Beyond the Premier:
Assessing Action Spotting Transfer Capability Across Diverse Domains

Bruno Cabado[1,2] Anthony Cioppa[3,4] Silvio Giancola[4] Andrés Villa[4]
Bertha Guijarro-Berdiñas[2] Emilio J. Padrón[2] Bernard Ghanem[4] Marc Van Droogenbroeck[3]
[1] Cinfo [2] CITIC Research Center, Universidade da Coruña [3] TELIM, University of Liège [4] IVUL, KAUST

Abstract

Football stands as one of the most successful sports in history thanks to the plethora of professional leagues broadcasted worldwide followed by avid fans, further fueled by the abundance of amateur and grassroots leagues across nearly every country, encompassing countless players who devote their time to the sport. Despite the tremendous amount of visual data available worldwide for developing automatic systems to extract game events, most efforts focus on the few professional league matches. However, the recording quality and broadcasts editing vary considerably across leagues, creating a disparity in the analytical capabilities of deep learning models. This paper delves into an analysis of how action spotting models transfer to diverse domains, analyzing the performance gap between various types of broadcasts. In particular, we investigate the transfer capability of state-of-the-art action spotting models across leagues, from amateur to professional, and broadcast quality, from AI-piloted camera to professional broadcast editing. Our analysis shows that transferring across leagues is challenging, with the most impactful feature being broadcasting editing quality. This analysis paper therefore seeks to spotlight this pressing issue and catalyze future research endeavors in the field of domain adaptation for action spotting methods.

1. Introduction

In recent years, the exponential growth in both the availability and accessibility of high-quality datasets has acted as a catalyst for unprecedented advancements in artificial intelligence (AI) across various domains. This phenomenon has been particularly pronounced within the realm of sports, where datasets have emerged as foundational pillars for AI-driven sports analytics. SoccerNet [25], among its contemporaries, continues to undergo iterative refinement, perpetually expanding its breadth and depth by incorporating additional data subsets tailored to address specific tasks or emerging analytical demands. This evolutionary trajectory has been instrumental in propelling innovations across a spectrum of areas within sports analytics, including player performance analysis, tactical insights, and decision-making processes within the sporting arena.

Figure 1. **Domain gap between different leagues and broadcast quality.** We showcase four domains: (1) professional league broadcast from the SoccerNet dataset, (2) professional league unedited main camera feeds from the Swiss Super League, (3) semi-professional league AI-piloted camera from the BGL Luxembourg National Division, and (4) amateur league AI-piloted camera from Sevilla. Our analysis shows that transferring action spotting models from professional broadcasts to amateur AI-piloted camera is challenging due to different recording and broadcast editing quality.

Despite the valuable contributions of datasets, a common characteristic is their pronounced bias towards elite matches. These datasets predominantly draw from top-tier professional leagues, thereby fostering a dataset landscape characterized by an overrepresentation of elite competition. While this emphasis on elite matches ensures unparalleled data quality, it concurrently gives rise to a lack of diversity in terms of both gameplay scenarios and broadcast quality. This inherent limitation presents formidable challenges, as techniques and models trained exclusively on such homog-

Contacts: bruno.cabado@udc.es, anthony.cioppa@uliege.be.

enized data may exhibit limited generalization capabilities when confronted with more diverse and heterogeneous domains, including lower-tier leagues, amateur competitions, or alternative broadcasts such as AI-piloted cameras.

Addressing this critical gap in the automatic detection of events in sports broadcast, the primary objective of this paper is to analyze the transfer capability of action spotting models within different domains. To do so, we explore four distinct domains to shed light on the challenge posed by domain gaps. These domains include: (1) professional league with edited broadcasts, encompassing the professional six leagues from the SoccerNet dataset, (2) unedited main camera feeds from the Swiss Super League, (3) AI-piloted camera feeds from the semi-professional BGL Luxembourg National Division, and (4) AI-piloted camera feeds from an amateur league in Sevilla. Our analysis reveals the performance gap when transferring action spotting models from professional broadcasts to amateur AI-piloted cameras, owing to disparities in recording and broadcast editing standards. By transcending the conventional dichotomy between elite and non-elite football, our research endeavors to augment the robustness, applicability, and real-world efficacy of AI-driven solutions in sports analytics and decision support systems.

Contributions. **(i)** We introduce a novel benchmark designed explicitly to evaluate domain transfer capabilities of action spotting models, encompassing four distinct domains spanning different leagues and broadcast formats. **(ii)** We provide a thorough analysis and insights into the inherent variations among the different leagues of the SoccerNet dataset providing insights on the opportunities for future research on domain adaptation of action spotting models. **(iii)** Leveraging state-of-the-art models trained on SoccerNet, we conduct a comprehensive evaluation of their adaptability to new domains, including amateur leagues and AI-piloted cameras

2. Related Work

2.1. Video understanding

Historically, video understanding has lagged behind image understanding due to the absence of expansive video datasets like ImageNet or CIFAR-100 [19, 39] in the video domain. The advent of significant video comprehension datasets, such as UCF101 [63], ActivityNet [6], YouTube-8M [1], and Kinetics [36]], has sparked increased interest in the field. Popular video understanding tasks include video classification [22, 35, 53], action recognition [58, 75], video captioning [24, 38, 78], and video generation [41]. The interest in crafting video models capable of capturing spatio-temporal features has grown notably. The Temporal Segment Network (TSN) [76] aggregates features across multiple temporal video segments to enhance recognition performance. In a parallel development, Tran *et al*. [68] introduced a novel spatio-temporal convolutional block, R(2+1)D, assessing its impact on action recognition models. A more recent innovation, the Multiscale Vision Transformer (MViT) [21, 42], synergizes convolutional neural networks (CNNs) and transformers in video classification to capture spatial and temporal nuances.

Recent challenges in the video domain [6] involve activity localization, *i.e*., identifying temporal boundaries of activities in long untrimmed videos. Two-stage methodologies, including proposal generation [5] and subsequent classification [4], have proliferated after object localization. SSN [81] models each action instance with a structured temporal pyramid, while TURN TAP [23] predicts action proposals and regresses their temporal boundaries. In a dynamic optimization approach, GTAN [46] fine-tunes the temporal scale of each action proposal using Gaussian kernels. Other methodologies like BSN [77], MGG [45], and BMN [43] use regression techniques to estimate activity boundaries, showcasing state-of-the-art performances on benchmark datasets such as ActivityNet 1.3 [6] and Thumos'14 [31]. On a different front, ActionSearch [2] addresses the spotting task iteratively, learning to predict the subsequent frame for spotting a given activity. In this paper, we study the transferability of video understanding models, particularly action spotting models accross different domains.

2.2. Sports understanding

The complexity of understanding sports videos has propelled the field into a prominent research area [51, 52, 66]. Initially, methodologies were centered on video classification [79], involving the identification of specific actions [37, 56] and the segmentation of diverse game phases [7, 14]. Other avenues of research delved into varied aspects of sports understanding, encompassing player detection [71], tracking [48], segmentation [13], and identification [61, 74], as well as tactics analysis in sports like football and fencing [65, 83]. Additional focal points included pass feasibility [3], 3D ball localization in basketball [69], and 3D shuttle trajectory reconstruction for badminton videos [44].

To support research in the field, several research groups released extensive datasets, such as the one of Pappalardo *et al*. [54], Yu *et al*. [80], SoccerTrack [57], SoccerDB [34], and DeepSportRadar [33, 70]. Recently, the SoccerNet dataset, introduced by Giancola *et al*. [25], offers benchmarks for more than 12 distinct tasks related to soccer video understanding. These tasks span action spotting [18], camera calibration [10, 47], player tracking [15] and re-identification [10, 49], multi-view foul recognition [28], dense video captioning [50], explainability [29], depth estimation [40], and game state reconstruction [62]. The

datasets serve as a platform for yearly competitions [16,26], fostering collaborative research in sports video understanding. In this paper, we leverage the SoccerNet action spotting dataset as it comprises 6 professional leagues edited broadcasts, as well as the videos from the SoccerNet-Tracking dataset comprising 9 available main camera feeds from the Swiss Super League. Furthermore, we explore two novel video datasets of AI-piloted camera in semi-professional and amateur leagues.

2.3. Action spotting

Detecting specific actions, known as action spotting, is a crucial aspect of understanding football videos, involving the precise localization of events like penalties, goals, or corners within untrimmed football broadcast videos. Unlike temporal activity localization [6], action spotting assigns a single timestamp to describe events, aligning with football rule definitions [32]. Notably, large-scale datasets like SoccerNet [25] have expanded to 17 classes to encompass all possible game actions [18].Giancola *et al*. [25] initially introduced the action spotting task and developed a first baseline, utilizing temporal pooling, later refining their approach by incorporating temporal context [27]. Rongved *et al*. [55] proposed a 3D ResNet applied directly to video frames in a 5-second sliding window fashion. Multimodal approaches by Vanderplaetse *et al*. [72] and Xarles *et al*. [26] combined visual and audio features. Cioppa *et al*. [11, 12] introduced a context-aware loss function to capture temporal context, while Vats *et al*. [73] employed a multi-tower CNN accounting for action location uncertainty, and Tomei *et al*. [67] fine-tuned feature extractors with a masking strategy for post-action frames.

The current state-of-the-art on SoccerNet-v2 is established by Denize *et al*. [20], winners of the 2023 SoccerNet challenge, using an end-to-end approach. This approach significantly surpasses Soares *et al*. [59, 60], an anchor-based approach, and Hong *et al*. [30], the first precise temporal spotting (PTS) method with an end-to-end trained feature extractor and spotting head works, the 2022 challenge winner and runner-ups respectively. The latter approach relies on a light-weight RegNet architecture with GSM [64] and GRU [9] modules. Other methods explored spatio-temporal encoders [17], graph-based layers [8], and transformer architectures [82].

3. Assessing Domain Transfer

In this section, we delve into the transfer capability of action spotting models across diverse domains, a critical aspect for extending their applicability beyond their initial training environments. This concept referes to *generalization*, *i.e*., the ability of a model to perform accurately across diverse, unseen domains, which is critical for robust action spotting systems. In this work, we explore how well state-of-the-art action spotting techniques maintain their performance when tested in new domains with specific domain shifts. This section is segmented several parts, including the formal definition of the problem, the description of the different domains in the SoccerNet action spotting dataset, and the novel datasets derived from broadcasts of Swiss league matches, Luxembourg league matches, and broadcasts from Cinfo, a company specialized in automated sports event production.

3.1. Problem definition

The transfer capability across diverse domain, or generalization, problem in action spotting can be formalized as follows: Let $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$ represent the training and testing datasets, respectively representing different domains, where each dataset consists of a set of video sequences $\mathbf{V}$ and corresponding action labels $\mathbf{A}$. The objective of an action spotting model f is to learn a mapping from video frames to action labels:

$$f : \mathbf{V} \rightarrow \mathbf{A}$$

The model f is trained on $\mathcal{D}_{train}$ with the goal of minimizing a loss function $\mathcal{L}$, which measures the discrepancy between the predicted actions $f(\mathbf{V})$ and the true actions $\mathbf{A}$:

$$\min_{f} \mathcal{L}(f(\mathbf{V}_{train}), \mathbf{A}_{train})$$

where $\mathbf{V}_{train}$ and $\mathbf{A}_{train}$ are the video sequences and action labels in $\mathcal{D}_{train}$, respectively.

The challenge in generalization arises when the model f is evaluated on $\mathcal{D}_{test}$, whose domain may differ significantly from $\mathcal{D}_{train}$ in terms of distribution, quality, or context. The generalization capability of f is then assessed based on its performance on $\mathcal{D}_{test}$:

$$\mathcal{P}(f) = \mathcal{L}(f(\mathbf{V}_{test}), \mathbf{A}_{test})$$

where $\mathbf{V}_{test}$ and $\mathbf{A}_{test}$ are the video sequences and action labels in $\mathcal{D}_{test}$, and $\mathcal{P}(f)$ represents the performance of the model.

The primary goal of this research is to assess the transfer capabilities of existing action spotting methods across different leagues and broadcast quality levels within $\mathcal{D}_{test}$, highlighting the impact on model performance.

3.2. The SoccerNet professional league domain

For this research, we leverage the SoccerNet dataset, which is composed of edited broadcasts from six professional European leagues.The breakdown of video quantities for each league is provided in Table 1 and further detailed in this section. Regarding annotations, SoccerNet encompasses 17 distinct classes, namely: “Ball out of play”, “Clearance”, “Corner”, “Direct free-kick”,

Partition	Games	Camera		Train Games					Validation Games					Test Games				
		Type	Num	2014	2015	2016	> 2019	Total	2014	2015	2016	> 2019	Total	2014	2015	2016	> 2019	Total
EPL	95	Human	> 1	4	25	29	0	58	1	12	6	0	19	1	12	5	0	18
UEFA	101	Human	> 1	22	29	11	0	62	8	10	2	0	20	7	6	6	0	19
Ligue 1	38	Human	> 1	1	2	24	0	27	0	1	8	0	9	0	0	2	0	2
Bundesliga	53	Human	> 1	5	10	16	0	31	2	4	2	0	8	1	4	9	0	14
Serie A	96	Human	> 1	7	5	44	0	56	3	1	14	0	18	1	3	18	0	22
La Liga	117	Human	> 1	6	18	42	0	66	6	9	11	0	26	6	9	10	0	25
Super League (SWL)	9	Human	1	0	0	0	0	0	0	0	0	0	0	0	0	0	9	9
BGL League	1	AI	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
CINFO	1	AI	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1

Table 1. **Datasets and domain statistics.** We detail the number of games per dataset, splitting them per league and per year. The first six leagues come from the SoccerNet dataset and are professional European leagues broadcasts. The three bottom leagues are captured by a single camera, with the BGL and CINFO datasets being from semi-professional and amateur leagues respectively and AI-piloted cameras.

"Foul", "Goal", "Indirect free-kick", "Kick-off", "Offside", "Penalty", "Red card", "Shots off target", "Shots on target", "Substitution", "Throw-in", "Yellow card", and "Yellow->red card'.

England - EPL. This subset contains 116 videos with 12,617 annotated events for training, 38 videos with 3,808 events for validation, and 36 videos with 4,016 events for testing. It is noteworthy that the validation set for this subset lacks examples of the class "Yellow->red card".

Europe - Champions League. This subset contains 124 videos with 13,549 annotated events for training, 40 videos with 4,357 events for validation, and 38 videos with 4,222 events for testing. Notably, the test set for this subset does not include examples of the class "Red card".

France - Ligue 1. This subset includes 54 videos with 5,465 annotated events for training, 18 videos with 1,827 events for validation, and 4 videos with 384 events for testing. The validation and testing sets for this subset lack examples of the classes "Red card' and "Yellow->red card".

Germany - BundesLiga. This subset comprises 62 videos with 7,258 annotated events for training, 16 videos with 1,786 events for validation, and 28 videos with 3,270 events for testing. The validation set for this subset does not contain examples of the classes "Red card' and "Yellow->red card".

Italy - Serie A. This subset encompasses 112 videos with 13,176 annotated events for training, 36 videos with 3,986 events for validation, and 44 videos with 5,147 events for testing. All subsets include examples of all classes.

Spain - LaLiga. This subset encompasses 132 videos with 14,395 annotated events for training, 52 videos with 5,683 events for validation, and 50 videos with 5,512 events for testing. All subsets include examples of all classes.

3.3. Novel datasets and domains

Alongside the SoccerNet action spotting dataset, we provide three extra domains, corresponding to unedited main camera and AI-piloted camera broadcasts from professional, semi-professional, and amateur leagues.

Switzerland - Super League. This dataset leverages the SoccerNet-Tracking data on which action spotting labels were annotated. It comprises 9 matches from the Swiss Super league, all of which are used as test set. These matches differ primarily from those in SoccerNet in that they feature only a single camera tracking the game. Notably, this tracking is performed by a human agent using a professional-grade camera. The dataset consists of 18 videos with 2,387 events. It is important to note that this dataset does not include examples of the classes "Red card' and "Yellow->red card".

Luxembourg - BGL. Similar to the Swiss League dataset, this subset consists of a single match used as test set. This match features a single camera tracking a BGL game. However, the tracking in this league is executed by an artificial agent moving the camera in real-time, and it employs a non-professional camera for broadcasts. The dataset comprises 2 videos with 165 events. Notably, this dataset lacks examples of the classes "Offside", "Penalty", "Red card", and "Yellow->red card".

Spain - Cinfo. Similarly, as in the previous section, this dataset consists of a single match used as test set. The is produced by the company Cinfo and, like the Swiss and Luxembourg BGL leagues, feature only a single camera tracking the game. However, similarly to the Luxembourg BGL league, the tracking in this dataset is performed by an artificial agent moving the camera in real-time, and it employs a non-professional camera for broadcasts. The dataset comprises 2 videos with 164 events. Notably, this dataset lacks examples of the classes "Direct free-kick", "Penalty", "Red card", "Yellow card", and "Yellow->red card".

4. Benchmark and Results

In this section, we first provide the implementation details of the evaluated methods. Then, we delve into an evaluation and analysis of the capacity of each state-of-the-art model to transfer knowledge from the data of each league within the SoccerNet dataset to all leagues therein (Cross-league transfer within SoccerNet), as well as to novel and challenging domains such as the Swiss Super League, the Luxembourg BLG, and Cinfo (transfer between leagues

outside of SoccerNet).

Metrics. This study employs the metrics established by [25], namely tight average-mAP and loose average-mAP. These metrics utilize the same average-mAP measure but with different temporal tolerances: 1-5 seconds and 5-60 seconds, respectively. It is important to highlight that tight average-mAP, being a stricter metric, prioritizes precise temporal localization. This precision is particularly valuable for tasks requiring accurate action spotting.

4.1. Implementation details

Pooling. Pooling is the first method in our analysis, drawing from the seminal work of Giancola *et al*. [25]. From their groundwork, we adopt two feature-based pooling methods, namely MaxPooling and NetVLAD, for action spotting tasks. Both during training and testing phases, we utilize the ResNET_TF2_PCA512 features. To ensure consistency and reproducibility, we adhere to a standardized set of hyperparameters across all models. These include a batch size of 256, a chunk size of 20 seconds, an evaluation frequency of 10 epochs, a number of input features of 512, the framerate of input features as 2 frames per second, the chunk size as 60 seconds, the learning rate (LR) set to 1e-03, and a patience of 10 epochs before reducing LR using the ReduceLROnPlateau strategy. Finally, for the NetVLAD method, we employ a Non-Maximum Suppression (NMS) mechanism with a window size of 20 seconds and a threshold of 0.5 to filter positive results effectively. Conversely, for MaxPooling, a threshold of 0.0 is utilized.

CALF. The second method is the one of Cioppa *et al*. [11], known as CALF. In both the training and testing phases of this model, we have also employed the ResNET_TF2_PCA512 features. To ensure reproducibility, we have adhered to specific hyperparameters: a batch size of 32, an evaluation frequency of 20, and chunks per epoch set to 18000, the number of input features set to 512, the dimension of the capsule network set to 16, a framerate of 2 for the input features, a chunk size set to 120 seconds, and a temporal receptive field of the network set to 40 seconds. Additionally, we have specified the weights for various components in the detection loss function: lambda_coord (weight of the coordinates of the event) set to 5.0, lambda_noobj (weight of the no object detection) set to 0.5, loss_weight_segmentation (weight of the segmentation loss compared to the detection loss) set to 0.000367, and loss_weight_detection (weight of the detection loss) set to 1.0. The learning rate (LR) is set to 1e-03 with a patience of 25 epochs before reducing LR using the ReduceLROnPlateau strategy.

NetVLAD++. The third method is NetVLAD++, proposed by Giancola *et al*. [27]. This advanced technique has played a pivotal role in refining action spotting methodologies, showcasing promising outcomes, particularly in temporal modeling. Throughout our experimentation, we utilize the ResNET_TF2 features for both training and testing phases. To ensure reproducibility, we utilized the following hyperparameters: an evaluation frequency set to 10, a framerate of 2 frames per second, a window size parameter of 15 seconds, a vocabulary size of 64, a Non-Maximum Suppression (NMS) window size of 30 seconds, and a threshold of 0.0 for NMS. Additionally, we employ a batch size of 256 and set the learning rate (LR) to 1e-03. We incorporate a patience of 10 epochs before LR reduction using the ReduceLROnPlateau strategy. Furthermore, setting the seed to 0 ensures reproducibility across experiments.

E2E-Spot. The fourth method under consideration is an end-to-end approach proposed by Hong *et al*. [30]. In our evaluation, we utilized the architectures that yielded the best results in the original work. Specifically, we employed the "rgb" modality alongside the CNN architectures "rny002_gsm" and "rny008_gsm". For temporal modeling in spotting, we employed the "gru" architecture in both cases. The hyperparameters used for both training and testing are as follows: a clip length of 100, a crop dimension of 224, a batch size of 8, gradient accumulation of 1, 3 epochs for warm-up, a maximum number of epochs set to 50, a learning rate of 0.001, validation criterion based on Mean Average Precision (mAP), a dilation distance of 0 during training, and the mixup augmentation technique enabled.

COMEDIAN. The fifth and final method, COMEDIAN, proposed by Denize *et al*. [20], serves as our benchmark in this study. This methodology stands out as the top-performing approach on SoccerNet, with its code publicly available, facilitating thorough evaluation. Notably, COMEDIAN represents another end-to-end method in our analysis. Although we have not trained models from scratch, we conduct an in-depth analysis of the capabilities of the models proposed in the original work: COMEDIAN Vivit, COMEDIAN Viswin, and ensembles thereof, on the new domains introduced in this study. This comprehensive examination allows us to assess the adaptability and effectiveness of COMEDIAN in action spotting tasks, leveraging its state-of-the-art performance as a reference point for comparison.

4.2. Cross leagues transfer within SoccerNet

In this set of experiments, we delve into an extensive analysis of cross-league transfer capability within the SoccerNet Action Spotting dataset. Our investigation aims to uncover the adaptability of various algorithms trained on specific leagues within SoccerNet when applied to different leagues. The comprehensive results presented in the Table 2 encapsulate the performance metrics, namely Tight-Average-mAP and Loose-Average-mAP, which offer insights into the transfer capability of cross-leagues. Upon analyzing the results, several noteworthy observations emerge,

Train	Test	EPL		UEFA		Ligue1		Bundesliga		SerieA		LaLiga	
		Average-mAP		Average-mAP		Average-mAP		Average-mAP		Average-mAP		Average-mAP	
		tight	loose	tight	loose	tight	loose	tight	loose	tight	loose	tight	loose
EPL	MaxPool	**2.56**	**15.87**	1.74	12.77	2.79	18.10	1.86	10.55	1.2	11.45	2.21	9.3
	NetVLAD	**3.82**	**24.02**	2.38	19.37	3.63	18.32	3.00	12.83	1.76	9.65	2.29	12.65
	NetVLAD++	**13.27**	**42.41**	12.14	39.12	13.68	44.62	11.28	36.26	7.37	32.69	9.2	32.4
	CALF	**7.12**	**31.33**	6.59	25.02	**9.48**	28.15	4.64	23.71	**5.61**	19.52	5.24	19.49
	E2E-Spot rny002_gsm	**53.09**	**62.77**	51.66	60.65	55.23	63.95	**46.33**	54.7	33.52	50.9	40.14	48.86
	E2E-Spot rny008_gsm	**53.53**	**63.22**	52.37	61.07	57.03	65.95	45.56	54.01	35.4	54.68	41.12	49.29
UEFA	MaxPool	1.71	11.20	**3.33**	**18.00**	**3.77**	20.24	1.86	13.28	1.9	12.36	2.85	10.66
	NetVLAD	3.09	13.27	4.50	**25.46**	3.89	25.37	**3.83**	17.32	2.27	11.75	3.12	15.86
	NetVLAD++	9.55	36.00	**14.49**	**46.74**	**19.57**	**56.85**	11.68	**39.78**	9.61	37.43	9.29	37.81
	CALF	4.77	22.66	**9.18**	**32.93**	7.52	28.21	5.85	24.77	5.5	23.6	5.78	20.52
	E2E-Spot rny002_gsm	46.84	57.81	**60.51**	**70.30**	**59.29**	**67.90**	46.23	**55.14**	34.04	55.29	44.10	52.74
	E2E-Spot rny008_gsm	50.00	60.55	**61.72**	**71.69**	**61.31**	**69.62**	**50.36**	**58.15**	34.05	54.06	44.65	54.78
Ligue1	MaxPool	1.65	9.81	1.70	12.51	3.74	**29.46**	1.51	10.47	1.13	10.49	1.85	9.12
	NetVLAD	1.75	5.51	2.93	10.24	**5.68**	**26.86**	1.69	6.64	0.68	3.67	2.66	6.58
	NetVLAD++	6.00	25.28	7.43	30.52	12.75	49.72	7.49	27.48	5.73	25.92	6.7	27.33
	CALF	2.73	12.48	3.32	15.88	6.56	**33.50**	2.97	12.6	3.25	11.33	3.03	12.89
	E2E-Spot rny002_gsm	26.28	36.05	32.83	42.36	56.91	67.33	28.65	36.22	21.19	34.28	26.77	35.12
	E2E-Spot rny008_gsm	29.31	38.39	36.62	44.70	58.50	65.42	28.73	35.82	21.99	35.67	29.93	37.33
Bundesliga	MaxPool	1.26	9.04	1.57	11.99	1.76	15.73	**2.47**	**16.86**	0.97	10.84	1.46	7.98
	NetVLAD	2.26	10.18	**4.67**	14.91	4.75	15.00	3.76	**23.49**	1.9	9.17	2.79	8.55
	NetVLAD++	7.99	27.92	10.06	32.68	10.40	40.12	11.55	37.65	6.64	27.86	7.18	27.85
	CALF	3.72	14.30	5.34	18.14	4.12	17.98	**9.06**	**31.77**	3.69	16.99	3.84	14.62
	E2E-Spot rny002_gsm	35.03	43.93	36.05	44.75	38.20	44.59	42.67	52.06	27.47	42.11	28.94	37.31
	E2E-Spot rny008_gsm	37.19	46.53	42.30	50.52	41.01	49.33	44.65	53.63	26.96	43.55	32.98	41.18
SerieA	MaxPool	1.34	9.67	2.10	12.23	1.68	17.63	1.66	10.87	**2.76**	**19.05**	2.03	10.22
	NetVLAD	2.4	10.85	3.47	18.87	4.37	16.90	2.54	13.37	**4.39**	**26.1**	2.5	11.26
	NetVLAD++	10.13	32.62	13.28	39.40	17.41	47.98	**12.96**	33.88	**11.04**	**44.5**	9.53	32.04
	CALF	3.37	17.13	4.15	21.46	6.36	31.54	3.64	17.7	5.21	**31.31**	3.29	16.86
	E2E-Spot rny002_gsm	37.70	49.00	43.46	53.11	46.82	56.53	44.77	54.01	**39.31**	**61.56**	36.89	46.51
	E2E-Spot rny008_gsm	38.62	51.81	47.50	61.12	48.94	59.17	42.36	51.38	**42.68**	**64.93**	41.87	52.06
LaLiga	MaxPool	1.42	9.75	1.75	14.04	1.95	17.31	1.8	10.95	1.73	11.61	**2.88**	**16.86**
	NetVLAD	1.65	9.76	2.75	15.37	4.68	15.69	2.74	11.54	2.66	12.15	**4.37**	**23.47**
	NetVLAD++	8.51	32.15	10.64	37.95	13.67	52.88	11.26	34.89	9.47	35.13	**11.72**	**40.99**
	CALF	4.11	19.77	6.17	26.45	8.11	32.23	3.81	21.5	4.67	22.81	**8.22**	**31.09**
	E2E-Spot rny002_gsm	43.12	53.38	53.85	63.95	55.23	64.00	42.80	50.53	35.3	55.7	**50.95**	**60.09**
	E2E-Spot rny008_gsm	45.10	55.20	54.35	62.78	55.66	64.41	46.51	53.73	36.89	57.27	**52.10**	**60.68**
SoccerNet	MaxPool	2.71	17.70	2.93	22.60	4.19	30.15	2.66	21.31	2.54	20.17	2.98	19.1
	NetVLAD	3.41	29.90	4.21	36.72	6.32	45.03	4.65	30.63	4.6	32.56	4.81	30.86
	NetVLAD++	17.54	56.87	14.54	60.79	18.38	66.00	15.42	55.27	9.39	54.14	11.43	50.95
	CALF	13.71	41.53	16.82	46.23	17.74	50.02	15.49	43.22	13.19	43.03	15.44	41.54
	E2E-Spot rny002_gsm	65.23	75.42	70.68	79.08	79.22	84.67	67.51	73.52	49.09	70.65	61.55	68.59
	E2E-Spot rny008_gsm	60.85	74.83	71.06	79.56	79.66	84.31	69.44	75.15	51.88	73.41	62.92	70.38

Table 2. **Intra league analyses within the SoccerNet action spotting dataset.** We analyze how state-of-the-art action spotting models trained on specific leagues of SoccerNet transfer on other leagues. We show the Tight-Average-mAP and Loose-Average-mAP metrics for comparison. Best performing training of a given test domain is shown in **bold**. The gray backgrounds depicts the same train/test domains.

Work	Model	SoccerNet Data		Super League		BGL League		CINFO	
		tight-AmAP	loose-AmAP	tight-AmAP	loose-AmAP	tight-AmAP	loose-AmAP	tight-AmAP	loose-AmAP
Pooling	MaxPooling	–	18.6	0.75	12.86	0.71	8.10	0.86	11.58
	NetVLAD	4.20	31.37	0.68	11.01	1.68	7.31	0.38	2.20
CALF	CALF	14.10	41.61	5.08	20.98	0.74	5.27	0.37	4.70
NetVLAD++	NetVLAD++	11.51	53.40	4.36	33.25	2.21	12.36	1.10	13.79
E2E-Spot	rny002_gsm	61.19	73.25	63.10	67.31	21.89	30.99	18.81	28.84
	rny008_gsm	61.82	74.05	70.02	73.92	24.09	32.66	19.56	30.32
	rny008_gsm Challenge	61.82	74.05	69.97	73.25	34.23	43.05	21.40	32.98
Comedian	ViViT Tiny	70.70	76.10	69.58	73.32	28.12	41.27	23.93	**40.13**
	ViSwin Tiny	71.60	76.60	**72.39**	**75.90**	**41.31**	**49.64**	**25.93**	36.14
	ViViT-T Ensemble	72.00	77.10	69.57	65.67	25.09	37.32	18.89	30.86
	ViSwin-T Ensemble	**73.10**	**77.60**	70.89	73.76	36.42	42.71	20.45	26.01

Table 3. **Cross-domain Transfer analyses.** We analyze how state-of-the-art action spotting models trained on SoccerNet transfer on the original and other domains. We show the Tight-Average-mAP and Loose-Average-mAP metrics for comparison. Best performing model of a given test domain is shown in **bold**.

shedding light on the efficacy of model transfer capability across diverse soccer leagues.

EPL to other leagues: Algorithms trained on the English Premier League (EPL) demonstrate competitive performance when transferred to other leagues, with NetVLAD++ and Spot rny008_gsm exhibiting notable performance across multiple leagues. **UEFA to other leagues:** Models trained on UEFA data also show promising transfer capability, especially with Spot rny008_gsm achieving impressive performance across different leagues. **Ligue1 to other leagues:** Transfer from Ligue1 to other leagues yields mixed results, with Spot rny002_gsm and Spot rny008_gsm showing relatively consistent performance. **Bundesliga to other leagues:** Algorithms trained on Bundesliga data generally exhibit good transfer capability, with Spot rny008_gsm consistently performing well across different leagues. **SerieA to other leagues:** Transfer from SerieA to other leagues yields mixed results, with Spot rny008_gsm demonstrating competitive performance across various leagues. **LaLiga to other leagues:** LaLiga-trained models demonstrate strong transfer capability, with Spot rny008_gsm consistently achieving high performance across different leagues.

Overall, the results highlight the potential of models trained on one league to perform well when transferred to other leagues within the SoccerNet dataset. Notably, end-to-end methods like Spot rny008_gsm exhibit more robustness and effectiveness in cross-league transfer scenarios.

4.3. Cross leagues transfer outside SoccerNet

In this set of experiments, we present the results of domain transfer conducted across different sports leagues, extending beyond the SoccerNet dataset. The evaluation encompasses two prominent soccer leagues, namely the Super League and the BGL League, along with data from the CINFO dataset. Table 3 showcases the performance metrics of various models across these diverse datasets. Results reveal notable variations in model performance across different leagues. For instance, in the Super League dataset, the "Spot" models, particularly rny002_gsm and rny008_gsm, exhibit high tight-AmAP scores of 63.10% and 70.02% respectively, indicating their effectiveness in this domain. Similarly, the "Comedian" models, such as ViSwin-T Ensemble, demonstrate competitive performance across all leagues, with tight-AmAP scores ranging from 70.89% to 73.10%. Interestingly, the feature-based models demonstrate strong performance on SoccerNet, while showing relatively lower performance in the Super League, BGL League, and CINFO datasets. Overall, these results underscore the importance of evaluating models across diverse datasets and sports leagues to assess their generalizability and robustness beyond the training domain.

5. Discussion

In this section, we delve into the implications of the results presented in Table 2 and Table 3, analyzing the strengths and limitations of the examined models, identifying factors influencing their performance, and proposing avenues for future research and model refinement.

Transfer capability across same domain Our investigation into the transfer capabilities across different domains within the SoccerNet dataset reveals compelling insights into the generalization ability of deep learning models in sports action analysis. Upon examining how models perform across various subsets of the dataset that we shown at Table 2, it becomes evident that models trained on a specific subset demonstrate notable generalization abilities when applied to other subsets. Notably, all models exhibit superior performance on their own subset's test set compared to others, indicating a degree of domain specificity. However, it is worth noting that the test sets for UEFA and

Ligue 1 subsets stand out due to their smaller number of classes—16 and 15, respectively—compared to the original 17. Despite this discrepancy, models trained on these subsets still showcase admirable generalization capabilities. Furthermore, our analysis underscores an interesting observation: none of the models trained on specific subsets surpass the performance of the original models, which were trained on the entire SoccerNet dataset, by a considerable margin. This performance gap ranges from less than 0.5% for older models like MaxPool and NetVLAD to over 10% for newer models like Spot. This suggests that while models trained on specific subsets can generalize well within their domain, they may struggle to match the overall performance of models trained on the entire dataset. Finally, we calculate the average discrepancy between the model results trained using data from all leagues and each individual league. Our analysis revealed that models trained with data from the UEFA league exhibited the smallest discrepancies, indicating that these models produce results closest to those derived from using the entire dataset. This observation may be attributed to the diverse stadiums across Europe where UEFA matches are held, each equipped with distinct broadcasting technologies. Consequently, this diversity enriches the dataset, enhancing the models' ability to generalize effectively.

Transfer capability across different domains The analysis presented in Table 3 highlights the models' ability to transfer knowledge from the original SoccerNet dataset to new domains. In the pooling-based methods, we observe intriguing patterns: despite MaxPooling exhibiting comparatively lower performance with a Loose Average-mAP of 18.6% compared to NetVLAD's 31.37%, it showcases superior generalization capabilities. MaxPooling experiences minimal performance degradation and even outperforms NetVLAD in the Swiss Super League subset. In the case of CALF and NetVLAD++, both models demonstrate similar performance, but CALF experiences a more pronounced drop in Tight-Average-mAP. Notably, the end-to-end methods exhibit excellent generalization in the Swiss Super League subset but experience substantial performance reductions in datasets with broadcast videos generated by cameras automatically piloted. This phenomenon suggests that while these models excel in specific domains, they may struggle to adapt to novel datasets, indicating potential challenges in generalization across diverse data distributions.

Feature-Based *vs*. End-to-End Methods Across all analyzed cases, as evidenced in both Table 2 and Table 3, a substantial gap exists between feature-based methods (MaxPooling, NetVLAD, NetVLAD++, and CALF) and end-to-end approaches (Spot and COMEDIAN). Feature-based methods exhibit a significant performance drop, particularly in the Tight-Average-mAP metric, where they only achieve a maximum of 5.08% in the Swiss Super League subset and less than 1.10% for the CINFO dataset. In contrast, end-to-end methods consistently maintain a Tight-Average-mAP above 18.89% across all cases, with the CINFO dataset posing the greatest challenge for both feature-based and end-to-end methods. This disparity underscores the inherent advantages of end-to-end approaches in preserving performance across diverse datasets, potentially attributed to their ability to learn more complex representations directly from raw data. However, it also highlights the need for further research to bridge the performance gap between feature-based and end-to-end methods, especially in challenging domains like the CINFO dataset.

6. Conclusion

In this study, we conducted an extensive analysis of state-of-the-art action spotting models applied to football videos, focusing on their performance within the SoccerNet dataset and on different domains. Our investigation revealed significant potential within SoccerNet for extracting valuable insights applicable across diverse domains. However, we also highlighted a notable performance gap between SoccerNet and other league levels and broadcast quality. To address this gap and achieve comparable results across domains, we suggest further exploration of techniques such as domain adaptation, continual learning, or transfer learning. These methodologies offer promising avenues for closing the performance disparity between datasets and expanding the application of AI in sports analysis to a broader spectrum of domains. This analysis paper therefore seeks to spotlight this pressing issue and catalyze future research endeavor.

Acknowledgment

This work was partly supported by Grants PID2019-109238GB-C2 and PID2022-136435NB-I00, funded by MCIN/AEI/ 10.13039/501100011033, PID2022 also funded by "ERDF A way of making Europe", EU, and the Xunta de Galicia (ED431C 2022/44, ED431C 2021/30 and ED431F 2021/11). This works was also 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. A. Cioppa is funded by the F.R.S.-FNRS.

References

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

[2] Humam Alwassel, Fabian Caba Heilbron, and Bernard Ghanem. Action search: Spotting actions in videos and its application to temporal action localization. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, pages 253–269, Cham, 2018. Springer International Publishing. 2

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

[4] S. Buch, Victor Escorcia, Bernard Ghanem, Li Fei-Fei, and Juan Carlos Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. In *British Machine Vision Conference*, 2017. 2

[5] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. Sst: Single-stream temporal action proposals. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6373–6382, 2017. 2

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

[7] Bruno Cabado, Bertha Guijarro-Berdiñas, and Emilio J. Padrón. Real-time classification of handball game situations. In *2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 686–691, 2022. 2

[8] Alejandro Cartas, Coloma Ballester, and Gloria Haro. A graph-based method for soccer action spotting using unsupervised player classification. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 93–102, Lisbon, Port., Oct. 2022. ACM. 3

[9] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches. In *Proc. SSST-8, Eighth Work. Syntax. Semant. Struct. Stat. Transl.*, pages 103–111, Doha, Qatar, 2014. Association for Computational Linguistics. 3

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

[11] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas B. Moeslund. A context-aware loss function for action spotting in soccer videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 13123–13133, Seattle, WA, USA, Jun. 2020. Inst. Electr. Electron. Eng. (IEEE). 3, 5

[12] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Floriane Magera, Olivier Barnich, Bernard Ghanem, and Marc 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*, pages 4532–4541, Nashville, TN, USA, Jun. 2021. 3

[13] Anthony Cioppa, Adrien Deliege, Maxime Istasse, Christophe De Vleeschouwer, and Marc Van Droogenbroeck. ARTHuS: Adaptive real-time human segmentation in sports through online distillation. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 2505–2514, Long Beach, CA, USA, Jun. 2019. Inst. Electr. Electron. Eng. (IEEE). 2

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

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

[16] Anthony Cioppa, Silvio Giancola, Vladimir Somers, Floriane Magera, Xin Zhou, Hassan Mkhallati, Adrien Deliège, Jan Held, Carlos Hinojosa, Amir M. Mansourian, Pierre Miralles, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdullah Kamal, Adrien Maglo, Albert Clapés, Amr Abdelaziz, Artur Xarles, Astrid Orcesi, Atom Scott, Bin Liu, Byoungkwon Lim, Chen Chen, Fabian Deuser, Feng Yan, Fufu Yu, Gal Shitrit, Guanshuo Wang, Gyusik Choi, Hankyul Kim, Hao Guo, Hasby Fahrudin, Hidenari Koguchi, Håkan Ardö, Ibrahim Salah, Ido Yerushalmy, Iftikar Muhammad, Ikuma Uchida, Ishay Be'ery, Jaonary Rabarisoa, Jeongae Lee, Jiajun Fu, Jianqin Yin, Jinghang Xu, Jongho Nang, Julien Denize, Junjie Li, Junpei Zhang, Juntae Kim, Kamil Synowiec, Kenji Kobayashi, Kexin Zhang, Konrad Habel, Kota Nakajima, Licheng Jiao, Lin Ma, Lizhi Wang, Luping Wang, Menglong Li, Mengying Zhou, Mohamed Nasr, Mohamed Abdelwahed, Mykola Liashuha, Nikolay Falaleev, Norbert Oswald, Qiong Jia, Quoc-Cuong Pham, Ran Song, Romain Hérault, Rui Peng, Ruilong Chen, Ruixuan Liu, Ruslan Baikulov, Ryuto Fukushima, Sergio Escalera, Seungcheon Lee, Shimin Chen, Shouhong Ding, Taiga Someya, Thomas B. Moeslund, Tianjiao Li, Wei Shen, Wei Zhang, Wei Li, Wei Dai, Weixin Luo, Wending Zhao, Wenjie Zhang, Xinquan Yang, Yanbiao Ma, Yeeun Joo, Yingsen Zeng, Yiyang Gan, Yongqiang Zhu, Yujie Zhong, Zheng Ruan, Zhiheng Li, Zhijian Huangi, and Ziyu Meng. SoccerNet 2023 challenges results. *arXiv*, abs/2309.06006, 2023. 3

[17] Abdulrahman Darwish and Tallal El-Shabrway. STE: Spatio-temporal encoder for action spotting in soccer videos. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 87–92, Lisbon, Port., Oct. 2022. ACM. 3

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

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

[20] Julien Denize, Mykola Liashuha, Jaonary Rabarisoa, Astrid Orcesi, and Romain Hérault. Comedian: Self-supervised learning and knowledge distillation for action spotting using transformers. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*, pages 530–540, January 2024. 3, 5

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

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

[23] Jiyang Gao, Zhenheng Yang, Chen Sun, Kan Chen, and Ram Nevatia. Turn tap: Temporal unit regression network for temporal action proposals. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 3648–3656, 2017. 2

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

[25] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. SoccerNet: A scalable dataset for action spotting in soccer videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 1792–179210, Salt Lake City, UT, USA, Jun. 2018. Inst. Electr. Electron. Eng. (IEEE). 1, 2, 3, 5

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

[27] Silvio Giancola and Bernard Ghanem. Temporally-aware feature pooling for action spotting in soccer broadcasts. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4490–4499, Nashville, TN, USA, Jun. 2021. 3, 5

[28] Jan Held, Anthony Cioppa, Silvio Giancola, Abdullah Hamdi, Bernard Ghanem, and Marc 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)*, pages 5086–5097, Vancouver, Can., Jun. 2023. Inst. Electr. Electron. Eng. (IEEE). 2

[29] Jan Held, Hani Itani, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. X-vars: Introducing explainability in football refereeing with multimodal large language models. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, Seattle, WA, USA, Jun. 2024. 2

[30] James Hong, Haotian Zhang, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Spotting temporally precise, fine-grained events in video. *arXiv*, abs/2207.10213, 2022. 3, 5

[31] Haroon Idrees, Amir R. Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. The thumos challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding*, 155:1–23, Feb. 2017. 2

[32] IFAB. Laws of the game. Technical report, The International Football Association Board, Zurich, Switzerland, 2022. 3

[33] Maxime Istasse, Vladimir Somers, Pratheeban Elancheliyan, Jaydeep De, and Davide Zambrano. DeepSportradar-v2: A multi-sport computer vision dataset for sport understandings. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 23–29, Ottawa, Ontario, Can., Oct. 2023. ACM. 2

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

[35] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale

video classification with convolutional neural networks. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1725–1732, Columbus, OH, USA, Jun. 2014. Inst. Electr. Electron. Eng. (IEEE). 2

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

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

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

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

[40] Arnaud Leduc, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc 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. 2

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

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

[43] Tianwei Lin, Xiao Liu, Xin Li, Errui Ding, and Shilei Wen. Bmn: Boundary-matching network for temporal action proposal generation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3888–3897, 2019. 2

[44] Paul Liu and Jui-Hsien Wang. MonoTrack: Shuttle trajectory reconstruction from monocular badminton video. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3512–3521, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

[45] Yuan Liu, Lin Ma, Yifeng Zhang, Wei Liu, and Shih-Fu Chang. Multi-granularity generator for temporal action proposal. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3599–3608, 2019. 2

[46] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. Gaussian temporal awareness networks for action localization. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 344–353, 2019. 2

[47] Floriane Magera, Thomas Hoyoux, Olivier Barnich, and Marc 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. 2

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

[49] Amir M. Mansourian, Vladimir Somers, Christophe De Vleeschouwer, and Shohreh Kasaei. Multi-task learning for joint re-identification, team affiliation, and role classification for sports visual tracking. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, page 103–112, Ottawa, Ontario, Can., Oct. 2023. ACM. 2

[50] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. SoccerNet-caption: Dense video captioning for soccer broadcasts commentaries. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 5074–5085, Vancouver, Can., Jun. 2023. Inst. Electr. Electron. Eng. (IEEE). 2

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

[52] Banoth Thulasya Naik, Mohammad Farukh Hashmi, Neeraj Dhanraj Bokde, and Zaher Mundher Yaseen. A comprehensive review of computer vision in sports: Open issues, future trends and research directions. *Appl. Sci.*, 12(9):1–49, Apr. 2022. 2

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

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

[55] Olav Rongved, Markus Stige, Steven Hicks, Vajira Thambawita, Cise Midoglu, Evi Zouganeli, Dag Johansen, Michael Riegler, and Pål 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. 3

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

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

[58] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, volume 27, pages 568–576, Dec. 2014. 2

[59] João V. B. Soares and Avijit Shah. Action spotting using dense detection anchors revisited: Submission to the SoccerNet challenge 2022. *arXiv*, abs/2206.07846, 2022. 3

[60] Joao V. B. Soares, Avijit Shah, and Topojoy Biswas. Temporally precise action spotting in soccer videos using dense detection anchors. In *ICIP*, pages 2796–2800, Bordeaux, France, Oct. 2022. Inst. Electr. Electron. Eng. (IEEE). 3

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

[62] Vladimir Somers, Victor Joos, Anthony Cioppa, Silvio Giancola, Seyed Abolfazl Ghasemzadeh, Floriane Magera, Baptiste Standaert, Amir Mohammad Mansourian, Xin Zhou, Shohreh Kasaei, Bernard Ghanem, Alexandre Alahi, Marc Van Droogenbroeck, and Christophe 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. 2

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

[64] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift networks for video action recognition. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 1099–1108, Seattle, WA, USA, Jun. 2020. Inst. Electr. Electron. Eng. (IEEE). 3

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

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

[67] Matteo Tomei, Lorenzo Baraldi, Simone Calderara, Simone Bronzin, and Rita Cucchiara. RMS-net: Regression and masking for soccer event spotting. In *IEEE Int. Conf. Pattern Recognit. (ICPR)*, pages 7699–7706, Milan, Italy, Jan. 2021. Inst. Electr. Electron. Eng. (IEEE). 3

[68] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 2

[69] Gabriel Van Zandycke and Christophe De Vleeschouwer. 3D ball localization from a single calibrated image. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3471–3479, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

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

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

[72] Bastien Vanderplaetse and Stephane Dupont. Improved soccer action spotting using both audio and video streams. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 3921–3931, Seattle, WA, USA, Jun. 2020. Inst. Electr. Electron. Eng. (IEEE). 3

[73] Kanav Vats, Mehrnaz Fani, Pascale Walters, David A. Clausi, and John Zelek. Event detection in coarsely annotated sports videos via parallel multi receptive field 1D convolutions. *arXiv*, abs/2004.06172, 2020. 3

[74] Kanav Vats, William McNally, Pascale Walters, David A. Clausi, and John S. Zelek. Ice hockey player identification via transformers and weakly supervised learning. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3450–3459, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2

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

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

[77] Shengbo Wang, Zhenjiang Miao, Wanru Xu, Cong Ma, and Miaomiao Li. Boundary sensitive and category sensitive network for temporal action proposal generation. In *2019 Chinese Automation Congress (CAC)*, pages 5194–5199, 2019. 2

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

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

[80] Junqing Yu, Aiping Lei, Zikai Song, Tingting Wang, Hengyou Cai, and Na Feng. Comprehensive dataset of broadcast soccer videos. In *IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, pages 418–423, Miami, FL, USA, Apr. 2018. Inst. Electr. Electron. Eng. (IEEE). 2

[81] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. Temporal action detection with structured segment networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2933–2942, 2017. 2

[82] He Zhu, Junwei Liang, Chengzhi Lin, Jun Zhang, and Jianming Hu. A transformer-based system for action spotting

in soccer videos. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 103–109, Lisbon, Port., Oct. 2022. ACM. 3

[83] Kevin Zhu, Alexander Wong, and John McPhee. FenceNet: Fine-grained footwork recognition in fencing. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3588–3597, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2