OSL-ActionSpotting: A Unified Library for Action Spotting in Sports Videos

Yassine Benzakour
Montefiore Institute, University of Liège
Email: yassine.benzakour@student.uliege.be

Bruno Cabado
Cinfo & CITIC Research Center,
University of A Coruña
Email: bruno.cabado@udc.es

Silvio Giancola
IVUL, KAUST
Email: silvio.giancola@kaust.edu.sa

Anthony Cioppa
Montefiore Institute, University of Liège
Email:anthony.cioppa@uliege.be

Bernard Ghanem
IVUL, KAUST
Email: bernard.ghanem@kaust.edu.sa

Marc Van Droogenbroeck
Montefiore Institute, University of Liège
Email: m.vandroogenbroeck@uliege.be

***Abstract*—Action spotting is crucial in sports analytics as it enables the precise identification and categorization of pivotal moments in sports matches, providing insights that are essential for performance analysis and tactical decision-making. The fragmentation of existing methodologies, however, impedes the progression of sports analytics, necessitating a unified codebase to support the development and deployment of action spotting for video analysis. In this work, we introduce *OSL-ActionSpotting*, a Python library that unifies different action spotting algorithms to streamline research and applications in sports video analytics. *OSL-ActionSpotting* encapsulates various state-of-the-art techniques into a singular, user-friendly framework, offering standardized processes for action spotting and analysis across multiple datasets. We successfully integrated three cornerstone action spotting methods into OSL-ActionSpotting, achieving performance metrics that match those of the original, disparate codebases. This unification within a single library preserves the effectiveness of each method and enhances usability and accessibility for researchers and practitioners in sports analytics. By bridging the gaps between various action spotting techniques, OSL-ActionSpotting significantly contributes to the field of sports video analysis, fostering enhanced analytical capabilities and collaborative research opportunities. The scalable and modularized design of the library ensures its long-term relevance and adaptability to future technological advancements in the domain.**

Keywords. Video understanding, Action spotting, Sports Analytics, Python library, Benchmark, Algorithms.

I. Introduction

Action spotting in sports videos is a critical task in sports analytics, allowing for the detailed analysis of player actions, game events, and overall performance. This process is instrumental in enhancing strategies, training programs, and audience engagement. However, the diversity of action spotting methodologies has led to a fragmented landscape where integrating these techniques is challenging. Existing systems often operate in isolation, leading to inefficiencies and difficulties in achieving a comprehensive analytical perspective. The necessity for consistency, efficiency, and scalability in sports analytics drives the need for a unified approach in action spotting. Moreover, the rapid evolution of video analysis technologies demands a flexible and adaptable framework that can incorporate new methods and adapt to changing analytical requirements.

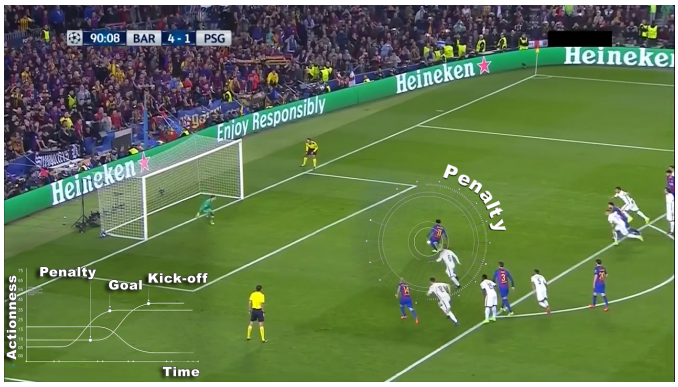

Fig. 1. **OSL-ActionSpotting** is a plug-and-play library that unifies action spotting algorithms. The design of OSL-ActionSpotting is inherently versatile, making it applicable to a broad spectrum of sports video analyses. This adaptability ensures that OSL-ActionSpotting can facilitate the development of novel action spotting techniques, and accelerate the deployment of these methods, providing a robust and comprehensive tool for researchers and analysts in various sports domains.

In response to these challenges, we propose OSL-ActionSpotting, a comprehensive library that integrates multiple state-of-the-art action spotting methods into a single, cohesive framework, illustrated in Figure 1. This library aims to standardize action spotting across various sports disciplines and enhances the accessibility and efficiency of these analyses for researchers and practitioners alike. The library is available at https://github.com/OpenSportsLab/OSL-ActionSpotting.

Contributions: (i) We release the first action spotting Python library that includes the most impactful algorithms developed in the literature, into a modularized framework that will empower future algorithmic development. **(ii)** We develop a video data loader based on DALI, that efficiently pre-process videos to be fed into our algorithm, enabling faster training of end-to-end algorithms. **(iii)** We formalize a novel action spotting dataset format in a unique JSON that enables the usage of OSL-ActionSpotting's algorithms on new video datasets.

II. Related Work

A. Sports Video Understanding

The field of sports video analysis has evolved into an important research area due to the intricate nature of sports video understanding [1], [2]. Early approaches primarily focused on video classification tasks [3], which included recognizing distinct actions [4] and categorizing various phases of gameplay [5]. Research expanded into different domains of sports video understanding, covering areas such as player detection [6], tracking [7], image segmentation [8], and player identification [9]. It also explored the analysis of tactics [10]. It investigated specific aspects like pass completion likelihood [11], 3D ball positioning in basketball [12], and the reconstruction of shuttle trajectories in badminton videos [13]. To facilitate research in this domain, numerous research collectives have made available comprehensive datasets, including those by Pappalardo *et. al.* [14], Yu *et. al.* [15], SoccerTrack [16], SoccerDB [17], and DeepSportRadar [18]. The SoccerNet dataset, introduced by Giancola *et. al.* [19], has set benchmarks for over 12 specific tasks in soccer video analytics. These include areas such as action spotting [20], camera calibration [21], [22], player tracking [23] and re-identification [21], multi-view foul detection [24], comprehensive video captioning [25], explainability [26], depth estimation [27] and game state reconstruction [28]. These datasets have become a cornerstone for yearly competitions [29], [30], promoting cooperative research efforts in the field of sports video analysis. In this work, we propose an easy-to-use library for action spotting, *i.e.* the task of localizing events in untrimmed videos, identified with a single timestamp.

B. Action spotting

Giancola *et. al.* [19] introduced the task of action spotting, defined as the process of localizing key actions within long untrimmed videos, such as penalties, goals, or corner kicks in football. This task differs from temporal activity localization, where activities are identified over durations [31]. In contrast, action spotting focuses on the precise moment of an event, with a single timestamp, in accordance with football regulations [32]. Giancola *et. al.* [19] propose a first baseline to solve this task based on learnable pooling methods, later improved with temporally-aware pooling [33]. Rongved *et. al.* [34] developed a method using 3D ResNet on video frames analyzed in a sliding window of five seconds. Multimodal techniques were explored by Vanderplaetse *et. al.* [35] and Xarles *et. al.* [29], who merged visual and auditory data for action detection. Further advancements include Cioppa *et. al.*'s [36], [37] context-aware loss function that leverages temporal dynamics, Vats *et. al.*'s [38] multi-tower CNN for managing action localization uncertainty, and Tomei *et. al.*'s [39] approach of refining feature extraction with a masking strategy targeting post-action frames. The current state-of-the-art on SoccerNet-v2 for published methods has been set by Denize *et. al.* [40], with their end-to-end methodology. This method outperforms previous leaders like Soares *et. al.* [41], who employed an anchor-based strategy, and Hong *et. al.* [42], pioneers of the precise temporal spotting (PTS) technique. The PTS approach integrates an end-to-end trainable feature extraction and spotting mechanism, utilizing a lightweight RegNet architecture enhanced with GSM [43] and GRU [44] modules. This surpassed the 2022 challenge's top contenders, showcasing advancements over earlier methods like spatio-temporal encoders [45], graph-based processing layers [46], and transformer models [47]. In this work, we re-implement feature-based and end-to-end methods from cornerstone action spotting algorithms [36], [33], [42], [48] into a unified library.

III. OSL-ActionSpotting

OSL-ActionSpotting is a streamlined, plug-and-play library tailored for efficient action spotting in sports videos, suitable for football and other sports. It merges various advanced action spotting algorithms into one framework, promoting modularity and facilitating further research development in this field.

A. Definitions

Action spotting in sports videos is the task of identifying and anchoring temporally specific actions semantic within a video. Let V be a video with frames $i_1, i_2, ..., i_T$, where T is the total number of frames. The objective is to detect the set of n actions $A = \{a_1, a_2, ..., a_n\}$ in V, each associated with a timestamp t_i. Formally, an action spotting algorithm is defined as a function $\mathcal{F} : V \rightarrow A \times T$, mapping the video V to a set of action-timestamp pairs (a_i, t_i). This function aims to accurately align the predicted timestamps with the actual occurrence times of actions in the video. The algorithmic structure can be segmented into three primary components: the backbone, the neck, and the head.

Backbone. The backbone is the fundamental part of the action spotting algorithm, primarily responsible for feature extraction. Given a video V with frames $i_1, i_2, ..., i_T$, the backbone processes these frames to extract a set of features $F = \{f_1, f_2, ..., f_T\}$. Formally, this can be represented as:

$$F = Backbone(i_1, i_2, ..., i_T)\,,$$

where F is the feature set corresponding to the frames in V.

Neck. The neck serves as an intermediate processing layer that refines and transforms the features F extracted by the backbone. Its purpose is to enhance the feature representation for more effective action spotting. The processed features F' can be expressed as:

$$F' = Neck(F)\,,$$

where F' are the neck features input to the spotting head.

Head. The head is the final component of the algorithm, responsible for action identification and timestamp assignment. It utilizes the refined features F' to identify actions A and their corresponding timestamps T. This can be formulated as:

$$(A, T) = Head(F')\,,$$

where A is the set of detected actions and T their timestamps.

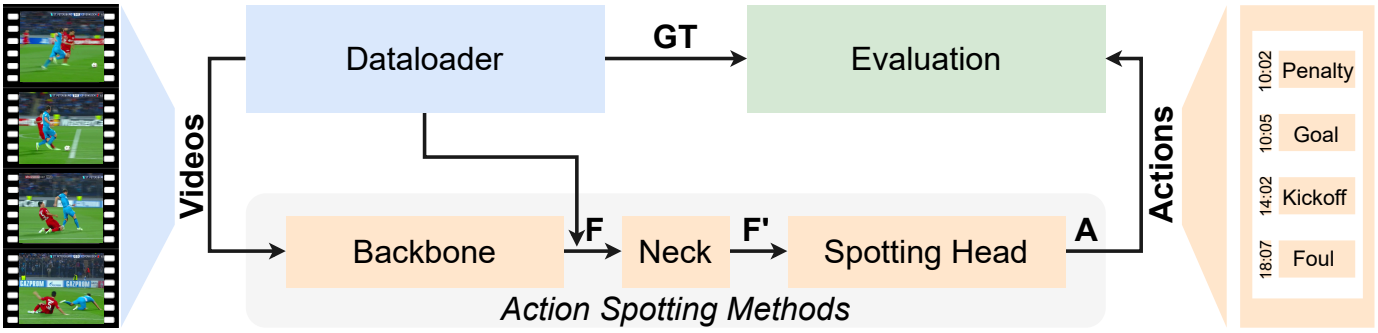

Fig. 2. **OSL-ActionSpotting** contains methods for efficient data loading, modularized action spotting, and comprehensive evaluation. The dataloader provides videos V or frame features F. The backbone converts the videos V into frame features F. The neck lifts the frame features F into neck features F', further processed by the spotting head that predicts the actions A. The evaluation compares the action predictions A with the ground truth GT.

Together, these components form the complete action spotting algorithm, where the backbone extracts raw features from video frames, the neck refines these features, and the head utilizes the refined features to detect and timestamp actions. The pipeline is illustrated in Figure 2.

B. Action Spotting Algorithms

The OSL-ActionSpotting library incorporates a variety of modular action spotting algorithms, each consisting of a backbone, neck, and head, to address different aspects of sports video analysis. The unified framework of OSL-ActionSpotting is designed to operate with these algorithms, Here, we detail some of the key action spotting algorithms already integrated into OSL-ActionSpotting.

Temporally-Aware Learnable Pooling [33]. *Backbone*: The features are pre-extracted offline using a CNN (*e.g.* ResNet), focusing on detailed spatial analysis of video frames. *Neck*: It applies learnable and temporal pooling to these features, emphasizing temporal dynamics and relevance to action spotting. The learnable aspect is based on feature clustering (*e.g.* NetVLAD), while the temporal aspect learns specific pooling before and after the action of interest. *Head*: It processes further the pooled features with linear layers to predict the actions and their timestamps as a clip classification task.

Context-Aware Loss Function (CALF) [36]. *Backbone*: CALF utilizes the same pre-extracted features based on deep CNNs to learn spatial frame features. *Neck*: The segmentation module learns the context of the action to spot, enhancing the features with semantic information around the actions. *Head*: CALF employs the context-enhanced features from the neck in its spotting module for precise action localization regression for action candidates within the video clips.

Precise Temporal Spotting (PTS) [42]. *Backbone*: PTS learns spatio-temporal features end-to-end with the remaining of the architecture. It typically relies on lighter spatial CNN such as RegNetY with intermediate temporal aggregation with GSM or TSM. *Head*: PTS does not contain any neck, it directly processes the backbone features in its recurrent spotting head (*e.g.* GRU), which achieves a precise localization of the actions through a frame classification task.

These algorithms have been incorporated in OSL-ActionSpotting, demonstrating the library's capability to cater to diverse needs in sports video analysis. Through the systematic integration of these diverse algorithms, OSL-ActionSpotting offers a robust and flexible solution for action spotting, adaptable to the specific requirements of different sports and analytical objectives.

C. Data Handling and Processing

In OSL-ActionSpotting, the data handling and processing are paramount to ensure reproducible performance of action spotting algorithms. The library accommodates two primary types of data loaders: feature-based and video-based, gathered into a unified action spotting dataset format.

Feature-Based Data Loader. This loader is designed for scenarios where features are pre-extracted and stored separately from the raw video data. It is optimized for quick access and loading of these features, facilitating efficient processing in use cases where the action spotting algorithms rely on pre-computed feature sets. The feature-based data loader streamlines the workflow by directly ingesting these features into the action spotting models, bypassing the need for a backbone and thus speeding up the analysis process.

Video-Based Data Loader. The video-based data loader, on the other hand, is tailored for processing raw video files directly, extracting features online as part of the action spotting pipeline. OSL-ActionSpotting leverages both the OpenCV video reader and the NVIDIA DALI (Data Loading Library) for this purpose, an efficient video data loader that significantly accelerates the preprocessing and feature extraction stages. DALI optimizes the data pipeline, handling tasks like decoding, cropping, and resizing the video directly on the GPU, which minimizes the latency and computational overhead associated with these operations. The DALI-integrated video-based data loader in OSL-ActionSpotting streamlines the training process by efficiently handling large-scale video datasets, ensuring fast and scalable video processing.

Unified action spotting dataset format. This new dataset format is a core aspect of OSL-ActionSpotting, and is a novelty for the field of action spotting. It is designed to harmonize the

input data structure across various action spotting datasets, and encapsulates the essential elements of action spotting. Inspired by the COCO format for image understanding, it is a single JSON file that includes the path of all videos or features, along with their annotations and associated metadata. We envision this format to facilitate the integration of new datasets in OSL-ActionSpotting, enabling easier training, fine-tuning, testing, and deployment on new domains.

In summary, OSL-ActionSpotting's data handling, with its feature-based and video-based data loaders, as well as its unified action spotting dataset JSON format, offers a streamlined and adaptable framework for efficient action spotting training on novel domain, optimizing data preparation whether using pre-extracted features or raw video content.

D. Evaluation metrics

In OSL-ActionSpotting, the evaluation framework for action spotting algorithms focuses on measuring their effectiveness and accuracy using two main metrics: loose and tight average mean Average Precision (mAP).

Loose Average mAP: This metric assesses the algorithm's performance with a lenient approach, allowing predictions to be considered correct if they fall within a broad temporal window around the actual event, ranging from 5 to 60 seconds. It gauges the algorithm's general ability to detect actions without requiring pinpoint temporal accuracy.

Tight Average mAP: Contrasting with the loose approach, the tight average mAP enforces a stringent temporal matching criterion, where predictions must closely align with the ground truth action timestamps, within a 1 to 5 second tolerance. This metric evaluates the algorithm's precision in accurately localizing actions in time.

These two mAP metrics provide a comprehensive view of an action spotting algorithm's performance in OSL-ActionSpotting, balancing between general detection capabilities and precise temporal localization. Through this evaluation approach, OSL-ActionSpotting ensures a robust assessment of action spotting algorithms, facilitating the refinement and enhancement of their performance in sports video analysis.

E. User-friendly Tools and Module

OSL-ActionSpotting utilizes the PyTorch Lightning framework to offer streamlined tools for training, inference, and evaluation, enhancing the ease of use and efficiency in training action spotting algorithms. The **training tools** facilitate model training with automated batch processing and checkpointing, allowing for flexible and efficient model development. The **inference tools** enable quick and accurate action detection in videos, optimized for both real-time and batch-processing environments. The **evaluation tools** provide detailed performance metrics to assess and compare the effectiveness of action spotting models within OSL-ActionSpotting. These tools, integrated in OSL-ActionSpotting with PyTorch Lightning, simplify the deep learning workflow, making advanced action spotting accessible and manageable for users across the sports analytics field.

IV. Experiments / Benchmark

We validate the reproducibility of our OSL-ActionSpotting implementation with respect to established methods.

Dataset. We focus our experiments on the SoccerNet-v2 action spotting dataset [20], as it is the most widely used dataset in the literature. Yet, more datasets can be tested with our library. SoccerNet-v2 provides videos for 500 football games from the big five European leagues, fully annotated with more than 110,458 temporally-anchored actions, from 17 common action classes. We refer to the SoccerNet-v2 paper [20] for a more detailed description of the dataset.

Methods. For the feature-based methods, we focus our experiments on the pre-extracted ResNet-152 features reduced with a PCA at a dimension of 512, provided with the SoccerNet-v2 dataset [20]. For the *Temporally-Aware Learnable Pooling* methods [33], we test the non-parametric pooling necks (MaxPooling, AvgPool), the learnable pooling necks (NetVLAD++, NetRVLAD++) and the temporally-aware pooling necks (Maxpool++, Avgpool++, NetVLAD++, NetRVLAD++). For the cluster-based approaches, we set the number of clusters to 64. The head is a single linear layer that projects the neck features to a dimension equal to the class number, followed by a softmax activation for the logits. We keep the training parameters similar to the original implementation. For the *Context-Aware Loss Function (CALF)* method [36], we define the segmentation module as the neck, and the spotting head as the head. We keep the training parameters and the temporal parameters similar to the original implementation. For the end-to-end approach *Precise Temporal Spotting (PTS)* [42], we use the 224p videos at 25fps as input. The backbone processes the frames at 2fps with RegNetY (RNY) and a GSM temporal shift. PTS does not have any neck, and we select the GRU head to accumulate the temporal information and predict the class activations. We keep the training parameters similar to the original implementation meant for the SoccerNet-v2 dataset.

Results. We present in Table I the results of our experiments. We reproduce the same performances as the original implementation in terms of loose and tight Average-mAP. We report the training time (per epoch between parenthesis), as well as the time for inference of the complete testing set of SoccerNet-v2 composed of 100 videos of 45min. The timing for all methods was estimated on 1 GPU RTX1080, 2 CPU cores and 32GB RAM, except for the PTS that required 4 GPU RTX A500 and 80GB of RAM.

V. Discussions

Performances reproduction. OSL-ActionSpotting successfully meets its primary goal of reproducing the performance metrics documented in existing literature, demonstrating its capability and reliability in the domain of sports video analysis.

DALI: accelerated video loading. Integrating DALI for video loading in OSL-ActionSpotting improved the processing of the video data, overcoming the inefficiencies of traditional CPU-based loading methods like OpenCV video and PyTorch image data loaders. This shift to GPU processing with

TABLE I
OSL-ACTIONSPOTTING: REPRODUCED RESULTS.

Method	Average-mAP		Timing (sec)	
	loose	tight	Training	Testing
Maxpool [33]	20.8	2.1	1268 (7)	118
Avgpool [33]	29.4	2.3	886 (7)	118
NetVLAD [33]	46.3	4.4	1328 (7)	119
NetRVLAD [33]	44.8	4.8	1371 (7)	118
Maxpool++ [33]	32.2	5.3	1172 (7)	118
Avgpool++ [33]	40.8	6.9	974 (7)	118
NetVLAD++ [33]	51.5	8.1	1185 (8)	120
NetRVLAD++ [33]	49.8	8.3	1147 (8)	118
CALF [36]	39.5	12.4	1639 (4)	26
PTS (RNY002) [42]	70.4	63.9	140700 (1200)	1500
PTS (RNY008) [42]	72.2	66.2	249780 (1500)	1800

DALI results in significant time improvement, streamlining OSL-ActionSpotting 's data pipeline. Compared to OpenCV video data loaders, best cases result in around 33% time improvement for the training steps and 50% improvement for each epoch's validation steps. Furthermore, compared to the PyTorch image data loader, we remove the need of pre-extracting frames from videos. Pre-extracting frame typically requires a significant extra volume space to store this data, typically around 10 fold, and presents a reading bottleneck due to a large number of file access. By adopting DALI, OSL-ActionSpotting enhances pre-processing efficiency and sets a new standard for speed, storage, and performance in sports video analysis, paving the way for more sophisticated and expansive video pre-processing for action spotting research, including larger resolution and higher frame rate.

JSON-based action spotting dataset format. The JSON-based action spotting dataset format in OSL-ActionSpotting marks a significant advance in standardizing data handling for sports analytics, aligning with algorithm needs by organizing timestamps, action labels, and metadata cohesively. This format enhances interoperability and eases integration with analytics tools, streamlining dataset preparation and parsing for efficient model training and evaluation. Standardization with OSL-ActionSpotting's JSON format boosts consistency in video analysis, supporting more scalable and collaborative research efforts, and facilitating innovation in sports analytics.

VI. CONCLUSION

OSL-ActionSpotting marks a novel development in sports video analysis by introducing the first Python library to consolidate major action spotting algorithms into a modular framework, fostering future algorithmic development. Its efficient video data loader, powered by DALI, streamlines video pre-processing, facilitating rapid training of end-to-end algorithms. Furthermore, the introduction of a novel JSON-based action spotting dataset format with OSL-ActionSpotting enhances the adaptability and application of its algorithms across various datasets and video analyses. We envision OSL-ActionSpotting as the one-stop Python library for action spotting algorithms, that will centralize the efforts of the sports analytics community into an easy-to-use modular library, accelerating the development and deployment of these tools.

ACKNOWLEDGMENT

A. Cioppa is funded by the F.R.S.-FNRS. This work was partly supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research through the Visual Computing Center (VCC) funding and the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI). Bruno Cabado wish to thanks the Axencia Galega de Innovación the grant received through its Industrial Doctorate program (23/IN606D/2021/2612054). CITIC is funded by Xunta de Galicia (ED431G 2019/01) and ERDF funds.

REFERENCES

[1] 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.*, vol. 159, pp. 3–18, Jun. 2017.

[2] B. T. Naik, M. F. Hashmi, N. D. Bokde, and Z. M. Yaseen, "A comprehensive review of computer vision in sports: Open issues, future trends and research directions," *Appl. Sci.*, vol. 12, pp. 1–49, Apr. 2022.

[3] F. Wu, Q. Wang, J. Bian, H. Xiong, N. Ding, F. Lu, J. Cheng, and D. Dou, "A survey on video action recognition in sports: Datasets, methods and applications," *arXiv*, vol. abs/2206.01038, 2022.

[4] 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*, (Salt Lake City, UT, USA), pp. 1846–1855, Jun. 2018.

[5] B. Cabado, B. Guijarro-Berdiñas, and E. J. Padrón, "Real-time classification of handball game situations," in *IEEE Int. Conf. Tools Artif. Intell. (ICTAI)*, (Macao, China), pp. 686–691, Oct. 2022.

[6] R. Vandeghen, A. Cioppa, and M. Van Droogenbroeck, "Semi-supervised training to improve player and ball detection in soccer," in *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, (New Orleans, LA, USA), pp. 3480–3489, Jun. 2022.

[7] A. Maglo, A. Orcesi, and Q.-C. Pham, "Efficient tracking of team sport players with few game-specific annotations," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, (New Orleans, LA, USA), pp. 3460–3470, Jun. 2022.

[8] A. Cioppa, A. Deliege, M. Istasse, C. De Vleeschouwer, and M. Van Droogenbroeck, "ARTHuS: Adaptive real-time human segmentation in sports through online distillation," in *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, (Long Beach, CA, USA), pp. 2505–2514, Jun. 2019.

[9] V. Somers, C. De Vleeschouwer, and A. Alahi, "Body part-based representation learning for occluded person Re-Identification," in *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, (Waikoloa, HI, USA), pp. 1613–1623, Jan. 2023.

[10] G. Suzuki, S. Takahashi, T. Ogawa, and M. Haseyama, "Team tactics estimation in soccer videos based on a deep extreme learning machine and characteristics of the tactics," *IEEE Access*, vol. 7, pp. 153238–153248, 2019.

[11] A. Arbués Sangüesa, A. Martín, J. Fernández, C. Ballester, and G. Haro, "Using player's body-orientation to model pass feasibility in soccer," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, (Seattle, WA, USA), pp. 3875–3884, Jun. 2020.

[12] G. Van Zandycke and C. De Vleeschouwer, "3D ball localization from a single calibrated image," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, (New Orleans, LA, USA), pp. 3471–3479, Jun. 2022.

[13] P. Liu and J.-H. Wang, "MonoTrack: Shuttle trajectory reconstruction from monocular badminton video," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, (New Orleans, LA, USA), pp. 3512–3521, Jun. 2022.

[14] L. Pappalardo, P. Cintia, A. Rossi, E. Massucco, P. Ferragina, D. Pedreschi, and F. Giannotti, "A public data set of spatio-temporal match events in soccer competitions," *Sci. Data*, vol. 6, pp. 1–15, Oct. 2019.

[15] J. Yu, A. Lei, Z. Song, T. Wang, H. Cai, and N. Feng, "Comprehensive dataset of broadcast soccer videos," in *IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, (Miami, FL, USA), pp. 418–423, Apr. 2018.

[16] A. Scott, I. Uchida, M. Onishi, Y. Kameda, K. Fukui, and K. Fujii, "SoccerTrack: A dataset and tracking algorithm for soccer with fish-eye and drone videos," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, (New Orleans, LA, USA), pp. 3568–3578, Jun. 2022.

[17] 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)*, (Seattle, WA, USA), p. 1–8, ACM, Oct. 2020.

[18] 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)*, (Ottawa, Ontario, Can.), pp. 23–29, ACM, Oct. 2023.

[19] 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)*, (Salt Lake City, UT, USA), pp. 1792–179210, Jun. 2018.

[20] 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 Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, (Nashville, TN, USA), pp. 4508–4519, Jun. 2021.

[21] 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*, vol. 9, pp. 1–9, Jun. 2022.

[22] F. Magera, T. Hoyoux, O. Barnich, and M. Van Droogenbroeck, "A universal protocol to benchmark camera calibration for sports," in *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, (Seattle, WA, USA), Jun. 2024.

[23] 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*, (New Orleans, LA, USA), pp. 3490–3501, Jun. 2022.

[24] 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)*, (Vancouver, Can.), pp. 5086–5097, Jun. 2023.

[25] 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)*, (Vancouver, Can.), pp. 5074–5085, Jun. 2023.

[26] 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 Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, (Seattle, WA, USA), Jun. 2024.

[27] 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 Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, (Seattle, WA, USA), Jun. 2024.

[28] V. Somers, V. Joos, S. Giancola, A. Cioppa, 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 Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, (Seattle, WA, USA), Jun. 2024.

[29] 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, and *et. al.*, "SoccerNet 2022 challenges results," in *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, (Lisbon, Port.), pp. 75–86, ACM, Oct. 2022.

[30] 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, and *et. al.*, "SoccerNet 2023 challenges results," *arXiv*, vol. abs/2309.06006, 2023.

[31] 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)*, (Boston, MA, USA), pp. 961–970, Jun. 2015.

[32] IFAB, "Laws of the game," tech. rep., The International Football Association Board, Zurich, Switzerland, 2022.

[33] S. Giancola and B. Ghanem, "Temporally-aware feature pooling for action spotting in soccer broadcasts," in *IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Nashville, TN, USA), pp. 4490–4499, Jun. 2021.

[34] 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*, vol. 3, pp. 1–25, Dec. 2021.

[35] 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*, (Seattle, WA, USA), pp. 3921–3931, Jun. 2020.

[36] 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)*, (Seattle, WA, USA), pp. 13123–13133, Jun. 2020.

[37] 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*, (Nashville, TN, USA), pp. 4532–4541, Jun. 2021.

[38] 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," *arXiv*, vol. abs/2004.06172, 2020.

[39] 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)*, (Milan, Italy), pp. 7699–7706, Jan. 2021.

[40] 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 Winter Conf. Appl. Comput. Vis. Work. (WACVW)*, (Waikoloa, HI, USA), pp. 530–540, Jan. 2024.

[41] 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)*, (Bordeaux, France), pp. 2796–2800, Oct. 2022.

[42] 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 of *Lect. Notes Comput. Sci.*, (Tel Aviv, Israël), pp. 33–51, Springer Nat. Switz., 2022.

[43] S. Sudhakaran, S. Escalera, and O. Lanz, "Gate-shift networks for video action recognition," in *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, (Seattle, WA, USA), pp. 1099–1108, Jun. 2020.

[44] K. Cho, B. van Merrienboer, D. Bahdanau, and Y. Bengio, "On the properties of neural machine translation: Encoder-decoder approaches," in *Workshop on Syntax, Semantics and Structure in Statistical Translation*, (Doha, Qatar), pp. 103–111, Association for Computational Linguistics, Oct. 2014.

[45] 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)*, (Lisbon, Port.), pp. 87–92, ACM, Oct. 2022.

[46] A. Cartas, C. Ballester, and G. Haro, "A graph-based method for soccer action spotting using unsupervised player classification," in *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, (Lisbon, Port.), pp. 93–102, ACM, Oct. 2022.

[47] 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)*, (Lisbon, Port.), pp. 103–109, ACM, Oct. 2022.

[48] B. Cabado, A. Cioppa, S. Giancola, A. Villa, B. Guijarro-Berdiñas, E. Padrón, B. Ghanem, and M. Van Droogenbroeck, "Beyond the Premier: Assessing action spotting transfer capability across diverse domains," in *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, (Seattle, WA, USA), Jun. 2024.