arXiv:submit/5893184 [cs.CV] 2 Oct 2024

Chapter 1

Deep learning for action spotting in association football videos

Silvio Giancola[1*], Anthony Cioppa[2*], Bernard Ghanem[1*], and
Marc Van Droogenbroeck[2*]

[1] *Center of Excellence for Generative AI, IVUL, KAUST, Saudi Arabia*
[2] *Montefiore Institute, Open-SportsLab, University of Liège, Belgium*

The task of *action spotting* consists in both identifying actions and precisely localizing them in time with a single timestamp in long, untrimmed video streams. Automatically extracting those actions is crucial for many sports applications, including sports analytics to produce extended statistics on game actions, coaching to provide support to video analysts, or fan engagement to automatically overlay content in the broadcast when specific actions occur. However, before 2018, no large-scale datasets for action spotting in sports were publicly available, which impeded benchmarking action spotting methods. In response, our team built the largest dataset and the most comprehensive benchmarks for sports video understanding, under the umbrella of *SoccerNet*. Particularly, our dataset contains a subset specifically dedicated to action spotting, called *SoccerNet Action Spotting*, containing more than 550 complete broadcast games annotated with almost all types of actions that can occur in a football game. This dataset is tailored to develop methods for automatic spotting of actions of interest, including deep learning approaches, by providing a large amount of manually annotated actions. To engage with the scientific community, the SoccerNet initiative organizes yearly challenges, during which participants from all around the world compete to achieve state-of-the-art performances. Thanks to our dataset and challenges, more than 60 methods were developed or published over the past five years, improving on the first baselines and making action spotting a viable option for the sports industry. This paper traces the history of action spotting in sports, from the creation of the task back in 2018, to the role it plays today in research and the sports industry.

1. Introduction

Sports video understanding has become an increasingly active research field[1–3] as the sports industry continues to grow into a major global entertainment sector. In 2023, the global sports market generated over 160$ billion in annual revenue for sports equipment alone, while the betting part peaked at 242$ billion.[4] Furthermore, new sectors such as eSport also open the potential for the market to

---

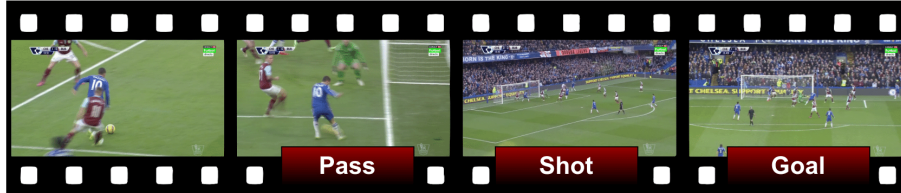*All authors contributed equally.

Fig. 1.: **Action spotting in football videos.** The task of *action spotting* consists in both identifying actions and precisely localizing them in time with a single timestamp in long, untrimmed video streams.

grow even more in the coming years.[5] Particularly, the rapid expansion in the volume of sports TV broadcasts has contributed to this demand, with the number of hours broadcasted in the United States alone growing more than fourfold between 2002 and 2017.[6] This surge in content has created a pressing need for automated analysis tools, as the manual review of video footage is both time-consuming and labor-intensive. Among various sports, football stands out as one of the most universally followed, with thousands of professional matches and millions of amateur games played annually. However, its dynamic and complex nature presents significant challenges for automatic analysis since much of the fine-grained understanding lies in subtle details. In response to these challenges, the scientific community has made remarkable strides over the past decade, tackling tasks such as player tracking[7,8] for individual performance evaluation, camera shot selection[9] and dense video captioning[10] for enhancing broadcast production, and camera calibration[11,12] for extracting physical metrics like speed and distance.

Action spotting focuses on identifying specific actions and precisely localizing them within long, untrimmed video streams using a single timestamp, marking the exact moment an action occurs,[9,13–15] as illustrated in Fig. 1. This approach contrasts with traditional temporal video understanding,[16] where actions are defined by a start and end time. In sports, however, these boundaries can often be ambiguous. For instance, determining when a goal "starts" and "ends" is blurry. *Does a goal begin with a pass or the shot itself? Does it end with the ball hitting the net or after the kick-off?* To address these ambiguities, Giancola *et al.*[13] introduced the novel task of action spotting in sports through the first version of the SoccerNet dataset. In this task, actions are annotated based on official football rules defined by the *IFAB Laws of the Game*,[17] such as the goal being marked at the precise moment the ball crosses the line or a free-kick occurring when the player strikes the ball. This focus on precise localization differs from temporal activity localization, where defining the exact boundaries of activities is less critical.[16]

Action spotting opens up a new area of research with a focus on precisely localizing actions, especially given the challenge of overlapping actions that may occur simultaneously. However, this task also brings inherent difficulties, including the sparsity of annotations since most moments in a match lack notable actions, and

the discontinuities between adjacent frames that may be visually similar while representing different actions.

In sports, action spotting offers significant value across multiple verticals. First, it plays a crucial role in *team strategy analysis*, by identifying key moments such as scoring opportunities, including corners or free kicks. This enables video analysts to efficiently extract and review historical data, gaining insights into how teams perform under specific circumstances. Second, action spotting contributes to *fairer refereeing*, especially by localizing and characterizing potential foul moments. This provides referees and VAR officials with evidence to ensure consistency and fairness throughout matches.[18,19] Third, action spotting *enhances broadcasting experiences*, allowing broadcasters to personalize highlights based on a viewer's interests or time constraints[20,21] by automatically retrieving the actions of interest, or simply identifying the most salient moments.

Fourth, it benefits *player scouting*, enabling teams to identify specific talents, such as a player skilled at scoring from headers during corner kicks. Automatically analyzing thousands of games therefore helps discover players with specialized abilities for particular sequences of actions. Lastly, sports *medical analysis* benefits from spotting actions that could impact player health, such as headers.[22] By examining historical data, long-term effects of these actions can be evaluated. Likewise, action spotting can also help coaches and medical staff in real time by identifying potential health risks for players during games due to repeated dangerous actions.

Deep learning has demonstrated impressive performance across a wide range of tasks, from image classification to video understanding.[23–26] However, these methods typically require large amounts of annotated data to train. To meet these data demands, deep learning models often rely on generic learning algorithms[27] trained on publicly available datasets[28,29] or fine-tuned algorithms trained on smaller, task-specific football datasets.[30] Before the introduction of SoccerNet,[13] most datasets related to association football were either small, scattered, or private, making it difficult to fairly compare different methods and slowing the progress of scientific research in the field. Furthermore, many publicly available datasets suffered from size limitations, restricting their ability to generalize across different games.[31,32] The football research community, therefore, faced a pressing need for large-scale, publicly accessible datasets to address these challenges.[33–37] Nevertheless, creating such datasets is both time-consuming and expensive, as annotating large amounts of video data requires significant resources.

In response to the need for large-scale, publicly accessible datasets in sports video understanding, Giancola *et al.*[13] introduced SoccerNet in 2018. The goal was to create an open-source dataset for reproducible research and benchmarking in football video analysis. The original dataset comprised 500 full broadcast football matches, totaling over 764 hours of footage from the six major European leagues: Serie A, La Liga, Premier League, Ligue 1, Bundesliga, and Champions League, spanning from 2014 to 2017. With this dataset, Giancola *et al.* introduced the novel

task of action spotting, focusing on the temporal localization of three key football actions: goals, cards, and substitutions. Over the years, SoccerNet has continuously expanded both in scope and content. SoccerNet-v2, presented by Deliège *et al.*[9] significantly increased the number of annotations and action classes, growing to include 110,458 actions across 17 classes, such as penalties, clearances, and ball out of play. The dataset also introduced annotations for camera shot changes, replay segments, and more, leading to new tasks like camera shot segmentation, boundary detection, and replay grounding.

The first SoccerNet challenge, organized in 2021, focused on action spotting and replay grounding, starting the dataset's interest in the sports community. In 2022, SoccerNet-v3[11] and SoccerNet-Tracking[7] were introduced, adding spatial annotations for players, the ball, and field elements across multiple views, alongside tasks such as pitch localization, camera calibration, player re-identification, and long-term multi-object tracking. In 2023, SoccerNet continued to expand with the addition of SoccerNet-Captions[10] for dense video captioning, SoccerNet-MVFouls[18] for multi-view foul recognition, and SoccerNet-BallActionSpotting[38] which identified first 2, then 12 different ball-related actions such as passes, drives, and headers. By 2024, further developments included SoccerNet-Depth[39] for monocular depth estimation, SoccerNet-XFoul[19] for multi-modal question-answer triplets about refereeing decisions, and SoccerNet-GSR[8] as the first open-source dataset for game state reconstruction.

Thanks to our dataset and challenges, more than 60 methods have been developed or published over the past five years for action spotting alone, improving on the initial baselines by Giancola *et al.*[13] and Cioppa *et al.*,[40] and making action spotting a viable option for the sports industry. To improve accessibility of action spotting for researchers and practitioners, we also published an open-source library called *OSL-Action Spotting*[41] that gathers several action spotting dataloaders, methods, and evaluation functions, under a common framework.

The remainder of this chapter is organized as follows. In Sec. 2, we mathematically formalize the task of action spotting. Section 3 provides a detailed description of the SoccerNet Action Spotting datasets, explaining their evolution and the different versions dedicated to action spotting. Section 4 presents the various methods developed for action spotting, highlighting key approaches and advancements over the years. In Sec. 5, we discuss the evaluation protocols and metrics used to assess the performance of action spotting methods. Section 6 offers a comprehensive benchmark, comparing the performance of state-of-the-art methods on the Soccer-Net datasets. Finally, in Sec. 7, we conclude the chapter by summarizing the key findings for action spotting in both research and industry.

2. Action Spotting: Definition of the Task

Action spotting is defined as the task of identifying and precisely localizing actions in time, *i.e.*, with a single timestamp, in long, untrimmed video streams. Given a

set of N untrimmed videos $\mathcal{V} = \{v^1, v^2, ..., v^n, ..., v^N\}$, the objective is to extract the set of all actions $\mathcal{A}^n = \{a_1^n, a_2^n, ..., a_k^n, , ..., a_{A^n}^n\}$ for each video n, where A^n is the total number of actions of interest in video n. All actions a_k^n are defined by two features: (1) a class c and (2) a timestamp t, indicating respectively the type of action and the exact moment the action happens in the video. Each class c is chosen among a set $\mathcal{C}$ of size C, corresponding to all possible classes of interest, while the timestamp is bounded by $t \in [0, L]$, typically expressed in milliseconds. From a practical perspective, the videos are generally represented as a succession of frames $v^n = \{i_1^n, i_2^n, ..., i_j^n, ..., i_{J^n}^n\}$, where J^n is the number of frames in video n. This process discretizes the possible values of the timestamps of each action following: $j = \lfloor \frac{t}{T} + \frac{1}{2} \rfloor + 1$, with T being the time between two frames, $j \in \{1, 2, ..., J^n\}$, and $\lfloor x + \frac{1}{2} \rfloor$ being the $round(x)$ function.

3. SoccerNet: Large-Scale Datasets for Action Spotting

Recently, the evolution of sports video analysis has demanded large-scale, richly annotated datasets to fuel research and enable the development of robust data-driven models. Recognizing the lack of such resources, we introduced comprehensive datasets dedicated to action spotting in football videos. Since its inception, Soccer-Net has become a cornerstone dataset for sports video understanding, growing in both scale and complexity over multiple iterations. Table 1 provides a summary of the datasets released for action spotting under the SoccerNet umbrella.

Table 1.: List of Action Spotting datasets published under the SoccerNet umbrella.

SoccerNet Dataset	year	# classes	# games	# annot.
SoccerNet Action Spotting (v1)[13]	2018	3	500	6,637
SoccerNet Action Spotting (v2)[9]	2021	17	500+50	110,458
SoccerNet Ball Action Spotting (2023)[38]	2023	2	7+2	11,041
SoccerNet Ball Action Spotting (2024)[42]	2024	12	7+2	11,041

3.1. *SoccerNet Action Spotting (v1): the foundation*

The original release of SoccerNet Action Spotting (v1), introduced by Giancola *et al.*[13] in 2018, was a pioneering dataset in the field of sports video analysis. The dataset contained 500 full broadcast football matches, totaling 764 hours of footage sourced from major European leagues like Serie A, Premier League, La Liga, Bundesliga, Ligue 1, and the UEFA Champions League. SoccerNet Action Spotting (v1) primarily focused on three main action classes for the task of action spotting: *Goals*, *Cards*, and *Substitutions*. Each of these actions was annotated with a timestamp marking the moment the action occurred with a 1-second precision. Different from traditional temporal activity localization tasks,[16] the actions are localized with a

single timestamp. The simplicity of this approach allowed researchers to get their hands on developing methods for spotting actions in football games, laying the groundwork for more complex future developments.

SoccerNet Action Spotting (v1) became a cornerstone dataset and benchmark in the domain of sports video analysis, enabling the comparison of various methods under standardized conditions. The dataset also provided pre-extracted features from the video frames, based on ResNET,[43] C3D[44] and I3D,[45] to ease computational demands, which was crucial at the time due to hardware limitations.

3.2. *SoccerNet Action Spotting (v2): increasing class diversity*

Building on the success of SoccerNet Action Spotting (v1), SoccerNet Action Spotting (v2) was introduced by Deliège *et al.*[9] in 2021. This version significantly expanded the dataset, both in terms of action classes and annotation complexity. Specifically, SoccerNet Action Spotting (v2) includes annotations for 17 different action classes: *Penalty, Kickoff, Goal, Substitution, Offside, Shot on target, Shot off target, Clearance, Ball out of play, Throw-in, Foul, Indirect free-kick, Direct free-kick, Corner, Yellow card, Red card,* and *Yellow to red card.*

The total number of annotated actions grew to 110,458 across the dataset, now annotated with 1-frame precision, offering a more comprehensive representation of football events. In addition to increasing the variety of actions, SoccerNet Action Spotting (v2) introduced new tasks, such as *replay grounding*, which consists in localizing the live action shown in a replay clip in a long untrimmed broadcast, and *camera shot segmentation*, which consists in temporally segmenting the different camera shots during a broadcast. These additions opened up new avenues for research in other areas, such as broadcast production, automatic video summarization, and foul analysis.[18,19,46]

This dataset also served in the first SoccerNet Action Spotting challenge in 2021, pushing the research community to develop state-of-the-art models and benchmark their methods for sports video analysis. Through this challenge, SoccerNet Action Spotting (v2) became a pivotal tool for advancing temporal spotting tasks in complex, real-world football scenarios.

3.3. *SoccerNet Ball Action Spotting (2023): towards more precise localization*

In 2023, SoccerNet introduced the Ball Action Spotting task.[38] Unlike the broader action spotting task, which targets major game-changing events such as goals or cards, Ball Action Spotting requires models to precisely localize frequent interactions with the ball, offering a finer granularity of analysis. This task added a significant layer of complexity, as it required models to detect not just prominent actions, but also subtle and rapid ball exchanges that are critical to the game.

The Ball Action Spotting task was introduced during the SoccerNet challenges

2023[38] and only focuses on 2 key actions: *pass* defined as the moment the ball leaves the feet of a player, who is passing the ball, and *drive*, defined as the moment the ball is received in the feet of another player, who is maintaining the control of the ball while moving through the field. Though these actions are frequent and subtle compared to more prominent events like goals or fouls, they are essential for understanding team strategies, game dynamics, and individual player performance.

The SoccerNet Ball Action Spotting dataset comprises 7 full games from the English football league with a total of 11,041 annotated timestamps, providing data for training and validation. Additionally, 2 extra games were held out as a separate evaluation set for the SoccerNet challenges 2023. This challenge pushed the research community to develop models capable of handling the increased temporal precision required for spotting ball actions, laying the groundwork for the future expansion of the task to incorporate more complex and diverse ball-related actions.

3.4. *SoccerNet Ball Action Spotting (2024): precise localization of fine-grained actions*

In 2024, the Ball Action Spotting task was further expanded to introduce more granular ball-related events, moving from just two action classes (pass and drive) to a more comprehensive set of 12 different ball-centric actions: *Pass, Drive, Header, High Pass, Out, Cross, Throw In, Shot, Ball Player Block, Player Successful Tackle, Free Kick,* and *Goal.* This extension significantly increases the task's complexity, providing researchers with an opportunity to develop models that can handle a wider range of interactions between players and the ball while maintaining precise temporal localization.

This expansion to 12 classes in SoccerNet Ball Action Spotting (2024) reflects a growing emphasis on ball interactions throughout a match, with each action class capturing specific, frequent events that are useful for understanding the flow of the game. The task highlights the diversity of ball interactions and underscores the necessity of precise temporal localization. Additionally, the 2024 version of the SoccerNet challenge introduces new difficulties, such as distinguishing between similar events like passes, high passes, or crosses. This requires a fine-grained understanding of both the visual and contextual aspects of the game, setting higher standards for model performance, and pushing for advancements in temporal detection and sequence modeling within sports video analysis.

The SoccerNet Ball Action Spotting (2024) dataset is composed of the same 7 games used in the 2023 version, and the same 11,041 annotated timestamps, but now extended with the 12 classes instead of 2. Similarly, the challenge set comprises 2 extra games that are withheld for benchmarking and challenge purposes.
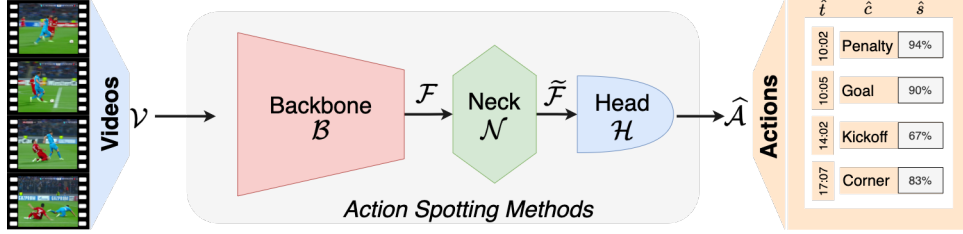
Fig. 2.: **Generic action spotting method pipeline.** A backbone $\mathcal{B}$ processes the videos $\mathcal{V}$ to extract features $\mathcal{F}$, refined by the neck $\mathcal{N}$ into action spotting discriminating features $\widetilde{\mathcal{F}}$ before being fed into an action spotting head $\mathcal{H}$ that produces action spotting predictions $\widehat{\mathcal{A}}$, each composed of a timestamp $\widehat{t}$, a class $\widehat{c}$, and a confidence score $\widehat{s}$.

3.5. *Summary*

SoccerNet has evolved into the largest and most comprehensive dataset for action spotting, enabling the sports video analysis community to develop and benchmark action spotting models. Over multiple iterations, it has expanded from basic action localization to fine-grained ball action spotting, driving innovation in precise temporal localization and contextual understanding of actions in association football videos. The continued growth of the SoccerNet challenges ensures that research in this field has continuously advanced each year, addressing increasingly complex aspects of sports video analysis.

4. Action Spotting Methods

Formally, an action spotting method $\mathcal{M}$ is defined as a function that maps a video v^n to a set $\widehat{\mathcal{A}^n} = \{\widehat{a_1^n}, \widehat{a_2^n}, ..., \widehat{a_k^n}, ..., \widehat{a_{\widehat{A^n}}^n}\}$ of $\widehat{A^n}$ predicted actions, following $\mathcal{M} : v^n \rightarrow \widehat{\mathcal{A}^n}$. Each predicted action $\widehat{a_k^n}$ is a triplet composed of a predicted action class $\widehat{c} \in \mathcal{C}$, a timestamp $\widehat{t} \in [0, L]$, alongside an optional confidence score $\widehat{s} \in [0, 1]$ with a default value of 1. Due to the discrete nature of videos, some methods will predict the timestamp in terms of frame number rather than milliseconds, following $\widehat{j} \in \{1, 2, ..., J^n\}$. Nevertheless, for evaluation purposes, the prediction is transformed back in millisecond following: $\widehat{t} = (\widehat{j} - 1) \times T$, with T being the time between two frames.

Typical action spotting methods consist of three primary components: the backbone $\mathcal{B}$, the neck $\mathcal{N}$, and the head $\mathcal{H}$, as illustrated in Figure 2. We detail the role of each component hereafter. The *backbone* serves as the core component for feature extraction. Given a video v^n composed of frames $\{i_1^n, i_2^n, ..., i_j^n, ..., i_{J^n}^n\}$, the backbone processes each frame and extracts a set of features $\mathcal{F}^n = \{f_1^n, f_2^n, ..., f_j^n, ..., f_{J^n}^n\}$, following $\mathcal{B} : v^n \rightarrow \mathcal{F}^n$. These features form the foundation of the action spotting representation, capturing the essential visual

information from the video. Following the backbone, the *neck* acts as an intermediary layer that refines and transforms the raw features $\mathcal{F}^n$ into a more discriminative form $\widetilde{\mathcal{F}^n}$ for action spotting and adapts the feature sizes to accommodate the input format of the head following $\mathcal{N} : \mathcal{F}^n \rightarrow \widetilde{\mathcal{F}^n}$. These refined features $\widetilde{\mathcal{F}^n}$ are then passed to the head. The *head* is the final stage of the algorithm, responsible for action identification and temporal localization. It uses the refined features $\widetilde{\mathcal{F}^n}$ to identify and localize the actions $\widehat{\mathcal{A}^n}$ following $\mathcal{H} : \widetilde{\mathcal{F}^n} \rightarrow \widehat{\mathcal{A}^n}$. Together, these components form a complete action spotting method $\mathcal{M} = \mathcal{H} \circ \mathcal{N} \circ \mathcal{B}$, where $(g \circ f)(x)$ denotes the composition $g(f(x))$. Some action spotting methods later apply some post-processing on the action set, such as Non-Maxima Suppression (NMS) to remove redundant predictions.

The action spotting methods developed in the last few years can be broadly categorized into *feature-based* and *end-to-end* approaches. In *feature-based* methods, the backbone $\mathcal{B}$ consists of pre-trained features extractor from commonly used classification networks like ResNet,[43] trained on large datasets such as ImageNet.[28] In this case, only the neck $\mathcal{N}$ and head $\mathcal{H}$ components are trained for the action spotting task. This approach is computationally efficient, but the action spotting performance is highly dependent on the quality of the pre-extracted features for discriminating actions. In contrast, *end-to-end methods* train the entire architecture, including the backbone $\mathcal{B}$, neck $\mathcal{N}$, and head $\mathcal{H}$, directly on the action spotting task. These methods often provide better performance since the backbone features are optimized for spotting actions on sports videos. However, they generally require significantly larger amounts of data and computational resources to be trained compared to feature-based methods.

In the following, we first describe several historical *feature-based methods* that utilize pre-trained backbones such as ResNet[43] to extract visual features before training task-specific neck and head components for action spotting. Following this, we explore more recent *end-to-end methods*, which jointly aim to optimize the backbone, neck, and head components. Through this progression, we illustrate the evolution of methods proposed for action spotting.

4.1. *Feature-based methods*

NetVLAD: learnable pooling layers.[13]
The NetVLAD method was introduced as the baseline for the action spotting task in SoccerNet-v1,[13] leveraging the popular NetVLAD[47] layer originally designed for image retrieval. NetVLAD is particularly effective for summarizing large amounts of visual data into compact representations, which made it a suitable choice for the SoccerNet Action Spotting (v1) dataset, where the task was to identify key moments such as goals, cards, and substitutions within long, untrimmed video sequences. The NetVLAD approach can be described as a feature-based method, where the backbone is pre-trained on large image datasets, and the focus is placed

on summarizing video frames into discriminative representations for action spotting.

In NetVLAD, the backbone consists of pre-extracted visual features from a ResNet model,[43] which was trained on the ImageNet dataset.[28] Instead of training the backbone on football videos directly, the features are extracted frame-by-frame from the input video. This approach significantly reduces computational overhead, as it avoids the need for feature extraction during training and testing. The features are eventually reduced in dimension through Principal Component Analysis (PCA) and then fed into the NetVLAD module, which acts as the neck of the model, aggregating the extracted frame-level features into a fixed-length representation. Instead of focusing on exact temporal localization, NetVLAD aggregates temporal information over a sampled window of video frames. The process involves: (i) extracting a temporal video clip from the untrimmed video in a sliding window fashion, (ii) assigning each frame-level feature to a set of learnable cluster centers, (iii) computing the residual between each feature and its assigned center, (iv) summing up these residuals to produce a compact, video clip-level representation. This process condenses the temporal information from a candidate video clip into a fixed-dimensional vector, which aggregates the temporally local receptive field. The head focuses solely on action classification. Once the feature representation pooled from a video clip is obtained, the head applies a multi-class classification layer to determine whether any of the predefined action classes (*e.g.*, goal, card, substitution) occurs within the clip. Since the model samples video clips from the untrimmed video in a sliding window fashion, it classifies each window without directly predicting the exact timestamp of the actions. Instead, the method outputs whether a particular action occurs within the sampled window. To manage the untrimmed and lengthy nature of football videos, NetVLAD applies a sliding window approach, producing an *actionness score* in time, *i.e.*, how likely the window represents an action of a certain class. In practice, the video is divided into overlapping temporal windows, and the model classifies each window independently. After classification, Non-Maxima Suppression (NMS) is used over the actionness score in time to refine the predictions by eliminating redundant detections across the overlapping windows. This ensures that only the most confident predictions are retained, reducing false positives and providing a cleaner set of action predictions.

The NetVLAD method served as an important baseline for the first version of SoccerNet. While it does not explicitly perform explicit temporal localization, it provided a practical and efficient framework for action spotting following a classification approach within a sliding window, demonstrating that pre-extracted features combined with temporal pooling could identify key moments in football matches to some extent. Its simplicity and computational efficiency made it a popular choice for early research, though it left room for the development of more advanced methods that could explicitly handle temporal localization and end-to-end training.

NetVLAD++: temporally-aware pooling.[48]

NetVLAD++ was introduced as an improvement over the original NetVLAD

method for action spotting, addressing a key limitation: the lack of temporal awareness regarding the context surrounding the actions in football broadcasts.[48] In NetVLAD, all frames in a video clip are treated equally, without distinguishing the context before or after an action. However, for many football events, the preceding and succeeding moments can provide different contextual cues for identifying actions. For instance, spotting a goal from a shot is ambiguous until the succeeding context is revealed. NetVLAD++ introduces a more sophisticated pooling strategy by learning separate clusters for the frames before and after the action, effectively making the model aware of temporal dynamics.

Like its predecessor, NetVLAD++ utilizes pre-extracted frame-level features from a ResNet encoder trained on the ImageNet dataset. The features are extracted from each frame of the video, serving as the raw input to the NetVLAD++ module. This keeps the model computationally efficient by focusing on feature aggregation rather than training a new backbone from scratch. The primary innovation in NetVLAD++ lies in its temporally-aware pooling mechanism. Rather than pooling features uniformly across the video clip, NetVLAD++ splits the candidate video clip from the sliding window into 2 parts: (i) the frames leading up to the action and (ii) the frames following the action. For each of these regions, NetVLAD++ learns separate NetVLAD clusters, allowing the model to capture the different types of information conveyed in each temporal segment. This separation helps the model focus on the action's immediate surroundings, which are often indicative of its occurrence. Following our goal example, the buildup (context before) and aftermath (context after) provide different information that complements the action itself. The NetVLAD++ layer thus pools features in a temporally-aware manner, creating separate, discriminative representations for the context before and after the action. This improvement allows the model to be more sensitive to the nuances of football events that often rely on contextual buildup or consequences. Once the temporally-aware features are pooled, the head of the model performs the same action classification along sliding windows with the Non-Maxima Suppression (NMS), refining the action spots.

The addition of temporally-aware feature pooling in NetVLAD++ significantly improves the performance of action spotting, particularly in cases where the context plays a crucial role in recognizing actions. By learning separate clusters for the pre-action and post-action frames, NetVLAD++ surpasses the performances of the original NetVLAD in detecting actions. This refinement paved the way for more advanced temporal modeling techniques and remains an influential approach in sports video understanding.

CALF: a context-aware loss function for action spotting.[40]

The CALF (Context-Aware Loss Function) method, developed by Cioppa *et al.*[40] introduces a novel approach for action spotting by focusing on the temporal context around actions. The two main contributions are (1) the direct regression of the timestamp value rather than classifying each video chunk followed by a NMS

post-processing as previously done in NetVLAD, and (2) the introduction of a context-aware loss function, which weights differently the loss associated to the frames surrounding a ground-truth action during training. The method considers overlapping 2-minutes video clips rather than ingesting the whole untrimmed video.

For the backbone, CALF uses frozen pre-trained feature extractors, like ResNet, to process each video frame and reduce the input dimensionality. These extractors, originally trained on large datasets such as ImageNet, output a set of features that encapsulate the visual information about each frame. Since the backbone is frozen, the features are extracted once and then stored as matrices with two dimensions: time and number of features. Following NetVLAD, the feature dimension is reduced through Principal Component Analysis (PCA) by a factor of four. The neck component in CALF is implemented as a trainable temporal convolutional neural network (CNN) that processes the features extracted by the backbone. This temporal CNN consists of multiscale convolutions that operate across time and feature dimensions. The purpose of this component is to fine-tune the raw spatial features by incorporating temporal dependencies across frames. In CALF, the head is composed of two modules. The first module is a segmentation head that predicts an actionness score for each frame, similar to NetVLAD. The segmentation head is guided during training by the context-aware loss function, which assigns a weight to the cross-entropy loss of each frame depending on the closest ground-truth action location. Specifically, if a frame is located just after an action, CALF assumes that there are still enough visual cues that the action occurred for the segmentation head to predict an action, with a smaller weight the further the frame is from the action. Just before the action, it is unsure whether the action will occur or not, so the weight is set to zero, letting the segmentation module decide freely whether to predict an action. Finally, the further away from the action, the more the segmentation head should predict that no action occurred. This loss therefore alleviates the issue of two similar consecutive frames needing to predict different actions. The predictions of the segmentation module are then concatenated with the neck features and passed to the second module, the spotting head. This head is composed of classical convolutions and max pooling operations to reduce dimensionality and format the output following a YOLO-like[49] encoding. Specifically, each video clip may predict a finite set of actions with a confidence score, a regressed timestamp representing the relative location of the action within the video clip, and the action class through a softmax activation. The ground-truth and predicted actions are associated through an iterative one-to-one matching algorithm. Successively, a mean squared error loss is operated on the confidence scores and timestamps, and a binary cross entropy on the class.

This method significantly improved upon NetVLAD on the first action spotting task of SoccerNet and the context-aware loss function has been successfully applied on top of other temporal activity localization models such as BSN[50] trained on the ActivityNet[16] dataset. The context-aware loss also allows for the discovery of

actions that are similar to actions of interest, such as goal tentatives thanks to the allowed doubt introduced before the action occurs.

Baidu features and transformers.[51]

Zhou *et al.*[51] propose a novel approach for the two-stage framework of action spotting, integrating fine-tuned backbones with transformer-based temporal detection necks and heads. Unlike previous methods like NetVLAD++ and CALF, which used generic ResNet features, this method fine-tunes multiple action recognition models specifically for football videos, ensuring improved performance in detecting key moments with greater precision.

The backbone $\mathcal{B}$ is a mixture-of-model between the encoders TPN,[52] GTA,[53] VTN,[54] irCSN,[55] and I3D-Slow,[56] fine-tuned on 5-second snippets from the Soccer-Net Action Spotting (v2) dataset, capturing the 17 classes of football events. This fine-tuning tailors the feature extractors to football dynamics, improving performance in detecting key actions. These features, extracted from a stack of frames rather than individual frames, are concatenated to provide richer inputs for temporal detection. For the neck $\mathcal{N}$, this method uses a transformer encoder consisting of three encoding layers with sine and cosine positional encodings. Each layer has four heads and a hidden dimension of 64. The transformer processes the combined features from the backbone, learning temporal dependencies between frames. This allows the model to capture subtle yet important contextual information for action spotting. The head $\mathcal{H}$ of the model focuses on action classification by directly predicting the 18 action class probabilities from the transformer-encoded features (17+ background). To refine the output, the method applies hyperparameter tuning, including adjustments to batch size, learning rate, feature dimension size, and the Non-Maxima Suppression (NMS) window size, ensuring that the model selects the most relevant predictions.

This method achieved the best performance in the first SoccerNet challenge in 2021, based on the SoccerNet Action Spotting (v2) dataset, excelling in both action spotting and replay grounding. Its combination of fine-tuned feature extraction and transformer-based temporal modelling raised the standard for action spotting, particularly in handling football's video manifold distribution. It demonstrated that combining deep attention mechanisms with domain-specific features improves action spotting in sports videos.

Dense Detection Anchors: precise action spotting in soccer videos.[57]

The method presented by Soares *et al.*[57] won the SoccerNet challenge 2022, based on the SoccerNet Action Spotting (v2) dataset. They introduce a novel approach to action spotting by focusing on precise localization of actions in football videos using dense detection anchors. This method significantly improves the temporal precision of action spotting compared to previous techniques by generating numerous temporal anchors and refining predictions through dense detection.

As with previous methods, Dense Detection Anchors relies on pre-extracted features from a backbone network $\mathcal{B}$, either the ResNet[43] trained on ImageNet or

Zhou *et al.*[51] features fine-tuned on SoccerNet. These frame-level features provide the input for the detection process. The core innovation of this method lies in the dense temporal anchoring strategy. The video is divided into a set of overlapping temporal segments, and within each segment, the method generates dense temporal anchors at regular intervals. These anchors serve as candidates for potential actions. By densely covering the timeline, the model ensures that no significant action is missed, even in the occurrence of several actions within short durations. For each anchor, features are extracted from the pre-extracted frame-level representations, and the model evaluates whether an action occurs within that temporal window. The neck component applies temporal convolutions to refine the anchor features and enhance their ability to discriminate actions across time. Once the anchor features are refined, the head of the model performs two tasks: (i) action classification: for each anchor, the model predicts the likelihood of an action occurring. The classification head outputs the predicted class for the action (*e.g.*, goal, card, substitution). (ii) temporal regression: in addition to classification, the model also refines the temporal localization of the action by regressing the precise timestamp relative to the anchor. This enables the model to predict the exact moment when the action happens, rather than simply classifying a broad temporal window. The dense anchoring and regression mechanisms ensure high temporal precision, improving the performances of action localization.

The method uses a multi-task loss function that combines classification and regression objectives. The classification loss penalizes incorrect action predictions while encouraging precise temporal localization, ensuring that the model identifies the correct action and pinpoints the exact moment it occurs, minimizing temporal ambiguity. After processing all dense detection anchors, Non-Maxima Suppression (NMS) is applied to remove redundant predictions.

The Dense Detection Anchors method was the top performer in the SoccerNet challenge 2022, significantly improving temporal precision in action spotting. Its high performance in handling complex football actions established it as a state-of-the-art technique for action spotting in sports video analysis.

4.2. *End-to-end methods*

E2E-Spot: spotting temporally precise, fine-grained events in video.[58]

Hong *et al.*[58] introduced the first end-to-end model specifically designed for the action spotting task, named *E2E-Spot*. This method brings several novel contributions to the field, particularly in its ability to reason both globally and locally about actions within a video. One of the key innovations of E2E-Spot is its compact architecture, which allows it to be trained on large temporal windows even with limited GPU memory. Despite its compactness, the method demonstrated impressive performance, securing second place in the SoccerNet challenge 2022 on Action Spotting.[59]

Unlike many of the earlier methods discussed in Section 4.1, which relied on pre-trained and frozen backbones, the backbone of E2E-Spot is fully trainable in an end-to-end manner. The authors explored several lightweight backbone architectures, including ResNet[43] (18 and 50), RegNet-Y[60] (200MF and 800MF), and ConvNeXt.[61] Another significant improvement in their approach was the integration of temporal information directly within the backbone. This was achieved by incorporating TSM[62] (Temporal Shift Module) or GSM[63] (Gate-Shift Module), which allow the model to capture temporal dynamics across frames. By fine-tuning the backbone specifically for the action spotting task, E2E-Spot gains a significant advantage over methods that rely on backbones trained on unrelated datasets.

The neck of the E2E-Spot model is designed to perform long-term temporal reasoning, providing a global understanding of the video sequence. To achieve this, the authors experimented with several architectures for this module, including GRU[64] (Gated Recurrent Unit), deeper GRU, MS-TCN[65] (Multi-Scale Temporal Convolutional Network), and ASFormer.[66] These architectures help the model aggregate information over long sequences, capturing the context necessary to accurately identify and localize actions within the video.

Finally, the head of E2E-Spot is simply a dense frame prediction. For every input frame, the model generates a corresponding output feature and a per-frame prediction. During training, a simple cross-entropy loss is applied to each frame's prediction using a one-hot encoding of the ground-truth action class, along with an extra class for background frames. This straightforward approach ensures that the model is directly optimized for frame-level accuracy. The results demonstrate that direct, end-to-end training of a simple and compact model like E2E-Spot can serve as a surprisingly strong end-to-end baseline for action spotting.

Gray-scale image stacking and transfer learning to cope with limited data.[67]

Ruslan Baikulov's method, which emerged as the winner of the SoccerNet Ball Action Spotting (2023) challenge,[67] showcases several innovative approaches tailored to the specific challenges of detecting ball-related actions. One of the key features of this method is the use of stacked grayscale images. In this approach, three consecutive frames are first converted from RGB to grayscale and then concatenated to form a new three-channel image. This transformation allows the model to focus on motion cues rather than color information, which is particularly advantageous for detecting subtle ball-related events where motion is a critical indicator.

The backbone of Baikulov's method employs a 2D convolutional encoder, EfficientNetV2-B0,[68] to extract per-frame features. These features are then passed to the neck, which uses a 3D convolutional encoder to merge temporal information across consecutive frames, effectively capturing motion dynamics. The merged features are concatenated along the temporal dimension and pooled using Generalized Mean (GeM) pooling to condense the information. Finally, the head consists of a simple linear classifier that predicts the presence and types of ball-related actions

based on the refined features of the neck.

Given the limited amount of data available in the Ball Action Spotting challenge, Baikulov's method leverages transfer learning to maximize performance. The 2D convolutional encoder is initialized with pre-trained weights from ImageNet,[28] while the 3D convolutional encoder and the linear classifier are trained from scratch. Initially, the model is trained on short sequences of 15 frames to manage computational complexity and memory constraints. The model is first pre-trained on the larger and more diverse SoccerNet Action Spotting (v2) dataset, which contains more videos and annotations, though with different action classes. After this initial training, transfer learning is applied by fine-tuning the model on the specific SoccerNet Ball Action Spotting (2023) dataset. Finally, to further enhance performance, the model undergoes fine-tuning on longer sequences of 33 frames, with the 2D backbone frozen to prevent overfitting and to allow the model to focus on longer temporal dynamics.

T-DEED: temporal-discriminability enhancer encoder-decoder for precise event spotting in sports videos.[69]

T-DEED, the winning method of the SoccerNet Ball Action Spotting (2024) challenge,[42] introduced by Artur Xarles *et al.*,[69] tackles several key challenges in action spotting, including improving the discriminability of frame representations, achieving high temporal resolution, and capturing information across multiple temporal scales to handle diverse action dynamics. This method represents a significant architectural advancement over previous approaches.

The backbone of T-DEED is designed to process fixed-length video clips and is built on a 2D RegNetY[60] architecture, augmented with Gate-Shift-Fuse[70] (GSF) modules. This backbone therefore produces spatio-temporal per-frame representations. The neck component of T-DEED consists of a temporally-discriminant encoder-decoder architecture, which leverages Scalable-Granularity Perception[71] (SGP) layers. The authors also include skip connections to preserve information from the initial layers. Additionally, the authors introduce a novel SGP-Mixer layer that accommodates inputs with distinct temporal scales. Finally, the refined features from the neck are passed to a classical action spotting classification head.[57,58,72] This head includes both a classification and displacement head, responsible for predicting the presence of actions and their precise temporal locations within the clip.

The combined architecture of T-DEED, with its focus on multiscale temporal processing and enhanced frame-level discriminability, marks a significant improvement over previous architectures in the domain of precise action spotting.

5. Evaluation and Metrics

An action spotting method $\mathcal{M}$ predicts a set of actions $\widehat{\mathcal{A}^n}$ for each video n. To evaluate the performance of $\mathcal{M}$, the set of predicted actions needs to be compared to the set of ground-truth actions for that video $\mathcal{A}^n$. An oracle method would

therefore satisfy the following condition: $\widehat{\mathcal{A}^n} = \mathcal{A}^n, \forall n \in \{1, ..., N\}$. Giancola *et al.*[13] proposed a first metric, called the average-mean Average Precision (*a-mAP*) which is inspired by classical object detection metrics in PASCAL VOC[73] as well as temporal activity localization metrics from the ActivityNet[16] challenges. This metric and some of its variants are still used nowadays to evaluate action spotting methods, including the SoccerNet challenges. The remaining of this section details the procedure to compute the *a-mAP* and its variants.

The first step consists in associating elements from the set of predicted actions $\widehat{\mathcal{A}^n}$ one-to-one to the set of ground-truth actions $\mathcal{A}^n$ for each video n of the evaluation dataset. To do so, Giancola *et al.*[13] proposed an iterative one-to-one matching algorithm. Since the predicted action set contains confidence scores $\widehat{s}$ associated with each action, they should be considered in the evaluation procedure. The *a-mAP* evaluates a parametric family of action spotting models, depending on a chosen confidence score threshold τ, which is a similar paradigm to what is typically done in the evaluation of object detection models. Therefore, one matching per considered threshold τ is first performed. Following the practice of Everingham *et al.*,[73] 200 threshold values equally spaced between $[0, 1[$ are chosen. For each threshold τ, a single ground-truth action a_k^n may be matched with one predicted action $\widehat{a_{k'}^n}$ only if three conditions are met: (1) the confidence score of that prediction is greater or equal than the threshold, *i.e.*, $\widehat{s} > \tau$, (2) they belong to the same class, *i.e.*, $\widehat{c} = c$, and (3) the prediction falls within a δ time tolerance around the ground-truth action, *i.e.*, $|\widehat{t} - t| \leq 0.5 \times \delta$. If the ground-truth action has several associated predictions, only the prediction with the highest confidence is matched with that ground-truth action. The pair of matched $(a_k^n, \widehat{a_{k'}^n})$ is then removed from the pool of available predicted/ground-truth actions and placed in a set of true-positive pairs $\mathcal{TP}_{\delta,\tau}^c$. If no predictions respect the conditions, the ground-truth action a_k^n is placed in a set of false negative ground-truth action $\mathcal{FN}_{\delta,\tau}^c$ The matching algorithm is then performed on the next ground-truth action until all ground-truth actions are either placed in the $\mathcal{TP}_{\delta,\tau}^c$ or the $\mathcal{FN}_{\delta,\tau}^c$ sets. All remaining unmatched predictions $\widehat{a_{k'}^n}$ are placed in a set of false positive predicted actions $\mathcal{FP}_{\delta,\tau}^c$.

For each class c and time interval δ, a set of Precision-Recall points $\mathcal{PR}_\delta^c$ is constructed by computing the precision and recall value for each τ following:

$$\mathcal{PR}_\delta^c = \left\{ \left(\frac{|\mathcal{TP}_{\delta,\tau}^c|}{|\mathcal{TP}_{\delta,\tau}^c| + |\mathcal{FP}_{\delta,\tau}^c|}, \frac{|\mathcal{TP}_{\delta,\tau}^c|}{|\mathcal{TP}_{\delta,\tau}^c| + |\mathcal{FN}_{\delta,\tau}^c|} \right) | \tau \in \{0, 0.005, 0.01, ..., 0.995\} \right\}.$$

The set of precision-recall points are then summarized in the Average-Precision (AP) metric following the 11-point approximation proposed in the PASCAL VOC challenge.[73] Specifically, this metric computes an approximation of the area under the precision-recall curve following:

$$AP_\delta^c = \frac{1}{11} \sum_{r'=0}^{10} \max_{(p,r)\in\Phi(r')} p \quad \text{with} \quad \Phi(r') = \left\{ (p,r) \in \mathcal{PR}_\delta^c | r \geq \frac{r'}{10} \right\}.$$

Afterward, the AP_δ^c are averaged across all classes into the mean-Average Precision ($mAP@\delta$) metric following:

$$mAP@\delta = \frac{1}{C} \sum_{c=1}^{C} AP_\delta^c \,.$$

This $mAP@\delta$ provides an evaluation of how an action spotting method identifies and localizes actions given a specific time tolerance δ around the ground-truth actions. The smaller the time tolerance, the more precise an action spotting method needs to be. This metric has proven particularly useful in the ball action spotting benchmarks and for ranking methods in the challenges, as actions related to the ball are very localized in time and therefore need to be precisely predicted. Hence, the $mAP@1$ is used to rank ball action spotting methods, meaning that actions need to be localized around the ground-truth actions within a 1-second tolerance.

Finally, Giancola *et al.*[13] propose to summarize the $mAP@\delta$ over a set of δ values into a single metric by computing an approximation of the area under the $(\delta, mAP@\delta)$ curve. To do so, they rely on a trapezoid integral approximation considering Δ time intervals equally spaced between δ_l and δ_h time tolerances. They define this new metric as the Average-mean Average Precision (a-mAP). The general formula to compute the a-mAP is given by:

$$a\text{-}mAP = \frac{1}{\Delta} \sum_{d=1}^{\Delta} \frac{mAP@(\delta_l + (d-1) \times \frac{\delta_h - \delta_l}{\Delta}) + mAP@(\delta_l + d \times \frac{\delta_h - \delta_l}{\Delta})}{2} \,.$$

The first metric introduced by Giancola *et al.*,[13] later called the loose Average-mean Average Precision (a-mAP_{loose}), considered $\Delta = 11$ time intervals of 5 seconds between tight time tolerances ($\delta_l = 5$ seconds) to loose ones ($\delta_h = 60$ seconds). A second stricter variant was later proposed during the SoccerNet Action Spotting 2022 and 2023 challenges[38] considering tighter time tolerances with $\Delta = 4$ time intervals of 1 second linearly spaced between $\delta_l = 1$ and $\delta_h = 5$ second time tolerances, called the a-mAP_{tight}. Let us note that for counting applications, the a-mAP_{loose} or the $mAP@\infty$ are sufficient as predicting the moment when the actions occur is not useful, while for video analysis of specific actions, the a-mAP_{tight} or the $mAP@1$ would be preferred, depending on the time tolerance required for the downstream application. As a final remark, SoccerNet-v2[9] introduced the concept of visible versus unshown actions, which correspond to actions that are respectively clearly visible on the broadcast or that happened but were not shown, *e.g.*, due to a replay or showing a different angle. The metrics described above can therefore be defined for each set separately, for analysis purposes. This is interesting to analyze whether methods can understand the rules of the game and still detect unshown actions based on context alone.

Table 2.: SoccerNet Action Spotting (v1)[13] leaderboard for 3 classes: goal, card, and substitution. The methods are ranked according to the test set a-mAP_{loose} metric. The a-AP_{loose} for each class is also displayed when available.

Methods	a-mAP_{loose}	Test Set a-AP_{loose}		
		Goal	Card	Substitution
Nakazawa *et al.*[74]	**81.6**	**87.1**	**63.3**	**94.3**
RMSNet[75a]	75.1	/	/	/
Si *et al.*[76]	66.8	/	/	/
Karimi *et al.*[77]	64.9	/	/	/
Mahaseni *et al.*[78]	63.3	/	/	/
CALF[13b]	62.5	/	/	/
Vats *et al.*[79]	60.1	/	/	/
Rongved *et al.*[80]	56.3	75.1	40.1	50.21
AudioVid[81]	56.0	/	/	/
NetVLAD[13c]	49.7	/	/	/
Rongved *et al.*[82]	32.0	/	/	/

[a]https://github.com/aimagelab/RMSNet_Soccer
[b]https://github.com/SoccerNet/sn-spotting/tree/main/Benchmarks/CALF
[c]https://github.com/SoccerNet/sn-spotting/tree/main/Benchmarks/Pooling

6. Benchmark Across Four Datasets: Summarizing 6 Years of Research

Since the creation of SoccerNet in 2018, numerous methods have been proposed to tackle the action spotting task, reflecting the evolution of both the dataset and the technological landscape of artificial intelligence. Initially, most methods were feature-based, leveraging the pre-extracted features provided by the SoccerNet Action Spotting (v1) dataset,[13] which focused on three classes of actions. At that time, end-to-end training on video data was challenging due to limitations in GPU VRAM memory and processing speed. As a result, the early methods focused on refining the neck and head components of the pipeline while relying on pre-extracted features from pre-trained backbones like ResNet. The original NetVLAD[13] method set a baseline, but there was a significant gap until the CALF[40] method introduced a context-aware loss function that substantially improved performance. Although several methods attempted to close this gap, it was not until newer architectures, such as attention mechanisms,[74] multi-scene encoders,[76] and metric learning,[77] were introduced a year later that better performances were achieved. Table 2 presents the results of these methods on the original action spotting (3 classes) benchmark.

The release of SoccerNet Action Spotting (v2),[9] which expanded the task to 17 classes, marked a significant step forward and laid the foundation for the first SoccerNet challenge on Action Spotting in 2021. This task provided the research community with a more challenging setup, leading to the development of new tech-

Table 3.: SoccerNet Action Spotting (v2) leaderboard for 17 actions. The methods are ranked according to the challenge set $a\text{-}mAP_{tight}$. *Main* considers all actions, *visible* only visible actions, and *unshown* unshown actions in the broadcast.

Methods	Test Set $a\text{-}mAP_{tight}$			$a\text{-}mAP_{loose}$			Challenge Set $a\text{-}mAP_{tight}$			$a\text{-}mAP_{loose}$		
	main	visible	unshown	main	visible	unshown	**main**	visible	unshown	main	visible	unshown
MEDet	/	/	/	/	/	/	**71.31**	76.29	54.09	78.56	81.67	69.13
mt_player	/	/	/	/	/	/	71.10	**77.22**	58.50	78.79	**82.02**	77.62
ASTRA[72a]	/	/	/	/	/	/	70.10	75.00	57.98	**79.21**	81.69	75.36
HCMUS-PK[83b]	62.49	69.04	30.42	73.98	80.09	45.05	69.38	75.50	53.31	76.15	80.08	66.62
team_ws_action	/	/	/	/	/	/	69.17	75.18	59.12	76.95	80.39	75.92
COMEDIAN[84c]	73.1	/	/	/	/	/	68.38	74.79	47.68	73.98	78.57	61.75
Soares et al.[57d]	65.07	71.24	36.62	78.59	83.89	54.15	68.33	73.22	**60.88**	78.06	80.58	**78.32**
DVP	/	/	/	/	/	/	66.95	74.68	53.81	73.61	79.15	67.38
E2E-Spot[58e]	/	/	/	/	/	/	66.73	74.84	53.21	73.62	79.16	67.42
AS&RG	/	/	/	/	/	/	64.88	70.31	53.03	72.83	76.08	72.35
MSAction	/	/	/	/	/	/	62.26	67.48	45.04	69.86	73.81	59.15
Faster-TAD[85]	61.10	25.50	54.09	/	/	/	/	/	/	/	/	/
Rkrystal	60.94	67.49	30.97	76.07	81.54	48.64	61.84	67.39	48.71	74.75	78.29	69.02
arturxe	57.28	63.57	29.04	72.13	77.11	45.62	60.56	65.75	53.00	71.72	75.15	69.91
cihe	60.51	66.17	31.91	75.33	80.20	46.74	59.97	64.51	53.80	72.95	76.29	71.95
SpotFormer[86]	60.9	48.6	31.0	81.5	48.6	31.9	/	/	/	/	/	/
STE-v2[87]	58.29	63.13	30.60	71.58	75.36	46.99	58.71	63.70	51.86	70.49	73.46	70.11
intro- and inter	/	/	/	/	/	/	53.30	58.48	47.30	66.59	68.75	67.09
memory	/	/	/	/	/	/	52.89	57.82	49.85	68.00	70.58	69.20
Zhu et al.[88f]	/	/	/	/	/	/	52.04	60.18	32.06	60.86	66.64	48.46
JAMY2 (AF_GRU)	/	/	/	/	/	/	51.97	58.05	44.29	63.12	65.98	61.66
tyru (GRU_CALF)	/	/	/	/	/	/	51.38	57.50	41.82	62.88	66.30	56.57
Zhou et al.[51g]	47.05	53.33	25.63	73.77	79.28	47.84	49.56	54.42	45.42	74.84	78.58	71.52
JAMY (LocPoint)	/	/	/	/	/	/	45.83	49.68	45.71	61.80	64.23	63.48
zqing	53.63	56.11	44.47	76.08	78.10	67.99	45.76	50.37	45.41	66.13	68.09	70.06
SIT	/	/	/	/	/	/	21.60	26.55	16.83	29.92	34.92	25.22
test_YYQ	/	/	/	/	/	/	12.73	14.13	11.21	54.21	58.75	48.55
ABCS	/	/	/	/	/	/	12.22	13.55	13.80	57.16	62.25	50.25
Visual Analysis of Humans	/	/	/	/	/	/	/	/	/	64.73	67.96	50.14
RMS-Net[75h]	/	/	/	63.49	68.88	38.02	/	/	/	60.92	64.09	56.61
MA-VLAD[89]	/	/	/	62.5	67.1	39.6	/	/	/	/	/	/
Shi et al.[76]	/	/	/	55.2	/	/	/	/	/	/	/	/
IdealCat	/	/	/	/	/	/	/	/	/	54.24	57.50	56.54
NetVLAD++[48i]	/	/	/	53.40	59.41	34.97	/	/	/	52.54	57.12	46.15
straw	/	/	/	49.79	56.35	31.14	/	/	/	51.65	57.03	45.33
CALF+PlayerLoc[90j]	/	/	/	46.8	/	/	/	/	/	/	/	/
CALF[40k]	/	/	/	41.61	43.54	28.88	/	/	/	42.22	43.51	37.91
AudioVid[81]	/	/	/	39.9	/	/	/	/	/	/	/	/
NetVLAD[13l]	/	/	/	31.37	34.30	23.27	/	/	/	30.74	32.99	23.27

[a] https://github.com/arturxe2/ASTRA
[b] https://github.com/Fsoft-AIC/UGLF
[c] https://github.com/juliendenize/eztorch
[d] https://github.com/yahoo/spivak
[e] https://github.com/jhong93/spot
[f] https://github.com/ArthurUnic/action_spotting
[g] https://github.com/baidu-research/vidpress-sport
[h] https://github.com/aimagelab/RMSNet_Soccer
[i] https://github.com/SoccerNet/sn-spotting/tree/main/Benchmarks/TemporallyAwarePooling
[j] https://github.com/SoccerNet/sn-spotting/tree/main/Benchmarks/CALF_Calibration_GCN
[k] https://github.com/SoccerNet/sn-spotting/tree/main/Benchmarks/CALF
[l] https://github.com/SoccerNet/sn-spotting/tree/main/Benchmarks/Pooling

niques. The first major performance boost came from Zhou *et al.*,[51] who fine-tuned pre-trained features on football data through a frame classification task. In the following year's challenge,[59] the evaluation metric was tightened from $a\text{-}mAP_{loose}$ to $a\text{-}mAP_{tight}$, requiring methods to focus on precise temporal localization. This led to significant breakthroughs, including the first end-to-end method proposed by Hong.[58] However, a feature-based method by Soares *et al.*[57] still emerged victorious, utilizing dense detection anchors to refine spotting predictions. The 2023 edition of the SoccerNet challenges[38] saw several methods surpass the 70% mark

Table 4.: Ball Action Spotting (2023) leaderboard for 2 classes. The methods are ranked according to the challenge set $mAP@1$.

| Methods | Test Set | | | | | | Challenge Set | | | | | |
| | mAP@ | | | | | $a\text{-}mAP_{tight}$ | mAP@ | | | | | $a\text{-}mAP_{tight}$ |
	1	2	3	4	5		1	2	3	4	5	
Ruslan Baikulov[a]	87.04	88.09	87.85	87.10	86.64	87.47	**86.47**	**87.98**	**88.28**	**88.18**	**87.95**	**87.91**
Wang *et al.*[91b]	86.37	87.16	87.15	85.80	85.48	86.51	83.39	85.19	85.81	86.00	86.19	85.45
BASIK	/	/	/	/	/	/	82.06	83.39	83.86	84.04	83.91	83.57
FC Pixel Nets	83.53	84.89	85.40	84.57	84.03	84.66	81.89	83.22	83.97	83.85	84.02	83.50
play	/	/	/	/	/	/	79.74	82.58	84.06	84.49	84.34	83.29
E2E-Spot[58c]	69.43	74.45	76.11	75.52	78.05	74.96	62.72	69.24	72.57	74.29	74.80	71.21

[a]https://github.com/lRomul/ball-action-spotting
[b]https://github.com/ZJLAB-AMMI/E2E-Spot-MBS
[c]https://github.com/jhong93/spot

Table 5.: Ball Action Spotting (2024) leaderboard for 12 classes. The methods are ranked according to the challenge set $mAP@1$.

| Methods | Test Set | | | | | | Challenge Set | | | | | |
| | mAP@ | | | | | $a\text{-}mAP_{tight}$ | mAP@ | | | | | $a\text{-}mAP_{tight}$ |
	1	2	3	4	5		1	2	3	4	5	
T-DEED[60a]	/	/	/	/	/	/	**73.39**	**76.82**	**77.83**	**78.42**	**78.51**	**77.25**
UniBw Munich - VIS	/	/	/	/	/	/	71.35	75.05	75.87	76.22	76.27	75.24
FS-TAHAKOM	/	/	/	/	/	/	67.09	70.22	71.26	71.74	71.59	70.64
MobiusLabs	/	/	/	/	/	/	62.53	65.46	67.01	67.15	67.27	66.13
AI4Sports	66.33	69.23	70.75	71.24	71.44	70.03	62.44	65.53	66.49	66.92	66.98	65.91
sota	/	/	/	/	/	/	62.44	65.68	66.69	67.48	67.35	66.19
SAIVA	/	/	/	/	/	/	56.74	60.23	60.98	61.86	61.75	60.58
Ruslan Baikulov[b]	59.70	62.73	63.74	64.47	64.87	63.31	56.15	60.30	61.20	61.86	61.91	60.60
blueblue111	/	/	/	/	/	/	3.66	4.08	4.12	4.17	4.01	4.05

[a]https://github.com/arturxe2/T-DEED
[b]https://github.com/lRomul/ball-action-spotting

in $a\text{-}mAP_{tight}$, with the winning entry employing a hybrid architecture that fused transformer and CNN-based encoders. The performances of these methods are summarized in Table 3.

In 2023, we introduced the first edition of the SoccerNet Ball Action Spotting challenge,[38] focusing on two classes: passes and drives. This new task presented three key challenges: the need to detect fast and subtle actions with minimal visual cues, the density of actions leading to difficulties in distinguishing between closely occurring actions, and the limited amount of training data available, which encouraged the use of semi-supervised, self-supervised, and transfer learning techniques. Participants in this challenge explored various strategies, including pre-training on action spotting videos and fine-tuning on ball action data, utilizing stacked sequences of grayscale images in the RGB channels, employing focal loss for label expansion, and model ensembling to combine different network variants. The results of these methods are detailed in Table 4.

In 2024, the complexity of the ball action spotting task[42] was increased by ex-

panding the number of classes from 2 to 12. This expansion required more granular categorization of actions, such as differentiating between various types of passes. Participants employed several advanced techniques to tackle this challenge. Similarly to 2023, key strategies included the use of grayscale images to simplify visual processing, the integration of 2D and 3D convolutional networks to capture both spatial and temporal features, and the application of transfer learning from related domains to enhance model training. The winning team, T-DEED,[69] introduced an innovative approach using a 2D backbone with Gate-Shift Fuse (GSF) modules for generating per-frame representations with local spatiotemporal information. Their method also included a temporally-discriminant encoder-decoder that refined per-frame tokens, increasing their discriminability within sequences while maintaining high temporal resolution. To address the issue of limited data, they trained their model on both SoccerNet Ball Action Spotting (2024) and the original SoccerNet Action Spotting (v2) datasets simultaneously, using a multitask training approach with dual prediction heads. The results of these methods are presented in Table 5.

Overall, the challenges organized around SoccerNet have attracted significant attention and fostered the development of innovative methods that have applications beyond sports video analysis. Notably, some winning methods, such as E2E-Spot by Hong *et al.*,[58] Dense Detection Anchors by Soares *et al.*,[57] and T-DEED by Xarles *et al.*,[69] have shown strong performance on other datasets, demonstrating that advancements made for sports can contribute to breakthroughs in more generic video understanding. This underscores the relevance of the SoccerNet dataset, benchmark, and challenges to the broader research community. By providing open-source data and organizing these challenges, SoccerNet has successfully built a large community of researchers dedicated to advancing the field of video understanding.

7. Conclusion

In this chapter, we provided a comprehensive overview of the task of *action spotting*, structured around three key pillars: datasets, methods, and evaluation metrics. We began by tracing the evolution of action spotting, highlighting the important role of the SoccerNet datasets in initiating extensive research in this domain. The introduction of these large-scale datasets, completed with detailed annotations, has been crucial in driving forward the field by providing a solid foundation for the development of innovative approaches.

Next, we explored how deep learning methods progressively approached the task of action spotting, discussing the general architecture and then describing specific methods published in the literature. Specifically, we examined both feature-based approaches, which rely on pre-trained backbones, and end-to-end methods that optimize the entire model architecture for the task. These methods have introduced significant advancements, each contributing unique insights and techniques that have pushed the state of the art in action spotting.

We also detailed the evaluation metrics, which are essential for assessing the

performance of action spotting methods, emphasizing the importance of precise temporal localization and the challenges it presents. Following this, we presented a retrospective analysis of six years of research, showcasing a thorough benchmark of methods developed in response to the four SoccerNet Action Spotting datasets showcased in open challenges. These challenges have been a driving force in the field, encouraging researchers to continuously improve upon the state of the art.

Overall, this chapter serves as a timestamped summary of the history of the SoccerNet Action Spotting datasets and challenges, cataloging over 60 methods that have been developed or published, many of which include publicly available code to facilitate further research. SoccerNet underscores the profound impact that open-source data, methods, and benchmarks have on advancing the field of video understanding, particularly in the context of sports.

Acknowledgements

References

1. T. B. Moeslund, G. Thomas, and A. Hilton, *Computer vision in sports.* (Springer, 2014). doi: 10.1007/978-3-319-09396-3. URL https://doi.org/10.1007/978-3-319-09396-3.

2. G. Thomas, R. Gade, T. B. Moeslund, P. Carr, and A. Hilton, Computer vision for sports: current applications and research topics, *Comput. Vis. Image Underst.* **159**, 3–18 (Jun., 2017). doi: 10.1016/j.cviu.2017.04.011. URL https://doi.org/10.1016/j.cviu.2017.04.011.

3. R. Gade, M. Merler, G. Thomas, and T. B. Moeslund, *The (Computer) Vision of Sports: Recent Trends in Research and Commercial Systems for Sport Analytics*, In *Computer Vision: Challenges, Trends, and Opportunities*, chapter 14, pp. 1–16. Chapman & Hall, (2024). doi: 10.1201/9781003328957. URL https://doi.org/10.1201/9781003328957.

4. Statista Research Department. Global sports market – statistics and facts. statista: The Statistics Portal (Mar., 2024). URL https://www.statista.com/topics/8468/global-sports-market/#topicOverview.

5. G. Gough. eSports market – statistics & facts. statista: The Statistics Portal (Jul., 2024). URL https://www.statista.com/topics/3121/esports-market/.

6. J. Stoll. Length of sports TV broadcast hours in the United States from 2002 to 2017. statista: The Statistics Portal (Jan., 2021). URL https://www.statista.com/statistics/290110/length-sports-tv-programming-available-usa/.

7. A. Cioppa, S. Giancola, A. Deliege, L. Kang, X. Zhou, Z. Cheng, B. Ghanem, and M. Van Droogenbroeck. 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, New Orleans, LA, USA (Jun., 2022). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvprw56347.2022.00393. URL https://doi.org/10.1109/CVPRW56347.2022.00393.

8. V. Somers, V. Joos, A. Cioppa, S. Giancola, S. A. Ghasemzadeh, F. Magera, B. Standaert, A. M. Mansourian, X. Zhou, S. Kasaei, B. Ghanem, A. Alahi, M. Van Droogenbroeck, and C. De Vleeschouwer. 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, Seattle, WA, USA (Jun., 2024). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvprw63382.2024.00334. URL https://doi.org/10.1109/CVPRW63382.2024.00334.

9. A. Deliège, A. Cioppa, S. Giancola, M. J. Seikavandi, J. V. Dueholm, K. Nasrollahi, B. Ghanem, T. B. Moeslund, and M. Van Droogenbroeck. SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pp. 4503–4514, Nashville, TN, USA (Jun., 2021). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvprw53098.2021.00508. URL https://doi.org/10.1109/CVPRW53098.2021.00508.

10. H. Mkhallati, A. Cioppa, S. Giancola, B. Ghanem, and M. Van Droogenbroeck. SoccerNet-caption: Dense video captioning for soccer broadcasts commentaries. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pp. 5074–5085, Vancouver, Can. (Jun., 2023). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvprw59228.2023.00536. URL https://doi.org/10.1109/CVPRW59228.2023.00536.

11. A. Cioppa, A. Deliège, S. Giancola, B. Ghanem, and M. Van Droogenbroeck, Scaling up SoccerNet with multi-view spatial localization and re-identification, *Sci. Data*. **9** (1), 1–9 (Jun., 2022). doi: 10.1038/s41597-022-01469-1. URL https://doi.org/10.1038/s41597-022-01469-1.

12. F. Magera, T. Hoyoux, O. Barnich, and M. Van Droogenbroeck. A universal protocol to benchmark camera calibration for sports. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, vol. abs/2309.06006, pp. 3335–3346, Seattle, WA, USA (Jun., 2024). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvprw63382.2024.00338. URL https://doi.org/10.1109/CVPRW63382.2024.00338.

13. S. Giancola, M. Amine, T. Dghaily, and B. Ghanem. SoccerNet: A scalable dataset for action spotting in soccer videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pp. 1792–179210, Salt Lake City, UT, USA (Jun., 2018). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvprw.2018.00223. URL https://doi.org/10.1109/cvprw.2018.00223.

14. F. Wu, Q. Wang, J. Bian, N. Ding, F. Lu, J. Cheng, D. Dou, and H. Xiong, A survey on video action recognition in sports: Datasets, methods and applications, *IEEE Trans. Multimedia*. **25**, 7943–7966, (2023). doi: 10.1109/tmm.2022.3232034. URL https://doi.org/10.1109/TMM.2022.3232034.

15. K. Seweryn, A. Wróblewska, and S. Łukasik, Survey of action recognition, spotting and spatio-temporal localization in soccer – current trends and research perspectives, *arXiv*. **abs/2309.12067**, (2023). doi: 10.48550/arXiv.2309.12067. URL https://doi.org/10.48550/arXiv.2309.12067.

16. F. Caba Heilbron, V. Escorcia, B. Ghanem, and J. Carlos Niebles. ActivityNet: A large-scale video benchmark for human activity understanding. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 961–970, Boston, MA, USA (Jun., 2015). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/CVPR.2015.7298698. URL https://doi.org/10.1109/CVPR.2015.7298698.

17. IFAB. Laws of the game. Technical report, The International Football Association Board, Zurich, Switzerland, (2022).

18. J. Held, A. Cioppa, S. Giancola, A. Hamdi, B. Ghanem, and M. Van Droogenbroeck. 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, Vancouver, Can. (Jun., 2023). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvprw59228.2023.00537. URL https://doi.org/10.1109/CVPRW59228.2023.00537.

19. J. Held, H. Itani, A. Cioppa, S. Giancola, B. Ghanem, and M. Van Droogenbroeck. 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, Seattle, WA, USA (Jun., 2024). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvprw63382.2024.00332. URL https://doi.org/10.1109/CVPRW63382.2024.00332.

20. J. Valand, H. Kadragic, S. Hicks, V. Thambawita, C. Midoglu, T. Kupka, D. Johansen, M. Riegler, and P. Halvorsen, AI-based video clipping of soccer events, *Machine learning & knowledge extraction.* **3**(4), 1–19 (Dec., 2021). doi: 10.3390/make3040049. URL https://doi.org/10.3390/make3040049.

21. J. Valand, H. Kadragic, S. Hicks, V. Thambawita, C. Midoglu, T. Kupka, D. Johansen, M. Riegler, and P. Halvorsen. Automated clipping of soccer events using machine learning. In *IEEE International Symposium on Multimedia (ISM)*, pp. 210–214, Naple, Italy (Nov.-Dec., 2021). doi: 10.1109/ISM52913.2021.00042. URL https://doi.org/10.1109/ISM52913.2021.00042.

22. S. Giancola, A. Cioppa, J. Georgieva, J. Billingham, A. Serner, K. Peek, B. Ghanem, and M. Van Droogenbroeck. Towards active learning for action spotting in association football videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pp. 5098–5108, Vancouver, Can. (Jun., 2023). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvprw59228.2023.00538. URL https://doi.org/10.1109/CVPRW59228.2023.00538.

23. S. Akan and S. Varlı, Use of deep learning in soccer videos analysis: survey, *Multimedia Systems.* pp. 1–19 (Dec., 2022). doi: 10.1007/s00530-022-01027-0. URL https://doi.org/10.1007/s00530-022-01027-0.

24. Z. Zhao, W. Chai, S. Hao, W. Hu, G. Wang, S. Cao, M. Song, J.-N. Hwang, and G. Wang, A survey of deep learning in sports applications: Perception, comprehension, and decision, *arXiv.* **abs/2307.03353**, (2023). doi: 10.48550/arXiv.2307.03353. URL https://doi.org/10.48550/arXiv.2307.03353.

25. D. Pontes, C. Costa, R. G. Faria, and J. H. Brito, Application of machine learning in soccer broadcast: A systematic review, *Soc. Sci. Res. Netw.* (2024). doi: 10.2139/ssrn.4870821. URL https://doi.org/10.2139/ssrn.4870821.

26. P. Pareek and A. Thakkar, A survey on video-based human action recognition: recent updates, datasets, challenges, and applications, *Artif. Intell. Rev.* **54**(3), 2259–2322 (Sept., 2020). doi: 10.1007/s10462-020-09904-8. URL https://doi.org/10.1007/s10462-020-09904-8.

27. M. Tan and Q. V. Le. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Int. Conf. Mach. Learn. (ICML)*, vol. 97, pp. 6105–6114, Long Beach, CA, USA (Jun., 2019). PMLR.

28. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 248–255, Miami, FL, USA (Jun., 2009). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/CVPR.2009.5206848. URL https://doi.org/10.1109/CVPR.2009.5206848.

29. T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft COCO: Common Objects in Context. In *Eur. Conf. Comput. Vis. (ECCV)*, vol. 8693, *Lect. Notes Comput. Sci.*, pp. 740–755. Springer (Sept., 2014). doi: 10.1007/978-3-319-10602-1_48. URL https://doi.org/10.1007/

978-3-319-10602-1_48.

30. A. Cioppa, A. Deliège, and M. 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*, pp. 1846–1855, Salt Lake City, UT, USA (Jun., 2018). doi: 10.1109/CVPRW.2018.00229. URL http://hdl.handle.net/2268/222427.

31. N. Homayounfar, S. Fidler, and R. Urtasun. Sports field localization via deep structured models. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4012–4020, Honolulu, HI, USA (Jul., 2017). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/CVPR.2017.427. URL https://doi.org/10.1109/CVPR.2017.427.

32. H. Biermann, J. Theiner, M. Bassek, D. Raabe, D. Memmert, and R. Ewerth. A unified taxonomy and multimodal dataset for events in invasion games. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, vol. 9, pp. 1–10. ACM (Oct., 2021). doi: 10.1145/3475722.3482792. URL https://doi.org/10.1145/3475722.3482792.

33. Y. Jiang, K. Cui, L. Chen, C. Wang, and C. Xu. SoccerDB: A large-scale database for comprehensive video understanding. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, p. 1–8, Seattle, WA, USA (Oct., 2020). ACM. doi: 10.1145/3422844.3423051. URL https://doi.org/10.1145/3422844.3423051.

34. N. Feng, Z. Song, J. Yu, Y.-P. P. Chen, Y. Zhao, Y. He, and T. Guan, SSET: a dataset for shot segmentation, event detection, player tracking in soccer videos, *Multimedia Tools Appl.* **79**(39), 28971–28992 (Oct., 2020). ISSN 1573-7721. doi: 10.1007/s11042-020-09414-3. URL https://doi.org/10.1007/s11042-020-09414-3.

35. G. Van Zandycke, V. Somers, M. Istasse, C. D. Don, and D. Zambrano. DeepSportradar-v1: Computer vision dataset for sports understanding with high quality annotations. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pp. 1–8, Lisbon, Port. (Oct., 2022). ACM. doi: 10.1145/3552437.3555699. URL https://doi.org/10.1145/3552437.3555699.

36. M. Istasse, V. Somers, P. Elancheliyan, J. De, and D. Zambrano. DeepSportradar-v2: A multi-sport computer vision dataset for sport understandings. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pp. 23–29, Ottawa, Ontario, Can. (Oct., 2023). ACM. doi: 10.1145/3606038.3616160. URL https://doi.org/10.1145/3606038.3616160.

37. E. J. Kassab, H. M. Solberg, S. Gautam, S. S. Sabet, T. Torjusen, M. Riegler, P. Halvorsen, and C. Midoglu. TACDEC: Dataset of tackle events in soccer game videos. In *Proc. ACM Multimedia Syst. Conf.*, p. 250–256, Bari, Italy (Apr., 2024). ACM. doi: 10.1145/3625468.3652166. URL https://doi.org/10.1145/3625468.3652166.

38. A. Cioppa, S. Giancola, V. Somers, F. Magera, X. Zhou, H. Mkhallati, A. Deliège, J. Held, C. Hinojosa, A. M. Mansourian, P. Miralles, O. Barnich, C. De Vleeschouwer, A. Alahi, B. Ghanem, M. Van Droogenbroeck, A. Kamal, A. Maglo, A. Clapés, A. Abdelaziz, A. Xarles, A. Orcesi, A. Scott, B. Liu, B. Lim, C. Chen, F. Deuser, F. Yan, F. Yu, G. Shitrit, G. Wang, G. Choi, H. Kim, H. Guo, H. Fahrudin, H. Koguchi, H. Ardö, I. Salah, I. Yerushalmy, I. Muhammad, I. Uchida, I. Be'ery, J. Rabarisoa, J. Lee, J. Fu, J. Yin, J. Xu, J. Nang, J. Denize, J. Li, J. Zhang, J. Kim, K. Synowiec, K. Kobayashi, K. Zhang, K. Habel, K. Nakajima, L. Jiao, L. Ma, L. Wang, L. Wang, M. Li, M. Zhou, M. Nasr, M. Abdelwahed, M. Liashuha, N. Falaleev, N. Oswald, Q. Jia, Q.-C. Pham, R. Song, R. Hérault, R. Peng, R. Chen, R. Liu, R. Baikulov, R. Fukushima, S. Escalera, S. Lee, S. Chen, S. Ding, T. Someya, T. B. Moeslund, T. Li, W. Shen, W. Zhang, W. Li, W. Dai, W. Luo, W. Zhao, W. Zhang, X. Yang, Y. Ma, Y. Joo, Y. Zeng, Y. Gan, Y. Zhu, Y. Zhong, Z. Ruan, Z. Li, Z. Huang, and

Z. Meng, SoccerNet 2023 challenges results, *Sports Eng.* **27**(2), 1–18 (July, 2024). doi: 10.1007/s12283-024-00466-4. URL https://doi.org/10.1007/s12283-024-00466-4.

39. A. Leduc, A. Cioppa, S. Giancola, B. Ghanem, and M. Van Droogenbroeck. 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, Seattle, WA, USA (Jun., 2024). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvprw63382.2024.00333. URL https://doi.org/10.1109/CVPRW63382.2024.00333.

40. A. Cioppa, A. Deliège, S. Giancola, B. Ghanem, M. Van Droogenbroeck, R. Gade, and T. B. Moeslund. A context-aware loss function for action spotting in soccer videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 13123–13133, Seattle, WA, USA (Jun., 2020). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvpr42600.2020.01314. URL https://doi.org/10.1109/CVPR42600.2020.01314.

41. Y. Benzakour, B. Cabado, S. Giancola, A. Cioppa, B. Ghanem, and M. Van Droogenbroeck. OSL-ActionSpotting: A unified library for action spotting in sports videos. In *IEEE Int. Work. Sport Technol. Res. (STAR)*, pp. 132–137, Lecco, Italy (Jul., 2024). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/star62027.2024.10635981. URL https://doi.org/10.1109/STAR62027.2024.10635981.

42. A. Cioppa, S. Giancola, V. Somers, V. Joos, F. Magera, J. Held, S. A. Ghasemzadeh, X. Zhou, K. Seweryn, M. Kowalczyk, Z. Mróz, S. Łukasik, M. Hałoń, H. Mkhallati, A. Deliège, C. Hinojosa, K. Sanchez, A. M. Mansourian, P. Miralles, O. Barnich, C. De Vleeschouwer, A. Alahi, B. Ghanem, M. Van Droogenbroeck, A. Gorski, A. Clapés, A. Boiarov, A. Afanasiev, A. Xarles, A. Scott, B. Lim, C. Yeung, C. Gonzalez, D. Rüfenacht, E. Pacilio, F. Deuser, F. S. Altawijri, F. Cachón, H. Kim, H. Wang, H. Choe, H. J. Kim, I.-M. Kim, J.-M. Kang, J. Tursunboev, J. Yang, J. Hong, J. Lee, J. Zhang, J. Lee, K. Zhang, K. Habel, L. Jiao, L. Li, M. Gutiérrez-Pérez, M. Ortega, M. Li, M. Lopatto, N. Kasatkin, N. Nemtsev, N. Oswald, O. Udin, P. Kononov, P. Geng, S. G. Alotaibi, S. Kim, S. Ulasen, S. Escalera, S. Zhang, S. Yang, S. Moon, T. B. Moeslund, V. Shandyba, V. Golovkin, W. Dai, W. Chung, X. Liu, Y. Zhu, Y. Kim, Y. Li, Y. Yang, Y. Xiao, Z. Cheng, and Z. Li, SoccerNet 2024 challenges results, *arXiv.* **abs/2409.10587**, (2024). doi: 10.48550/arXiv.2409.10587. URL https://doi.org/10.48550/arXiv.2409.10587.

43. K. He, X. Zhang, S. Ren, and J. Sun, Deep residual learning for image recognition, *arXiv.* **abs/1512.03385**, (2015). doi: 10.48550/arXiv.1512.03385. URL https://doi.org/10.48550/arXiv.1512.03385.

44. D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 4489–4497, Santiago, Chile (Dec., 2015). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/iccv.2015.510. URL https://doi.org/10.1109/ICCV.2015.510.

45. J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4724–4733, Honolulu, HI, USA (Jul., 2017). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvpr.2017.502. URL https://doi.org/10.1109/CVPR.2017.502.

46. J. Held, A. Cioppa, S. Giancola, A. Hamdi, C. Devue, B. Ghanem, and M. Van Droogenbroeck, Towards AI-powered video assistant referee system (VARS) for association football, *arXiv.* **abs/2407.12483**, (2024). doi: 10.48550/arXiv.2407.12483. URL https://doi.org/10.48550/arXiv.2407.12483.

47. R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, NetVLAD: CNN architecture for weakly supervised place recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(6), 1437–1451 (Jun., 2018). doi: 10.1109/tpami.2017.2711011. URL

https://doi.org/10.1109/TPAMI.2017.2711011.

48. S. Giancola and B. Ghanem. Temporally-aware feature pooling for action spotting in soccer broadcasts. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4490–4499, Nashville, TN, USA (Jun., 2021). doi: 10.1109/CVPRW53098.2021.00506.

49. J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, You only look once: Unified, real-time object detection, *arXiv.* **abs/1506.02640** (Jun., 2016). URL https://arxiv.org/abs/1506.02640.

50. T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang. BSN: Boundary sensitive network for temporal action proposal generation. In *Eur. Conf. Comput. Vis. (ECCV)*, vol. 11208, *Lect. Notes Comput. Sci.*, pp. 3–21. Springer Int. Publ., (2018). doi: 10.1007/978-3-030-01225-0_1. URL https://doi.org/10.1007/978-3-030-01225-0_1.

51. X. Zhou, L. Kang, Z. Cheng, B. He, and J. Xin, Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection, *arXiv.* **abs/2106.14447**, (2021). doi: 10.48550/arXiv.2106.14447. URL https://doi.org/10.48550/arXiv.2106.14447.

52. C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou. Temporal pyramid network for action recognition. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 588–597, Seattle, WA, USA (Jun., 2020). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvpr42600.2020.00067. URL https://doi.org/10.1109/CVPR42600.2020.00067.

53. B. He, X. Yang, Z. Wu, H. Chen, S.-N. Lim, and A. Shrivastava. GTA: Global temporal attention for video action understanding. In *Br. Mach. Vis. Conf. (BMVC)*, Virtual conference, (2021).

54. D. Neimark, O. Bar, M. Zohar, and D. Asselmann. Video transformer network. In *IEEE/CVF Int. Conf. Comput. Vis. Work. (ICCV Work.)*, pp. 3156–3165, Montréal, Can. (Oct., 2021). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/iccvw54120.2021.00355. URL https://doi.org/10.1109/ICCVW54120.2021.00355.

55. D. Tran, H. Wang, M. Feiszli, and L. Torresani. Video classification with channel-separated convolutional networks. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 5551–5560, Seoul, South Korea (Oct., 2019). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/iccv.2019.00565. URL https://doi.org/10.1109/ICCV.2019.00565.

56. C. Feichtenhofer, H. Fan, J. Malik, and K. He. SlowFast networks for video recognition. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 6201–6210, Seoul, South Korea (Oct., 2019). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/iccv.2019.00630. URL https://doi.org/10.1109/ICCV.2019.00630.

57. J. V. B. Soares, A. Shah, and T. Biswas. Temporally precise action spotting in soccer videos using dense detection anchors. In *IEEE Int. Conf. Image Process. (ICIP)*, pp. 2796–2800, Bordeaux, France (Oct., 2022). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/icip46576.2022.9897256. URL https://doi.org/10.1109/ICIP46576.2022.9897256.

58. J. Hong, H. Zhang, M. Gharbi, M. Fisher, and K. Fatahalian. Spotting temporally precise, fine-grained events in video. In *Eur. Conf. Comput. Vis. (ECCV)*, vol. 13695, *Lect. Notes Comput. Sci.*, pp. 33–51, Tel Aviv, Israël, (2022). Springer Nat. Switz. doi: 10.1007/978-3-031-19833-5_3. URL https://doi.org/10.1007/978-3-031-19833-5_3.

59. S. Giancola, A. Cioppa, A. Deliège, F. Magera, V. Somers, L. Kang, X. Zhou, O. Barnich, C. De Vleeschouwer, A. Alahi, B. Ghanem, M. Van Droogenbroeck, A. Darwish, A. Maglo, A. Clapés, A. Luyts, A. Boiarov, A. Xarles, A. Orcesi, A. Shah, B. Fan, B. Comandur, C. Chen, C. Zhang, C. Zhao, C. Lin, C.-Y. Chan, C. C. Hui, D. Li, F. Yang, F. Liang, F. Da, F. Yan, F. Yu, G. Wang, H. A. Chan, H. Zhu, H. Kan, J. Chu, J. Hu, J. Gu, J. Chen, J. V. B. Soares, J. Theiner, J. De Corte, J. H. Brito,

J. Zhang, J. Li, J. Liang, L. Shen, L. Ma, L. Chen, M. Santos Marques, M. Azatov, N. Kasatkin, N. Wang, Q. Jia, Q. C. Pham, R. Ewerth, R. Song, R. Li, R. Gade, R. Debien, R. Zhang, S. Lee, S. Escalera, S. Jiang, S. Odashima, S. Chen, S. Masui, S. Ding, S.-w. Chan, S. Chen, T. El-Shabrawy, T. He, T. B. Moeslund, W.-C. Siu, W. Zhang, W. Li, X. Wang, X. Tan, X. Li, X. Wei, X. Ye, X. Liu, X. Wang, Y. Guo, Y. Zhao, Y. Yu, Y. Li, Y. He, Y. Zhong, Z. Guo, and Z. Li. SoccerNet 2022 challenges results. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pp. 75–86, Lisbon, Port. (Oct., 2022). ACM. doi: 10.1145/3552437.3558545. URL https://doi.org/10.1145/3552437.3558545.

60. I. Radosavovic, R. P. Kosaraju, R. Girshick, K. He, and P. Dollar. Designing network design spaces. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 10425–10433, Seattle, WA, USA (Jun., 2020). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvpr42600.2020.01044. URL https://doi.org/10.1109/CVPR42600.2020.01044.

61. Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie. A ConvNet for the 2020s. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11966–11976, New Orleans, LA, USA (Jun., 2022). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvpr52688.2022.01167. URL https://doi.org/10.1109/CVPR52688.2022.01167.

62. J. Lin, C. Gan, and S. Han. TSM: Temporal shift module for efficient video understanding. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 7082–7092. Inst. Electr. Electron. Eng. (IEEE) (Oct., 2019). doi: 10.1109/iccv.2019.00718. URL https://doi.org/10.1109/ICCV.2019.00718.

63. S. Sudhakaran, S. Escalera, and O. Lanz. Gate-shift networks for video action recognition. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 1099–1108, Seattle, WA, USA (Jun., 2020). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvpr42600.2020.00118. URL https://doi.org/10.1109/cvpr42600.2020.00118.

64. J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *Adv. Neural Inf. Process. Syst. (NeurIPS) Work.*, pp. 1–9, Montréal, Can. (Dec., 2014). Curran Assoc. Inc.

65. Y. Abu Farha and J. Gall. MS-TCN: Multi-stage temporal convolutional network for action segmentation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3570–3579, Long Beach, CA, USA (Jun., 2019). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvpr.2019.00369. URL https://doi.org/10.1109/CVPR.2019.00369.

66. F. Yi, H. Wen, and T. Jiang. ASFormer: Transformer for action segmentation. In *Br. Mach. Vis. Conf. (BMVC)*, pp. 1–15, Virtual conference (Nov., 2021).

67. R. Baikulov. Solution for SoccerNet ball action spotting challenge 2023. https://github.com/lRomul/ball-action-spotting (Jun., 2023).

68. M. Tan and Q. V. Le. EfficientNetV2: Smaller models and faster training. In *Int. Conf. Mach. Learn. (ICML)*, vol. 139, *Proc. Mach. Learn. Res.*, pp. 10096–10106 (Jul., 2021). URL https://proceedings.mlr.press/v139/tan21a.html.

69. A. Xarles, S. Escalera, T. B. Moeslund, and A. Clapés, T-DEED: Temporal-discriminability enhancer encoder-decoder for precise event spotting in sports videos, *arXiv.* **abs/2404.05392**, (2024). doi: 10.48550/arXiv.2404.05392. URL https://doi.org/10.48550/arXiv.2404.05392.

70. S. Sudhakaran, S. Escalera, and O. Lanz, Gate-shift-fuse for video action recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* **45**(9), 10913–10928 (Sept., 2023). doi: 10.1109/tpami.2023.3268134. URL https://doi.org/10.1109/TPAMI.2023.3268134.

71. D. Shi, Y. Zhong, Q. Cao, L. Ma, J. Lit, and D. Tao. TriDet: Temporal action de-

tection with relative boundary modeling. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 18857–18866, Vancouver, Can. (Jun., 2023). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvpr52729.2023.01808. URL https://doi.org/10.1109/CVPR52729.2023.01808.

72. A. Xarles, S. Escalera, T. B. Moeslund, and A. Clapés. ASTRA: An Action Spotting TRAnsformer for soccer videos. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, p. 93–102, Ottawa, Ontario, Can. (Oct., 2023). ACM. doi: 10.1145/3606038.3616153. URL https://doi.org/10.1145/3606038.3616153.

73. M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, The PASCAL visual object classes (VOC) challenge, *Int. J. Comput. Vis.* **88**(2), 303–338 (Jun., 2010). doi: 10.1007/s11263-009-0275-4. URL https://doi.org/10.1007/s11263-009-0275-4.

74. H. Minoura, T. Hirakawa, T. Yamashita, H. Fujiyoshi, M. Nakazawa, Y. Chea, and B. Stenger. Action spotting and temporal attention analysis in soccer videos. In *International Conference on Machine Vision and Applications (MVA)*, pp. 1–6, Aichi, Japan (Jul., 2021). doi: 10.23919/MVA51890.2021.9511342. URL https://doi.org/10.23919/MVA51890.2021.9511342.

75. M. Tomei, L. Baraldi, S. Calderara, S. Bronzin, and R. Cucchiara. RMS-net: Regression and masking for soccer event spotting. In *IEEE Int. Conf. Pattern Recognit. (ICPR)*, pp. 7699–7706, Milan, Italy (Jan., 2021). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/icpr48806.2021.9412268. URL https://doi.org/10.1109/icpr48806.2021.9412268.

76. Y. Shi, H. Minoura, T. Yamashita, T. Hirakawa, H. Fujiyoshi, M. Nakazawa, Y. Chae, and B. Stenger. Action spotting in soccer videos using multiple scene encoders. In *IEEE Int. Conf. Pattern Recognit. (ICPR)*, pp. 3183–3189, Montréal, Can. (Aug., 2022). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/icpr56361.2022.9956667. URL https://doi.org/10.1109/ICPR56361.2022.9956667.

77. A. Karimi, R. Toosi, and M. A. Akhaee. Soccer video event detection using metric learning. In *International Conference on Computer and Knowledge Engineering (ICCKE)*, pp. 048–052, Mashhad, Islamic Republic of Iran (Nov., 2022). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/iccke57176.2022.9959985. URL https://doi.org/10.1109/ICCKE57176.2022.9959985.

78. B. Mahaseni, E. Faizal, and R. Raj, Spotting football events using two-stream convolutional neural network and dilated recurrent neural network, *IEEE Access.* **9**, 61929–61942 (Apr., 2021). doi: 10.1109/ACCESS.2021.3074831. URL https://doi.org/10.1109/ACCESS.2021.3074831.

79. K. Vats, M. Fani, P. Walters, D. A. Clausi, and J. Zelek. Event detection in coarsely annotated sports videos via parallel multi receptive field 1D convolutions. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pp. 3856–3865, Seattle, WA, USA (Jun., 2020). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvprw50498.2020.00449. URL https://doi.org/10.1109/CVPRW50498.2020.00449.

80. O. Rongved, M. Stige, S. Hicks, V. Thambawita, C. Midoglu, E. Zouganeli, D. Johansen, M. Riegler, and P. Halvorsen, Automated event detection and classification in soccer: The potential of using multiple modalities, *Machine Learning and Knowledge Extraction.* **3**(4), 1–25 (Dec., 2021). doi: 10.3390/make3040051. URL https://doi.org/10.3390/make3040051.

81. B. Vanderplaetse and S. Dupont. Improved soccer action spotting using both audio and video streams. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pp. 3921–3931, Seattle, WA, USA (Jun., 2020). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/cvprw50498.2020.00456. URL https://doi.org/

10.1109/CVPRW50498.2020.00456.

82. O. Rongverd, S. Hicks, V. Thambawita, H. Stensland, E. Zouganeli, D. Johansen, M. Midoglu, C. Riegler, and P. Halvorsen, Using 3D convolutional neural networks for real-time detection of soccer events, *International Journal of Semantic Computing.* **15**(2), 161–187 (Jun., 2021). doi: 10.1142/S1793351X2140002X. URL https://doi.org/10.1142/S1793351X2140002X.

83. K. H. Tran, P. V. Do, N. Q. Ly, and N. Le, Unifying global and local scene entities modelling for precise action spotting, *arXiv.* **abs/2404.09951**, (2024). doi: 10.48550/arXiv.2404.09951. URL https://doi.org/10.48550/arXiv.2404.09951.

84. J. Denize, M. Liashuha, J. Rabarisoa, A. Orcesi, and R. Hérault. COMEDIAN: Self-supervised learning and knowledge distillation for action spotting using transformers. In *IEEE/CVF Winter Conf. Appl. Comput. Vis. Work. (WACVW)*, pp. 518–528, Waikoloa, HI, USA (Jan., 2024). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/wacvw60836.2024.00060. URL https://doi.org/10.1109/WACVW60836.2024.00060.

85. S. Chen, C. Chen, W. Li, X. Tao, and Y. Guo, Faster-TAD: Towards temporal action detection with proposal generation and classification in a unified network, *arXiv.* **abs/2204.02674**, (2022). doi: 10.48550/arXiv.2204.02674. URL https://doi.org/10.48550/arXiv.2204.02674.

86. M. Cao, M. Yang, G. Zhang, X. Li, Y. Wu, G. Wu, and L. Wang. SpotFormer: A transformer-based framework for precise soccer action spotting. In *Int. Work. Multimedia Signal Process. (MMSP)*, pp. 1–6, Shanghai, China (Sept., 2022). Inst. Electr. Electron. Eng. (IEEE). doi: 10.1109/mmsp55362.2022.9948888. URL https://doi.org/10.1109/MMSP55362.2022.9948888.

87. A. Darwish and T. El-Shabrway. STE: Spatio-temporal encoder for action spotting in soccer videos. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pp. 87–92, Lisbon, Port. (Oct., 2022). ACM. doi: 10.1145/3552437.3555704. URL https://doi.org/10.1145/3552437.3555704.

88. H. Zhu, J. Liang, C. Lin, J. Zhang, and J. Hu. A transformer-based system for action spotting in soccer videos. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pp. 103–109, Lisbon, Port. (Oct., 2022). ACM. doi: 10.1145/3552437.3555693. URL https://doi.org/10.1145/3552437.3555693.

89. N. Feng, Y. Tang, Z. Song, J. Yu, Y.-P. P. Chen, and W. Yang, MA-VLAD: a fine-grained local feature aggregation scheme for action recognition, *Multimedia Systems.* **30**(3), 1–13 (May, 2024). doi: 10.1007/s00530-024-01341-9. URL https://doi.org/10.1007/s00530-024-01341-9.

90. A. Cioppa, A. Deliège, S. Giancola, F. Magera, O. Barnich, B. Ghanem, and M. Van Droogenbroeck. Camera calibration and player localization in SoccerNet-v2 and investigation of their representations for action spotting. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pp. 4532–4541, Nashville, TN, USA (Jun., 2021). doi: 10.1109/CVPRW53098.2021.00511. URL http://hdl.handle.net/2268/259026.

91. L. Wang, H. Guo, and B. Liu, A boosted model ensembling approach to ball action spotting in videos: The runner-up solution to CVPR'23 SoccerNet challenge, *arXiv.* **abs/2306.05772**, (2023). doi: 10.48550/arXiv.2306.05772. URL https://doi.org/10.48550/arXiv.2306.05772.