26
ResNet [24]	Max	Single	0.22	0.25
R(2+1)D [55]	Mean	Single	0.28	0.30
R(2+1)D [55]	Max	Single	0.27	0.29
MViT [15, 37]	Mean	Single	0.32	0.23
MViT [15, 37]	Max	Single	0.29	0.27
ResNet [24]	Mean	Multi	0.34	0.25
ResNet [24]	Mean	Multi	0.32	0.24
R(2+1)D [55]	Mean	Multi	0.34	0.30
R(2+1)D [55]	Max	Multi	0.39	0.31
MViT [15, 37]	Mean	Multi	0.38	0.31
MViT [15, 37]	Max	Multi	**0.43**	**0.34**

Table 5. **Multi-view video offence and severity classification**. We evaluate our VARS with different feature encoders and pooling methods on a single and multi-task setup. BA stands for the balanced accuracy.

(a) Elbowing - Yellow card

(b) Tackling - Yellow card

(c) Standing Tackling - No card

(d) Standing Tackling - Red card

Figure 4. **Example of fouls.** (a) The defender uses his arm as a tool to gain an unfair advantage and ignores the potential danger for his opponent. (b) The defender makes a tackle while taking the risk of his opponent being injured. (c) The defender tries to play the ball in no dangerous way. (d) The defender has no intention to play the ball and only aims to harm his opponent.

for the safety of his opponent and therefore resulting in a yellow card. In contrast, Figures 4c and 4d depict fouls that are visually more similar than the previous two fouls, yet one resulted in "No card" while the other resulted in a "Red card". Minor differences such as the point of contact, the speed of the foul, the distance to the ball, and the intention to play the ball or not, can lead to different classifications.

Multi-task classifier. Training a multi-task classifier on related tasks allows the model to utilize the learned information from one task to improve the performance on other tasks. In the bottom part of Table 5, we can see that the multi-task classifier outperforms the single-task classifier regardless of the feature encoder or pooling technique for offence severity classification. Using ResNet to extract spatial features for the type of foul and the offence severity classification does not perform well for either task. The body movements over time and the speed of the players involved in an action are important factors that can greatly impact the outcome of the classification. The multi-task classifier combined with MViT as encoder and max pooling shows promising results in classifying actions into their corresponding offence severity class. Furthermore, the multi-task classifier shows similar results as obtained for the single-task type of foul classification.

Performance	Viewing Setup				
	L	R1	L+R1	R1+R2	L+R1+R2
Acc_{T1}	0.31	0.47	0.50	0.56	**0.57**
$\text{Acc}_{T1}@2$	0.54	0.68	0.70	0.69	**0.72**
BA_{T1}	0.29	0.38	0.36	**0.44**	0.39
Acc_{T2}	0.38	0.39	**0.43**	0.39	0.40
$\text{Acc}_{T2}@2$	0.67	0.70	0.72	0.73	**0.75**
BA_{T2}	0.38	0.27	0.34	0.27	**0.39**

Table 6. **Single *vs*. multi-view classification**. We compare the performance for single vs multi-views and the influence of the type of view (Live L and replay R). We use MViT [15, 37] as feature extractor and max pooling. For both tasks, the best performance is mostly obtained with all three views. BA stands for balanced accuray, T1 stands for task 1 (foul classification) and T2 stands for task 2 (offence severity classification).

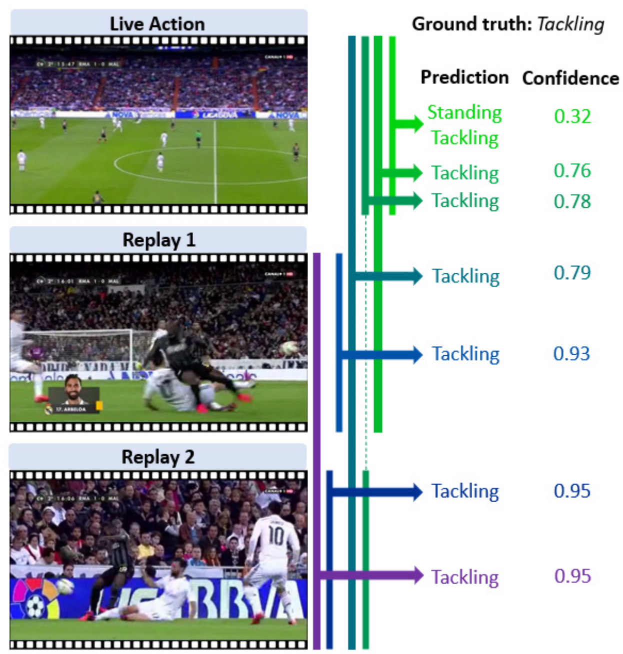

Figure 5. **Qualitative results.** VARS predictions for different combinations of views as input.The best performance is obtained with the two replay views.

5.3. Detailed analysis

Single *vs*. multi-view analysis. We now study the improvement of using multiple views over a single view. To do so, we first created a subset of the test set for which we have clips with two replays and one live action. As evidenced by the top part of Table 6, the type of view has a significant impact on the VARS's ability to detect the correct type of foul. Although the live-action view alone provides worse performance than the replays, combining the live-action view with a replay improves the accuracy slightly compared to using only the replay view in both tasks. This implies that even a poor-quality view can slightly improve the performance. A highly informative view can boost the performance, as we can see by comparing the two replays with a single replay for the type of foul classification. For the offence and severity classification, the VARS seems to benefit more from live actions compared to replays for the offense and severity classification task. One possible explanation is that for the live actions, the VARS takes into account the position of the action on the field, allowing it to learn that the likelihood of a 'No card' or 'Yellow card' is higher in specific areas of the field. For both tasks, we achieved better results by using multiple views, and for most of the metrics, the best performance was achieved by using a live-action clip with two replays. This demonstrates the effectiveness of using multiple views to improve model performance in the type of foul and offence severity classification.

In Figure 5, we show the predictions of the foul classification models while changing the number and type of views. By only using the live action, the VARS is not able to detect the correct type of foul, as confirmed in Table 6. By adding 1 or 2 replays as input to the model, it is able to detect the foul class with a confidence score ranging from 76% to 95%. By analyzing the confidence scores, we can see that the view has a big impact of the prediction, which agrees with the results found in Table 6.

Temporal analysis. We investigated the temporal context needed to identify fouls and offence severity. In particular, we increased the video length, by reducing the frame rate, in order to maintain the same number of frames to process. Table 7, shows the results of the temporal analysis. We observed that as we increase the temporal context while decreasing the frame rates, the performance of our model decreases. This is likely because the most useful information for our classification tasks is concentrated within a narrow temporal window immediately preceding and following the foul. Adding more temporal context to the model results in the inclusion of frames that do not offer much additional information. By default, we used a frame rate of 16 frames per second, with a temporal context of 1 second, which seemed to strike the best balance between capturing sufficient temporal information and excluding unnecessary frames.

Frame rate (FPS) **Temporal context**	**5** **3.2s**	**8** **2.0s**	**12** **1.3s**	**16** **1.0s**
Accuracy (Foul class.)	0.36	0.38	0.44	0.47
Accuracy (Off. sev. class.)	0.39	0.41	0.43	0.43

Table 7. **Temporal analysis.** We experiment with various temporal context while maintaining a fixed number of 16 frames. In all scenarios, we include 8 frames before and after the foul.

Per class analysis. We further analyze the performance per class. The confusion matrices for both tasks are in the supplementary material. We saw that performance varies considerably across classes. For the fine-grained foul classification, the VARS struggles to distinguish between illegal arm movements due to their shared characteristics. It performs well in detecting "Tackling", but often confuses it with "Dive" due to the challenge of distinguishing genuine from deceptive actions in soccer games. The most difficult class for the VARS is "Challenge", as it shares visual similarities with many other classes, making proper generalization during training difficult. Regarding offence classification, the VARS tends to make bad predictions in neighboring classes of the ground truth. For instance, it may classify a foul as "Offence + Yellow card" instead of "Offence + No card". However, the model struggles with "Offence + Red card" due to the limited number of samples in the dataset.

6. Conclusion

In summary, our Video Assistant Referee System (VARS) has the potential to bring about a significant improvement in soccer refereeing by ensuring fairness and accuracy at all levels of professional and amateur play. VARS utilizes the latest advances in multi-view video analysis and provides referees with real-time feedback and assists them in making informed decisions that can impact the outcome of soccer games. To prove the effectiveness of VARS, we introduced a novel dataset, SoccerNet-MVFoul, that curates relevant fouls in soccer broadcasts from multiple views and includes foul properties. Our benchmarking results demonstrate that VARS can recognize foul characteristics based on multi-view video processing. By integrating the specific requirements of referees, VARS offers an unbiased and reliable decision-making process for soccer matches.

Acknowledgement. This work was partly supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research through the Visual Computing Center (VCC) funding and the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI). A. Cioppa is funded by the F.R.S.-FNRS.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. YouTube-8M: A large-scale video classification benchmark. *CoRR*, abs/1609.08675, 2016. 2

[2] Adrià Arbués Sangüesa, Adriàn Martín, Javier Fernández, Coloma Ballester, and Gloria Haro. Using player's body-orientation to model pass feasibility in soccer. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3875–3884, Seattle, WA, USA, Jun. 2020. Inst. Electr. Electron. Eng. (IEEE). 2

[3] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 961–970, Boston, MA, USA, Jun. 2015. Inst. Electr. Electron. Eng. (IEEE). 2

[4] Alejandro Cartas, Coloma Ballester, and Gloria Haro. A graph-based method for soccer action spotting using unsupervised player classification. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 93–102, Lisbon, Port., Oct. 2022. ACM. 2

[5] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Scaling up SoccerNet with multi-view spatial localization and re-identification. *Sci. Data*, 9(1):1–9, Jun. 2022. 2, 3

[6] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas B. Moeslund. A context-aware loss function for action spotting in soccer videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 13123–13133, Seattle, WA, USA, Jun. 2020. Inst. Electr. Electron. Eng. (IEEE). 2

[7] Anthony Cioppa, Adrien Deliège, Noor Ul Huda, Rikke Gade, Marc Van Droogenbroeck, and Thomas B. Moeslund. Multimodal and multiview distillation for real-time player detection on a football field. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 3846–3855, Seattle, WA, USA, Jun. 2020. 2

[8] Anthony Cioppa, Adrien Deliège, and Marc Van Droogenbroeck. A bottom-up approach based on semantics for the interpretation of the main camera stream in soccer games. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 1846–1855, Salt Lake City, UT, USA, Jun. 2018. 2

[9] Anthony Cioppa, Silvio Giancola, Adrien Deliege, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. SoccerNet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 3490–3501, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[10] Abdulrahman Darwish and Tallal El-Shabrway. STE: Spatio-temporal encoder for action spotting in soccer videos. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 87–92, Lisbon, Port., Oct. 2022. ACM. 2

[11] Tom Decroos, Lotte Bransen, Jan Van Haaren, and Jesse Davis. Actions speak louder than goals. In *ACM SIGKDD Int. Conf. Knowl. Discov. & Data Min.*, page 1851–1861. ACM, Jul. 2019. 2

[12] Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 4508–4519, Nashville, TN, USA, Jun. 2021. Best CVSports paper award. 2, 3

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 248–255, Miami, FL, USA, Jun. 2009. Inst. Electr. Electron. Eng. (IEEE). 2, 6

[14] Carlos Esteves, Yinshuang Xu, Christine Allec-Blanchette, and Kostas Daniilidis. Equivariant multi-view networks. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 1568–1577, Seoul, South Korea, Oct. 2019. Inst. Electr. Electron. Eng. (IEEE). 3

[15] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 6804–6815, Montreal, QC, Canada, Oct. 2021. Inst. Electr. Electron. Eng. (IEEE). 2, 6, 7

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast networks for video recognition. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 6201–6210, Seoul, South Korea, Oct. 2019. Inst. Electr. Electron. Eng. (IEEE). 2

[17] FIFA. Goal-line technology. https://www.fifa.com/technical/football-technology/football-technologies-and-innovations-at-the-fifa-world-cup-2022/goal-line-technology, 2022. 1

[18] FIFA. Semi automated offside technology. https://www.fifa.com/technical/football-technology/football-technologies-and-innovations-at-the-fifa-world-cup-2022/semi-automated-offside-technology, 2023. 2

[19] Lianli Gao, Zhao Guo, Hanwang Zhang, Xing Xu, and Heng Tao Shen. Video captioning with attention-based LSTM and semantic consistency. *IEEE Trans. Multimedia*, 19(9):2045–2055, Sept. 2017. 2

[20] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. SoccerNet: A scalable dataset for action spotting in soccer videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 1792–179210, Salt Lake City, UT, USA, Jun. 2018. Inst. Electr. Electron. Eng. (IEEE). 2, 3

[21] Silvio Giancola, Anthony Cioppa, Adrien Deliège, Floriane Magera, Vladimir Somers, Le Kang, Xin Zhou, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdulrahman

Darwish, Adrien Maglo, Albert Clapés, Andreas Luyts, Andrei Boiarov, Artur Xarles, Astrid Orcesi, Avijit Shah, Baoyu Fan, Bharath Comandur, Chen Chen, Chen Zhang, Chen Zhao, Chengzhi Lin, Cheuk-Yiu Chan, Chun Chuen Hui, Dengjie Li, Fan Yang, Fan Liang, Fang Da, Feng Yan, Fufu Yu, Guanshuo Wang, H. Anthony Chan, He Zhu, Hongwei Kan, Jiaming Chu, Jianming Hu, Jianyang Gu, Jin Chen, João V. B. Soares, Jonas Theiner, Jorge De Corte, José Henrique Brito, Jun Zhang, Junjie Li, Junwei Liang, Leqi Shen, Lin Ma, Lingchi Chen, Miguel Santos Marques, Mike Azatov, Nikita Kasatkin, Ning Wang, Qiong Jia, Quoc Cuong Pham, Ralph Ewerth, Ran Song, Rengang Li, Rikke Gade, Ruben Debien, Runze Zhang, Sangrok Lee, Sergio Escalera, Shan Jiang, Shigeyuki Odashima, Shimin Chen, Shoichi Masui, Shouhong Ding, Sin-wai Chan, Siyu Chen, Tallal El-Shabrawy, Tao He, Thomas B. Moeslund, Wan-Chi Siu, Wei Zhang, Wei Li, Xiangwei Wang, Xiao Tan, Xiaochuan Li, Xiaolin Wei, Xiaoqing Ye, Xing Liu, Xinying Wang, Yandong Guo, Yaqian Zhao, Yi Yu, Yingying Li, Yue He, Yujie Zhong, Zhenhua Guo, and Zhiheng Li. SoccerNet 2022 challenges results. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 75–86, Lisbon, Port., Oct. 2022. ACM. 2, 3

[22] Abdullah Hamdi, Faisal AlZahrani, Silvio Giancola, and Bernard Ghanem. MVTN: Learning multi-view transformations for 3D understanding. *CoRR*, abs/2212.13462, 2022. 6

[23] Abdullah Hamdi, Silvio Giancola, and Bernard Ghanem. MVTN: Multi-view transformation network for 3D shape recognition. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 1–11, Montreal, QC, Canada, Oct. 2021. Inst. Electr. Electron. Eng. (IEEE). 3

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 770–778, Las Vegas, NV, USA, Jun. 2016. Inst. Electr. Electron. Eng. (IEEE). 6, 7

[25] Yutaro Honda, Rei Kawakami, Ryota Yoshihashi, Kenta Kato, and Takeshi Naemura. Pass receiver prediction in soccer using video and players' trajectories. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3502–3511, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[26] James Hong, Haotian Zhang, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Spotting temporally precise, fine-grained events in video. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 13695 of *Lect. Notes Comput. Sci.*, pages 33–51. Springer Nat. Switz., 2022. 2

[27] IFAB. Laws of the game. Technical report, The International Football Association Board, Zurich, Switzerland, 2022. 4, 13, 15

[28] Yudong Jiang, Kaixu Cui, Leilei Chen, Canjin Wang, and Changliang Xu. SoccerDB: A large-scale database for comprehensive video understanding. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, page 1–8. ACM, Oct. 2020. 2

[29] Asako Kanezaki, Yasuyuki Matsushita, and Yoshifumi Nishida. RotationNet: Joint object categorization and pose estimation using multiviews from unsupervised viewpoints. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5010–5019, Salt Lake City, UT, USA, Jun. 2018. Inst. Electr. Electron. Eng. (IEEE). 3

[30] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1725–1732, Columbus, OH, USA, Jun. 2014. Inst. Electr. Electron. Eng. (IEEE). 2

[31] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 2, 3, 6

[32] Muhammad Zeeshan Khan, Summra Saleem, Muhammad A. Hassan, and Muhammad Usman Ghanni Khan. Learning deep C3D features for soccer video event detection. In *Int. Conf. Emerg. Technol. (ICET)*, pages 1–6, Islamabad, Pakistan, Nov. 2018. 2

[33] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 706–715, Venice, Italy, Oct. 2017. Inst. Electr. Electron. Eng. (IEEE). 2

[34] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images, 2009. Technical report, University of Toronto. 2

[35] Junnan Li, Yongkang Wong, Qi Zhao, and Mohan S Kankanhalli. Unsupervised learning of view-invariant action representations. *Advances in neural information processing systems*, 31, 2018. 3

[36] Yitong Li, Martin Min, Dinghan Shen, David Carlson, and Lawrence Carin. Video generation from text. *AAAI Conf. Artif. Intell.*, 32(1), Apr. 2018. 2

[37] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4804–4814, 2022. 2, 6, 7

[38] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(10):2684–2701, Oct. 2020. 3

[39] Adrien Maglo, Astrid Orcesi, and Quoc-Cuong Pham. Efficient tracking of team sport players with few game-specific annotations. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3460–3470, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[40] Javier Mallo, Pablo Gonzalez Frutos, Daniel Ju'arez, and Enrique Navarro. Effect of positioning on the accuracy of decision making of association football top-class referees and assistant referees during competitive matches. *Journal of Sports Sciences*, 30(13):1437–1445, Sept. 2012. 4, 5

[41] Thomas B. Moeslund, Graham Thomas, and Adrian Hilton. *Computer vision in sports*. Springer, 2014. 2

[42] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4694–4702, Boston, MA, USA, Jun. 2015. Inst. Electr. Electron. Eng. (IEEE). 2

[43] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. A public data set of spatio-temporal match events in soccer competitions. *Sci. Data*, 6(1):1–15, Oct. 2019. 2

[44] Hossein Rahmani, Arif Mahmood, Du Huynh, and Ajmal Mian. Histogram of oriented principal components for cross-view action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 38(12):2430–2443, Dec. 2016. 3

[45] Himangi Saraogi, Rahul Anand Sharma, and Vijay Kumar. Event recognition in broadcast soccer videos. In *Indian Conference on Computer Vision, Graphics and Image Processing*, pages 1–7, Dec. 2016. 2

[46] Atom Scott, Ikuma Uchida, Masaki Onishi, Yoshinari Kameda, Kazuhiro Fukui, and Keisuke Fujii. SoccerTrack: A dataset and tracking algorithm for soccer with fish-eye and drone videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3568–3578, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[47] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 27, pages 568–576, Dec. 2014. 2

[48] João V. B. Soares and Avijit Shah. Action spotting using dense detection anchors revisited: Submission to the SoccerNet challenge 2022. *CoRR*, abs/2206.07846, 2022. 2

[49] Joao V. B. Soares, Avijit Shah, and Topojoy Biswas. Temporally precise action spotting in soccer videos using dense detection anchors. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 2796–2800, Bordeaux, France, Oct. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[50] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body part-based representation learning for occluded person Re-Identification. In *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pages 1613–1623, Waikoloa, HI, USA, Jan. 2023. Inst. Electr. Electron. Eng. (IEEE). 2

[51] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 2

[52] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik Learned-Miller. Multi-view convolutional neural networks for 3D shape recognition. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 945–953, Santiago, Chile, Dec. 2015. Inst. Electr. Electron. Eng. (IEEE). 3

[53] Genki Suzuki, Sho Takahashi, Takahiro Ogawa, and Miki Haseyama. Team tactics estimation in soccer videos based on a deep extreme learning machine and characteristics of the tactics. *IEEE Access*, 7:153238–153248, 2019. 2

[54] Graham Thomas, Rikke Gade, Thomas B. Moeslund, Peter Carr, and Adrian Hilton. Computer vision for sports: current applications and research topics. *Comput. Vis. Image Underst.*, 159:3–18, Jun. 2017. 2

[55] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 6450–6459, Salt Lake City, UT, USA, Jun. 2018. Inst. Electr. Electron. Eng. (IEEE). 2, 6, 7

[56] UEFA. New age for referees. https://www.uefa.com/insideuefa/news/0194-0f844fe11b14-503126864653-1000--new-age-for-referees/, Aug. 2003. 1

[57] Gabriel Van Zandycke, Vladimir Somers, Maxime Istasse, Carlo Del Don, and Davide Zambrano. DeepSportradar-v1: Computer vision dataset for sports understanding with high quality annotations. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 1–8, Lisbon, Port., Oct. 2022. ACM. 2

[58] Renaud Vandeghen, Anthony Cioppa, and Marc Van Droogenbroeck. Semi-supervised training to improve player and ball detection in soccer. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 3480–3489, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[59] Shruti Vyas, Yogesh S. Rawat, and Mubarak Shah. Multi-view action recognition using cross-view video prediction. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 12372 of *Lect. Notes Comput. Sci.*, pages 427–444. Springer Int. Publ., 2020. 3

[60] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. Cross-view action modeling, learning, and recognition. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 2649–2656, Columbus, OH, USA, Jun. 2014. Inst. Electr. Electron. Eng. (IEEE). 3

[61] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 9912 of *Lect. Notes Comput. Sci.*, pages 20–36. Springer Int. Publ., 2016. 2

[62] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(11):2740–2755, Nov. 2019. 2

[63] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. Video captioning via hierarchical reinforcement learning. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4213–4222, Salt Lake City, UT, USA, Jun. 2018. Inst. Electr. Electron. Eng. (IEEE). 2

[64] Xin Wei, Ruixuan Yu, and Jian Sun. View-GCN: View-based graph convolutional network for 3D shape analysis. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1847–1856, Seattle, WA, USA, Jun. 2020. Inst. Electr. Electron. Eng. (IEEE). 3

[65] Fei Wu, Qingzhong Wang, Jian Bian, Haoyi Xiong, Ning Ding, Feixiang Lu, Jun Cheng, and Dejing Dou. A survey on video action recognition in sports: Datasets, methods and applications. *CoRR*, abs/2206.01038, 2022. 2

[66] Junqing Yu, Aiping Lei, Zikai Song, Tingting Wang, Hengyou Cai, and Na Feng. Comprehensive dataset of

broadcast soccer videos. In *IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, pages 418–423, Miami, FL, USA, Apr. 2018. Inst. Electr. Electron. Eng. (IEEE). 2

[67] He Zhu, Junwei Liang, Chengzhi Lin, Jun Zhang, and Jianming Hu. A transformer-based system for action spotting in soccer videos. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 103–109, Lisbon, Port., Oct. 2022. ACM. 2