# Towards an AI-Powered Video Assistant Referee System (VARS) for Association Football

Jan Held[1], Anthony Cioppa[1], Silvio Giancola[2], Abdullah Hamdi[3], Christel Devue[1], Bernard Ghanem[2], and Marc Van Droogenbroeck[1]

[1] University of Liege (ULiège), Belgium
[2] King Abdullah University of Science and Technology (KAUST), Saudi Arabia
[3] University of Oxford, United Kingdom

**Abstract.** Over the past decade, the technology used by referees in football has improved substantially, enhancing the fairness and accuracy of decisions. This progress has culminated in the implementation of the Video Assistant Referee (VAR), an innovation that enables backstage referees to review incidents on the pitch from multiple points of view. However, the VAR is currently limited to professional leagues due to its expensive infrastructure and the lack of referees worldwide. In this paper, we present the Video Assistant Referee System (VARS) that leverages the latest findings in multi-view video analysis. Our VARS achieves a new state-of-the-art on the *SoccerNet-MVFoul* dataset by recognizing the type of foul in 50% of instances and the appropriate sanction in 46% of cases. Finally, we conducted a comparative study to investigate human performance in classifying fouls and their corresponding severity and compared these findings to our VARS. The results of our study highlight the potential of our VARS to reach human performance and support football refereeing across all levels of professional federations.

**Keywords:** Football · Soccer · Artificial Intelligence · Automated Decision · Video Assistant Referee · Referee Success Rate · Human study.

## 1 Introduction

In recent years, technology has played an increasing role in football, revolutionizing how the game is played, coached, and officiated. This transformation extends into the domain of sports video analysis, which encompasses a diverse range of challenging tasks, including player detection and tracking [3, 25, 27], spotting actions in untrimmed videos [1, 2, 11, 12], camera calibration [22], player re-identification in occluded scenarios [24], or dense video captioning for football broadcasts commentaries [23]. Similar to many other fields in which deep learning has been used, the advancements in sports video understanding heavily rely on the availability of large-scale datasets. SoccerNet [4, 5, 9, 10, 13, 14, 18] stands among the largest and most comprehensive sports dataset, with extensive annotations for video understanding in football.

In refereeing, the most significant evolution was introduced by the Video Assistant Referee (VAR) in 2016. The system involves a team of referees located in a video operation room outside the stadium. These referees have access to all available camera views and check all decisions taken by the on-field referee. If the VAR indicates a probable “clear and obvious error” (*e.g.* when the referee misses a penalty or a red card, gives a yellow card to the wrong player, etc.), it will be communicated to the on-field referee who can then review his decision in the referee review area before taking a final decision. The VAR helps to ensure greater fairness in the game by reducing the impact of incorrect decisions on the outcome of games. Notably, in 8% of the matches, the VAR has a decisive impact on the result of the game [6] and it slightly reduces the unconscious bias of referees towards home teams [15]. Controversial referee mistakes like the famous “hand of God” goal by Diego Maradona during the quarter-final match Argentina versus England of the 1986 FIFA World Cup, Josip Šimunić getting three yellow cards in a single game at the 2006 FIFA World Cup, or Thierry Henry’s handball preventing the Republic of Ireland from qualifying for the World Cup could have been avoided with the VAR and would have changed football history.

Despite its potential benefits, the use of the VAR technology remains limited to professional leagues. The infrastructure of the VAR is expensive, including multiple cameras to analyze the incident from different angles, video operation rooms in various locations, and VAR officials hired to analyze the footage. In addition to the upfront costs of the infrastructure, there is also an ongoing expense associated with using the VAR. The officials who serve as Video Assistant Referees require specialized training and monetary compensation following each game. Given the implementation and operational costs of VAR, its use is currently restricted to professional leagues. A further obstacle is the shortage of referees worldwide. In Germany, there were only 50,241 active referees during the 2020/2021 season, whereas the number of games played each weekend was around 90,000 [7, 29]. The introduction of the VAR requires an additional team of referees per game, which is not feasible for semi-professional or amateur leagues. Finally, each referee interprets the Laws of the Game [16] slightly differently, resulting in different decisions for similar actions. Given that the video assistant referee changes from one game to another, inconsistencies may arise, with the VAR making different decisions for similar actions across different matches.

In this paper, we present the “Video Assistant Referee System” (VARS), which could support or extend the current VAR. Our VARS fulfills the same objectives and tasks as the VAR. By analyzing fouls from a single or a multi-camera video feed, it indicates a probable “clear and obvious error”, and can communicate this information to the referee, who will then decide whether to initiate a “review”. The proposed VARS automatically analyzes potential incidents that can then be shown to the referee in the referee review area. Just like the regular VAR, our VARS serves as a support system for the referee and only alerts him in the case of potential game-changing mistakes, but the final decision remains in the hands of the main referee. The main benefit of our VARS is that it no longer requires additional referees, making it the perfect tool for leagues that do not have enough financial or human resources.

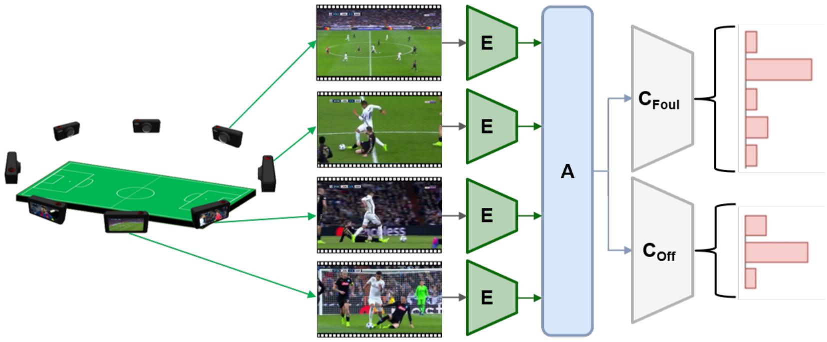

**Fig. 1. Architecture of our Video Assistant Referee System.** From multi-view video clips input, our system encodes per-view video features (**E**), aggregates the view features (**A**), and classifies different properties ($\mathbf{C}_{\mathbf{Foul}}$ and $\mathbf{C}_{\mathbf{Off}}$). This figure is a version adapted of [13] with permission from the original authors.

**Contributions.** We summarize our contributions and novelties as follows: **(i)** We propose an upgraded version of the *VARS* presented by Held *et al.* [13]. We introduce an attention mechanism on the different views and calculate an importance score to allocate more attention to more informative views before aggregating the views. **(ii)** We present a thorough study on the influence of using multiple views and different types of camera views on the performance of our VARS. **(iii)** We present a comprehensive human study where we compare the performance of human referees, football players, and our VARS on the task of type of foul classification and offense severity classification. Our human study also illustrates the subjectivity of refereeing decisions by examining the inter-rater agreement among referees. Note that our study was approved by the University of Liège's ethics committee (2223-080/5624).

## 2 Methodology

We propose an upgraded version of the Video Assistant Referee System, which adds an advanced pooling technique to combine the information from multiple views, extracting the most relevant information based on our attention mechanisms. The architecture is shown in Figure 1. Formally, the VARS takes multiple video clips $\mathbf{v} = \{v_i\}_1^n$ as input. Each video clip shows the same action from $n$ different perspectives. Each clip $v_i$ is fed into a video encoder **E** to extract a spatio-temporal feature vector $f_i$ of dimension $d$ for each clip $v_i$. All feature vectors $f_i$ are then stored in a matrix **f** as follows: $\mathbf{f} = \left[f_1, f_2, ..., f_n\right]^T$. An aggregation block **A** takes **f** as input and outputs a single multi-view representation **R**. A multi-head classifier, $\mathbf{C}^{\text{foul}}$ and $\mathbf{C}^{\text{off}}$, simultaneously predicts the fine-grained type of foul class and the offense severity class. For each task, the VARS selects the value with the highest confidence from the respective confidence vector as

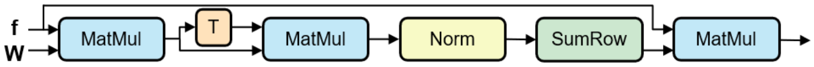

**Fig. 2. Architecture of the attention block.** “MatMul” represents matrix multiplication, “T” denotes transpose, “Norm” signifies normalization, and “SumRow” indicates the process of summing each row.

the final prediction, following:

$$\mathbf{VARS}^t \leftarrow \mathrm{argmax}\, \mathbf{C}^t_{\theta_{C^t}}(\mathbf{R}), \forall t \in \{\mathrm{foul}, \mathrm{off}\}, \tag{1}$$

where $\theta_{C^t}$ corresponds to the parameters of the classification head for task $t \in \{\mathrm{foul}, \mathrm{off}\}$. The model is trained by minimizing the unweighted summation of both task losses $\mathcal{L}^{\mathrm{foul}}$ and $\mathcal{L}^{\mathrm{off}}$.

**Video Encoder E.** Based on the work presented in [13], the best performance is obtained with a video encoder that extracts spatial and temporal features. We use the state-of-the-art video encoder MViT [8, 19] pretrained on Kinetics [17], which incorporates a transformer-based architecture with a multiscale feature representation, allowing it to capture spatial and temporal information.

**Multi-view aggregation block A.** The original paper [13] used simple mean or max pooling operations to gather the multi-view information into a unique representation. A major drawback of these pooling approaches is that the combination of the feature vectors is fixed and ignores the relationship between the views. Instead, we propose a new aggregation technique based on an attention mechanism to model such relationships.

Our approach is inspired by the “Integrating Block” presented in [28], where each view is associated with an attention score. However, instead of aggregating multi-view images, we extend the operation to multi-view videos. Technically, we assign an attention score to each view and then calculate the final representation by a weighted sum of the feature vectors. There exist several strategies to assign an attention score to a view. A first naive approach consists of passing each feature vector individually into a learned function. However, this would neglect the relationships between the views and would not provide a relative attention score of the views. A better approach consists of determining the attention score of each view based on its relationships with the other views. To do so, we first take the dot product (denoted by $\cdot$) of $\mathbf{f}$ multiplied by a matrix $W \in \mathbb{R}^{dxd}$ of trainable weights and its transpose: $\mathbf{S} = \mathbf{f}W \cdot (\mathbf{f}W)^T$.

By multiplying the matrix $\mathbf{f}$ with its transpose, we compute the dot product between each pair of feature vectors, which measures the similarity between two vectors. The obtained symmetric similarity matrix $\mathbf{S}$ is of dimension $n \times n$, where the value at row $i$ and column $j$ corresponds to the similarity score between view $i$ and view $j$. A higher score indicates a higher similarity between the vectors, while a lower score suggests a lower similarity. Next, we normalize the similarity scores to get a probability-like distribution, by passing the matrix $\mathbf{S}$ through a $ReLU$ layer and dividing it by the sum of the matrix $\mathbf{S}$, following:

$\mathbf{N} = \frac{ReLU(\mathbf{S})}{\sum_{i=1}^{n}\sum_{j=1}^{n} ReLU(\mathbf{S}_{i,j})}$ . To obtain the attention score for each view, we sum the values in each row of the normalized similarity matrix $\mathbf{N}$. The attention score for a view $i$ represents the sum of its normalized similarity scores with all other views, reflecting how similar it is to all views collectively. Consequently, the resulting attention score captures a view's overall relevance within the set of views. The reasoning behind this approach is that if a view is highly similar to many other views, it is considered important because it shares visual content with multiple views. If a view is dissimilar to other views, it might be considered less important since it does not contribute significantly to the collective visual information. Formally, we take the sum per row to obtain the attention score $\mathbf{A}$ per view: $\mathbf{A} = \sum_{i=1}^{n} \mathbf{N}_{i,j}$ , where $\mathbf{A}$ is a vector of size $n$, where the value $j$ corresponds to the attention score of the view $j$ regarding all other views and itself. The final representation is given by the sum of the extracted feature vector weighted by their calculated attention score, following $\mathbf{R}_i = \sum_{j=1}^{n} \mathbf{f}_{i,j} \times \mathbf{A}_j$ .

**Classification heads C.** A multi-task classification approach is used to classify simultaneously the type of foul, whether it is an offense or not, and its severity. As both tasks are related, learning them together can lead to improved generalization and a better understanding of each task. The model can leverage the relationships between the two tasks to make better predictions. Each classification head consists of two dense layers and takes as input the aggregated representation. The output is a vector whose dimensions correspond to the number of classes in each of the classification problems.

# 3 Experiments

## 3.1 Experimental setup

**Tasks.** We test our VARS on the two classification tasks introduced by the SoccerNet-MVFouls dataset [13]: *Fine-grained foul classification*, which is the task of classifying a foul into one of 8 fine-grained foul classes (*i.e.*, "Standing tackling", "Tackling", "High leg", "Pushing", "Holding", "Elbowing", "Challenge", and "Dive/Simulation"), and *Offence severity classification*, which is the task of classifying whether an action is an offence, as well as the severity of the foul, defined by four classes: "No offence", "Offence + No card", "Offence + Yellow card", and "Offence + Red card".

**Data.** The SoccerNet-MVFoul dataset contains 3,901 actions, composed of at least two videos, the live action and at least one replay, see Figure 1. The views were manually synchronized by a human and no pre-processing of the video clips is necessary. Our VARS is trained on clips of 16 frames, mostly 8 frames before the foul and 8 after the foul, spanning one second temporally with a spatial dimension re-scaled to 224 × 224 pixels. This approach is chosen because of the high computational cost associated with using a larger number of frames. Future research could explore whether an increase in frame rate or a larger temporal context enhances performance.

**Training details.** The encoder **E** is pre-trained as detailed in the methodology, and the classifier **C** is trained from scratch, both end-to-end. We use a cross-entropy loss, optimized with Adam on a batch size of 6 samples. The learning rate starts at $5e^{-5}$ and is multiplied by 0.3 every 3 steps. To artificially increase the dataset size, we use data augmentation and a random temporal shift to have a flexible number of frames used before and after the foul frame annotation during training. The model begins to overfit after 7 epochs and requires approximately 8 hours of training time on a single NVIDIA V100 GPU.

**Evaluations metrics.** To evaluate the performance of the VARS, SoccerNet-MVFouls uses the classification accuracy, defined as the ratio of correctly classified actions to the total number of actions. As SoccerNet-MVFouls [13] is unbalanced, the authors also suggest a balanced accuracy (BA), which is defined as $\mathrm{BA} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{P_i}$ , with $N$ being the number of classes, $TP_i$ the number of True Positives and $P_i$ the number of Positives for class $i$. To ensure a fair comparison, we use the same training, validation, and test sets as those used in the original paper [13].

### 3.2 Main results

Table 1 shows the results obtained for the fine-grained foul and the offense severity classification task. Compared to the fixed combination of the feature vectors (mean or max pooling), our novel attention mechanism enhances the model's ability to identify and classify the type of foul by 5% and the balanced accuracy by 1%. This demonstrates the effectiveness of combining the feature vectors of the different views based on their importance compared to max or mean pooling. Similarly, the attention mechanism improves the model's performance to determine if an action is a foul and the corresponding severity by 3% and the balanced accuracy remains the same compared to max pooling. One might argue that the performance increase is based on the supplementary parameters introduced by the attention mechanism. However, the attention mechanism only adds an extra 0.1% of parameters to the model compared to when using max and mean pooling. This suggests that the performance increase derives from the use of the attention mechanism rather than the introduction of additional parameters.

### 3.3 Detailed analysis

**Sensitivity analysis.** We first investigate the impact of the training dataset size on the performance of our two classification tasks. Figure 3 shows the evolution of the accuracy regarding different training dataset sizes. For each dataset size, we independently trained and tested the model 10 times to avoid any epistemic uncertainty bias. The tests were all performed on the same test set. As expected, we observe that increasing the dataset size improves the accuracy of our VARS. For the type of foul classification, we notice a significant improvement in accuracy with increasing dataset size, especially at the beginning. However, the accuracy reached a plateau between 40% and 80% of the data. Interestingly,

**Table 1. Multi-task classification.** Attention pooling sets a new benchmark on the *SoccerNet-MVFoul* dataset for all the evaluation metrics and tasks. The type of foul classification accuracy is increased by 5% while the balanced accuracy (BA) is increased by 1%. We have an increment of 3% for the offense severity classification, while the balanced accuracy stays the same.

| | | **Type of Foul** | | **Offense Severity** | |
|---|---|---|---|---|---|
| Feature extraction | Pooling | Acc. | BA | Acc. | BA |
| ResNet | Mean | 0.30 | 0.27 | 0.34 | 0.25 |
| ResNet | Max | 0.32 | 0.27 | 0.32 | 0.24 |
| R(2+1)D | Mean | 0.32 | 0.34 | 0.34 | 0.30 |
| R(2+1)D | Max | 0.34 | 0.33 | 0.39 | 0.31 |
| MViT | Mean | 0.44 | 0.40 | 0.38 | 0.31 |
| MViT | Max | 0.45 | 0.39 | 0.43 | **0.34** |
| MViT | Attention | **0.50** | **0.41** | **0.46** | **0.34** |

we observed a sudden increase in accuracy when we increased the dataset size from 80% to 100%. This may be due to our unbalanced dataset. The dataset contains numerous “Standing tacklings” and “Tacklings”, while other labels are underrepresented. Increasing the dataset size from 40% to 80% may not have improved accuracy if the model still struggles to generalize to certain actions due to limited samples. Increasing the dataset size to 100% could have provided the model with the additional data necessary to better generalize actions. Moreover, Figure 3 reveals that our VARS is significantly more prone to epistemic uncertainty for smaller datasets, as indicated by the high standard deviation.

In contrast, the offense severity curve in Figure 3 initially shows a sharp increase, but later demonstrates a slower growth. Yet, with each increase in the dataset size, the accuracy improves, which confirms that more data would further improve the performance. The reason for this lies in the significant variability in the visual appearance of an offense with “No card”, “Yellow card”, or “Red card”. For instance, a yellow card can be the outcome of a tackle, or it can be the result of a player holding an opponent’s shirt. Although both instances may result in a yellow card, their visual representations differ significantly. To accurately determine whether an action is an offense or not and the corresponding severity, the model needs plenty of examples to learn the underlying distribution.

**Qualitative results.** Figure 4 shows the prediction of our VARS on two examples with a 3-view setup. In both examples, the VARS correctly determines the type of foul and correctly classifies both actions as a foul with the correct severity. Furthermore, the attention scores offer valuable insights into the contribution of different views or camera angles to the decision-making process of the model. In both cases, the “live action clips” have the lowest attention score, confirming our intuition that they were filmed from too far away to make an accurate decision. Both replays have a similar attention score, as they both offer a lot of information to the model. However, we can see that the most informative view has a slightly higher attention score. The attention score provides insight on which views contribute the most to classifications and helps us better under-

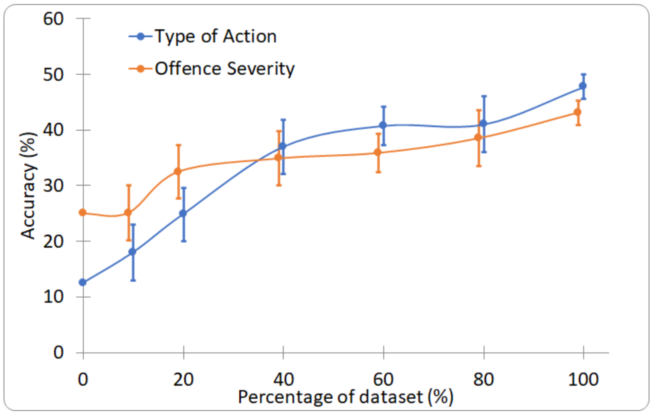

**Fig. 3. Performance evaluation for different dataset sizes.** 100% of the dataset corresponds to 2,319 actions. For each dataset size, we independently trained and tested the model 10 times. The tests were all performed on the same test set. The error bar corresponds to the standard deviation. For 0% of the dataset, we indicate the accuracy by taking a random decision.

stand how the model processes the visual data. This interpretability is especially important when the VARS is used in practice, as it is essential for fans, players, and referees to understand the reasoning behind decisions and feel confident that the technology is improving the fairness and integrity of the sport. Finally, the attention scores assigned to each view can assist broadcasters in automatically selecting the optimal camera angle for broadcasting purposes. Furthermore, it can support the VAR and helps accelerate the review process by automatically proposing the most informative camera perspective. This is particularly useful at a professional level, where the VAR can have up to 30 different camera perspectives at their disposal, making finding the optimal camera a challenge on its own. The attention scores would provide valuable information by highlighting the views that are more likely to provide crucial details, to accelerate the decision-making process during the VAR review.

## 4 Human study

In contrast to classical classification tasks that involve well-defined and easily separable classes, determining whether an action in football constitutes a foul may be subjective. Despite the definitions and regulations provided by the *Laws of the Game* [16], the rule book published by the IFAB regarding when an action in football is considered a foul and its corresponding severity, these guidelines are still open to interpretation, leading to differing opinions about the same action. In practice, many actions fall into this gray area where both interpretations, foul or no foul, could be considered correct. In this study, we first analyze whether and how the performance of our VARS aligns with human performance (*i.e.*,

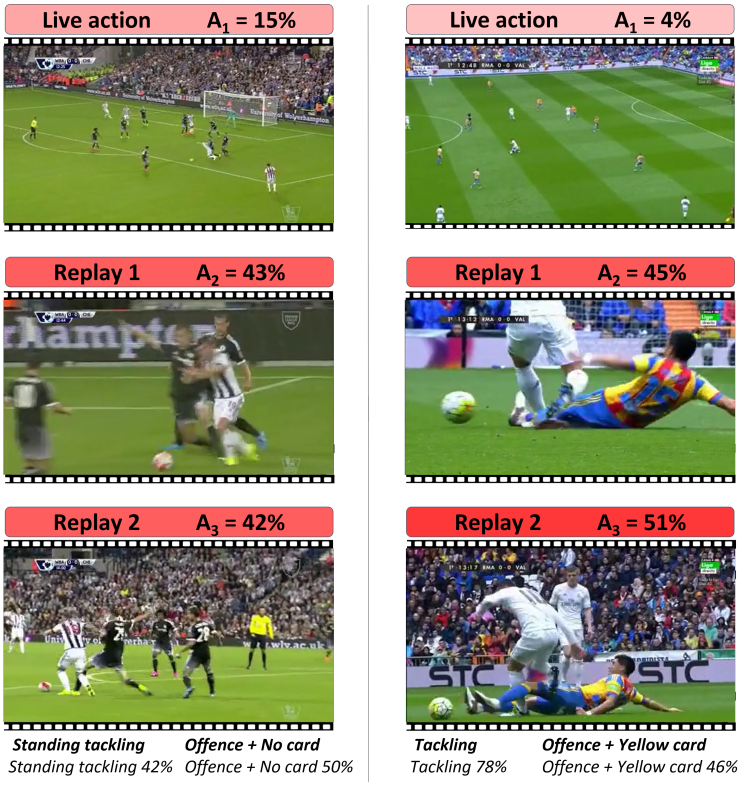

**Fig. 4. Qualitative results.** VARS prediction on two examples where the attention score of each view is given in percentage. The ground truth is given in bold and the model prediction with the confidence is given in italic.

referees and football players) by comparing the accuracy of the type of foul and offense severity classifications between VARS and our human participants. Secondly, we conduct an inter-rater agreement analysis of human decisions to quantify the extent of agreement among our human participants.

**Experimental setup.** The study involves two distinct groups of participants with different expertise in football: "Players" and "Referees". The first group consisted of 15 male individuals aged 18 or older (with a mean M = 23.06 and a standard deviation SD = 3.49 years), who had been playing football for a minimum of three years (M = 8.71 and SD = 3.32 years). The second group consisted of 15 male individuals aged 18 or older (M = 25.33 and SD = 4.51 years), who are certified football referees and have officiated in at least 200 official games (from 223 to 1150 games). Both groups analyzed 77 actions, each presented with three different camera perspectives simultaneously. The participants could review the clips several times and watch the actions in slow motion or frame-by-frame, without any time restriction. To reduce bias, the actions were shown in a dif-

**Table 2. Accuracy comparison between referees, players, and our VARS.** The survey was performed on a subset of the test set of size 77. The time is given in seconds and represents the average time needed to make a decision. A rating of 5 indicates high confidence, while a rating of 1 indicates low confidence.

| | **Type of Foul** | | **Offense Severity** | | **Time** |
|---|---|---|---|---|---|
| | Accuracy | Confidence | Accuracy | Confidence | |
| Players | **75%** | 3.6 | 58% | 3.3 | 41.53 |
| Referees | 70% | 3.7 | **60%** | 3.6 | 38.01 |
| VARS | 60% | - | 51% | - | **0.12** |

ferent random order to each participant. For each action, we measured the time taken by the participants to make their decision. This time was measured from the moment the participants started the video until they clicked on the 'Next video' button. For each action, the participants had the same classification task as presented in Section 3.1. Specifically, they had to determine the type of foul, if the action was a foul or not, and the corresponding severity. For each action, we use the annotations from the SoccerNet-MVFoul dataset as ground truth to determine the accuracy for each participant. An important note is that the participants have a clear advantage over our VARS as they view clips lasting 5 seconds, with a frame rate of 25 fps, while our model gets a 1-second clip at 16 fps as input. All analyses were performed using the JASP software.

### 4.1 Comparison to human performance

Table 2 shows the average accuracy compared to the ground truth of players, referees, and our VARS, respectively. These results align with similar studies [21, 26], where the referees had an overall decision accuracy ranging from 45% to 80%. In terms of the type of foul categorization, players (M = 0.752, SD = 0.055) were numerically more accurate than referees (M = 0.704, SD = 0.120), but this difference was not statistically significant, as shown by an independent samples Student t-test, $t(28) = 1.421$, $p = 0.166$, $d = 0.519$, 95% CI = $[-0.214$ - $1.243]$. Mean confidence levels in these categorizations were comparable between players (M = 3.64, SD = 0.28) and referees (M = 3.71, SD = 0.32), $t(28) < 1$.

As for determining if an action corresponds to a foul and the corresponding severity, referees were slightly more accurate (M = 0.594, SD = 0.091) than players (M = 0.582, SD = 0.061). However, this difference was not statistically significant, $t(28) = -0,401$, $p = 0.691$, $d = -0.147$, 95% CI = $[-0.862$ - $0.571]$. Although the accuracy of players and referees was comparable, referees were more confident in their severity judgments (M = 3.67, SD = 0.36) than players (M = 3.33, SD = 0.39), $t(28) = -2.3$, $p = 0.029$, $d = -0.839$, 95% CI = $[-1.581$ - $-0.084]$. Referees' higher confidence might be due to their specific experience in assessing fouls and their severity on the field.

Overall, our results suggest that the accuracy of players and referees is comparable. The Bayesian version of the Student t-test provides support for this null hypothesis with Bayes factors BF10 of 0.732 and 0.366 for the type of foul and

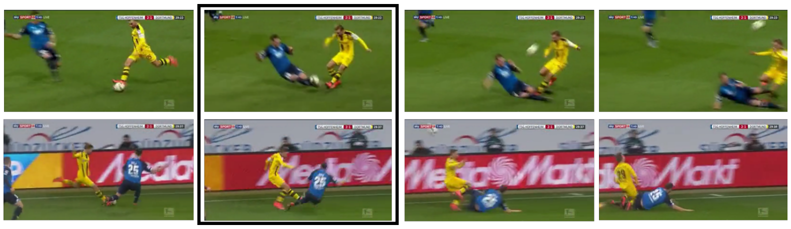

**Fig. 5. Example of the subjectivity of human choices.** Decisions taken by our participants: “No offense”, “Offense + No card”, and “Offense + Yellow card”.

offense severity task, respectively. There is a possibility that this lack of difference between groups is due to power issues, *i.e.*, the sample size being too small. Replication studies conducted on larger groups would be valuable in revealing potential differences between the two human groups.

As we do not have a standard deviation for the VARS, we conducted two One-Sample t-tests to compare its performance against humans (players and referees were grouped as their accuracy was comparable). For action categorization, humans (M = 0.728, SD = 0.095) were significantly more accurate than our VARS (M = 0.597), $t(29) = 7.556$, $p < .001$, $d = 1.379$, 95% CI = [0.870 - 1.876]. Humans were also more accurate (M = 0.588, SD = 0.081) than our VARS (M = 0.508) for offense severity judgments, $t(29) = 5.492$, $p < .001$, $d = 1.003$, 95% CI = [0.556 - 1.437]. This difference in performance might be due to differences in training between our VARS and humans. Players and referees have accumulated an extensive amount of experience in football, through officiating, playing, and watching the game for countless hours. In contrast, our VARS has only been trained on an unbalanced training set of 2,916 actions, where some types of labels only occur a few times. For example, there are only 27 fouls with a red card in the training set, making it difficult for the model to precisely learn the difference between a foul with a yellow card and one with a red card. Considering the difficulty of the task and the significant experience disadvantage of our VARS compared to humans, the current results are promising. Further, it is notable that our VARS only requires $120ms$ to reach a decision, which is more than 300 times faster than humans. Both referees and players require a similar amount of time to decide. On average, players take around 41.53 seconds and referees 38.01 seconds, which is similar to the average time of 46 seconds taken for the VAR to decide as reported by López [20].

### 4.2 Inter-rater agreement

We now investigate the reliability and consistency of humans in determining whether an action constitutes a foul and its severity. To assess the level of consensus among humans, we calculated inter-rater agreement in each group for the severity classification task. Since determining if an action is a foul and assessing its severity is the most important task, we only focus on evaluating inter-rater

**Table 3. Similarity analysis of the results for Offense Severity classification.** Among high-level referees, 28% of cases result in three different decisions being made for the same action. For referee talents, this percentage even increases to 38%. This shows how difficult it is to determine whether an action is a foul and assess its severity.

| **Number of different decisions** | **1** | **2** | **3** | **4** |
|---|---|---|---|---|
| High-level referees | 16% | 56% | 28% | 0% |
| Referee talents | 2% | 60% | 38% | 0% |

agreement for this aspect. To quantify the inter-rater agreement, we calculated the unweighted average Cohen's kappa, which measures the agreement between multiple individuals. The referees achieved an unweighted average Cohen's kappa of 0.213, indicating weak agreement. Similarly, players' agreement was weak, with a score of 0.223. This suggests limited consistency among both groups in their assessments. Among our 15 referees, 7 are officiating at a high level (in the highest league of their country). These referees are called "high-level referees" in the following. All other referees are called "referee talents". Table 3 shows the consensus in each subgroup for the offense severity classification task. As can be seen, high-level and referee talents reached a consensus between themselves for only 16% and 2% of the actions, respectively. For most cases, multiple decisions were made for the same action, indicating the difficulty in determining whether an action should be classified as a foul and assessing its severity. Particularly among referee talents, 38% of actions resulted in three different decisions (out of four possible decisions to take) for the same action. Figure 5 shows an example of an action where all three decisions "No offense", "Offense + No card" and "Offense + Yellow card" were taken. Certain referees do not award a free-kick when the defender plays the ball. However, other referees believe that even if the defender plays the ball, he disregards the danger to, or consequences for, an opponent and awards a yellow card. These findings underscore the complexity and subjectivity inherent in refereeing decisions, highlighting the potential for further research to improve consistency and fairness in officiating.

## 5 Conclusion

Distinguishing between a foul and no foul and determining its severity is a complex and subjective task relying on individual interpretation of the *Laws of the Game* [16]. Despite the challenges posed by this complex task and an unbalanced training dataset, our solution demonstrates promising results. While we have not reached human-level performance yet, we believe that VARS holds the potential to assist and support referees across all levels of professionalism in the future.

**Acknowledgments.** 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, Belgium.

## References

1. Cabado, B., Cioppa, A., Giancola, S., Villa, A., Guijarro-Berdiñas, B., Padrón, E.J., Ghanem, B., Van Droogenbroeck, M.: Beyond the Premier: Assessing action spotting transfer capability across diverse domains. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW). vol. 27, pp. 3386–3398 (Jun 2024)
2. Cioppa, A., Deliège, A., Giancola, S., Ghanem, B., Van Droogenbroeck, M., Gade, R., Moeslund, T.B.: A context-aware loss function for action spotting in soccer videos. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 13123–13133 (Jun 2020)
3. Cioppa, A., Deliège, A., Ul Huda, N., Gade, R., Van Droogenbroeck, M., Moeslund, T.B.: Multimodal and multiview distillation for real-time player detection on a football field. In: IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports. pp. 3846–3855. Seattle, WA, USA (Jun 2020)
4. Cioppa, A., Giancola, S., Deliege, A., Kang, L., Zhou, X., Cheng, Z., Ghanem, B., Van Droogenbroeck, M.: SoccerNet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In: IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports. pp. 3490–3501 (Jun 2022)
5. Cioppa, A., Giancola, S., Somers, V., Magera, F., Zhou, X., Mkhallati, H., Deliège, A., Held, J., Hinojosa, C., Mansourian, A.M., Miralles, P., Barnich, O., De Vleeschouwer, C., Alahi, A., Ghanem, B., Van Droogenbroeck, M., et al.: SoccerNet 2023 challenges results. Sports Eng. **27**(2), 1–18 (July 2024)
6. De Dios Crespo, J.: The Contribution of VARs to Fairness in Sport, chap. 2, pp. 23–35. Routledge, New York City, NY, USA (Jun 2021)
7. Deutscher Fußball-Bund (DFB): Anzahl aktiver schiedsrichter/-innen bis 2022. https://www.dfb.de/verbandsstruktur/dfb-zentrale/ (2022)
8. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 6804–6815 (Oct 2021)
9. Gautam, S., Sarkhoosh, M.H., Held, J., Midoglu, C., Cioppa, A., Giancola, S., Thambawita, V., Riegler, M.A., Halvorsen, P., Shah, M.: SoccerNet-echoes: A soccer game audio commentary dataset. Int. Symp. Multimedia (ISM) pp. 71–78 (Dec 2024)
10. Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: SoccerNet: A scalable dataset for action spotting in soccer videos. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW). pp. 1792–179210 (Jun 2018)
11. Giancola, S., Cioppa, A., Georgieva, J., Billingham, J., Serner, A., Peek, K., Ghanem, B., Van Droogenbroeck, M.: Towards active learning for action spotting in association football videos. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW). pp. 5098–5108 (Jun 2023)
12. Giancola, S., Ghanem, B.: Temporally-aware feature pooling for action spotting in soccer broadcasts. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW). pp. 4485–4494 (Jun 2021)
13. Held, J., Cioppa, A., Giancola, S., Hamdi, A., Ghanem, B., Van Droogenbroeck, M.: VARS: Video assistant referee system for automated soccer decision making from multiple views. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW). pp. 5086–5097 (Jun 2023)
14. Held, J., Itani, H., Cioppa, A., Giancola, S., Ghanem, B., Van Droogenbroeck, M.: X-VARS: Introducing explainability in football refereeing with multi-modal large language models. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW). vol. 9, pp. 3267–3279 (Jun 2024)

15. Holder, U., Ehrmann, T., König, A.: Monitoring experts: insights from the introduction of video assistant referee (VAR) in elite football. Journal of Business Economics **92**(2), 285–308 (Aug 2021)
16. IFAB: Laws of the game. Tech. rep., The International Football Association Board, Zurich, Switzerland (2022)
17. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset. arXiv **abs/1705.06950** (2017)
18. Leduc, A., Cioppa, A., Giancola, S., Ghanem, B., Van Droogenbroeck, M.: SoccerNet-Depth: a scalable dataset for monocular depth estimation in sports videos. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW). vol. 12, pp. 3280–3282 (Jun 2024)
19. Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., Feichtenhofer, C.: MViTv2: Improved multiscale vision transformers for classification and detection. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR). pp. 4794–4804 (Jun 2022)
20. López, A.M.: Average time needed for a video assistant referee (VAR) intervention in Brazil in 2019 and 2020. https://www.statista.com/statistics/1010093/average-time-video-assistant-referee-checking-brazil/ (Feb 2023)
21. MacMahon, C., Helsen, W.F., Starkes, J.L., Weston, M.: Decision-making skills and deliberate practice in elite association football referees. Journal of Sports Sciences **25**(1), 65–78 (Jan 2007)
22. Magera, F., Hoyoux, T., Barnich, O., Van Droogenbroeck, M.: A universal protocol to benchmark camera calibration for sports. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW). pp. 3335–3346 (Jun 2024)
23. Mkhallati, H., Cioppa, A., Giancola, S., Ghanem, B., Van Droogenbroeck, M.: SoccerNet-caption: Dense video captioning for soccer broadcasts commentaries. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW). pp. 5074–5085 (Jun 2023)
24. Somers, V., De Vleeschouwer, C., Alahi, A.: Body part-based representation learning for occluded person Re-Identification. In: IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV). pp. 1613–1623 (Jan 2023)
25. Somers, V., Joos, V., Cioppa, A., Giancola, S., Ghasemzadeh, S.A., Magera, F., Standaert, B., Mansourian, A.M., Zhou, X., Kasaei, S., Ghanem, B., Alahi, A., Van Droogenbroeck, M., De Vleeschouwer, C.: SoccerNet game state reconstruction: End-to-end athlete tracking and identification on a minimap. In: IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW). pp. 3293–3305 (Jun 2024)
26. Spitz, J., Put, K., Wagemans, J., Williams, A.M., Helsen, W.F.: Visual search behaviors of association football referees during assessment of foul play situations. Cognitive Research: Principles and Implications **1**(1) (Oct 2016)
27. Vandeghen, R., Cioppa, A., Van Droogenbroeck, M.: Semi-supervised training to improve player and ball detection in soccer. In: IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports. pp. 3480–3489 (Jun 2022)
28. Yang, Z., Wang, L.: Learning relationships for multi-view 3D object recognition. In: IEEE Int. Conf. Comput. Vis. (ICCV). pp. 7504–7513 (Oct 2019)
29. Zeppenfeld, B.: Anzahl aktiver schiedsrichter / schiedsrichterinnen des deutschen fußball bundes (DFB) von 2018/2019 bis 2022/2023. https://de.statista.com/statistik/daten/studie/1243626/umfrage/dfb-anzahl-aktiver-schiedsrichter/ (Jun 2023)