Action Anticipation from SoccerNet Football Video Broadcasts

Mohamad Dalal[1,*]   Artur Xarles[2,3,*]   Anthony Cioppa[4]   Silvio Giancola[5]
Marc Van Droogenbroeck[4]   Bernard Ghanem[5]   Albert Clapés[2,3]   Sergio Escalera[1,2,3]
Thomas B. Moeslund[1]

[1]Aalborg University   [2]Universitat de Barcelona   [3]Computer Vision Center   [4]University of Liège   [5]KAUST

Abstract

Artificial intelligence has revolutionized the way we analyze sports videos, whether to understand the actions of games in long untrimmed videos or to anticipate the player's motion in future frames. Despite these efforts, little attention has been given to anticipating game actions before they occur. In this work, we introduce the task of action anticipation for football broadcast videos, which consists in predicting future actions in unobserved future frames, within a five- or ten-second anticipation window. To benchmark this task, we release a new dataset, namely the Soccer-Net Ball Action Anticipation dataset, based on SoccerNet Ball Action Spotting. Additionally, we propose a Football Action ANticipation TRansformer (FAANTRA), a baseline method that adapts FUTR, a state-of-the-art action anticipation model, to predict ball-related actions. To evaluate action anticipation, we introduce new metrics, including mAP@δ, which evaluates the temporal precision of predicted future actions, as well as mAP@∞, which evaluates their occurrence within the anticipation window. We also conduct extensive ablation studies to examine the impact of various task settings, input configurations, and model architectures. Experimental results highlight both the feasibility and challenges of action anticipation in football videos, providing valuable insights into the design of predictive models for sports analytics. By forecasting actions before they unfold, our work will enable applications in automated broadcasting, tactical analysis, and player decision-making. Our dataset and code are publicly available at https://github.com/MohamadDalal/FAANTRA.

1. Introduction

Artificial intelligence and computer vision have significantly advanced sports analytics, improving player tracking, tactical evaluation, and event detection. These advancements support applications ranging from real-time match
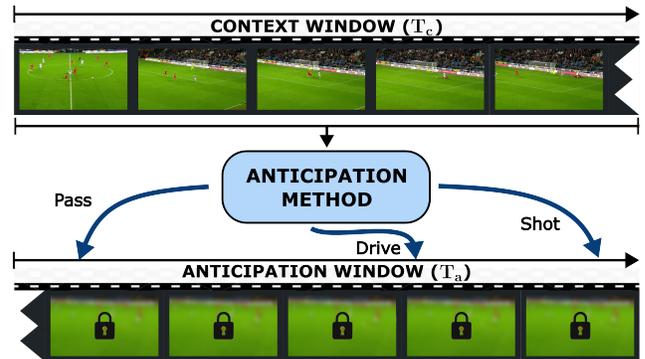
*Equal contribution.

Figure 1. **Overview of our new action anticipation task for sports.** Action anticipation aims to predict and temporally localize future actions in an *anticipation window* of T_a seconds using information from a preceding observed *context window* of T_c seconds. Unlike action spotting, where models can access the entire video sequence to detect actions, action anticipation requires predicting future events without access to future frames.

analysis to automated content generation. Much progress has been made in *action recognition* [53], which classifies actions, and *action spotting* [19], which localizes them in time. Yet, *action anticipation*, which predicts actions before they occur, remains largely unexplored. Anticipation is a key aspect in football, and integrating this predictive capability into AI can enhance game understanding and enable applications such as: (1) automated broadcasting, where cameras proactively adjust based on anticipated actions, (2) team tactical analysis, where coaches gain insights into possible future plays, and (3) player decision-support systems, which help athletes anticipate opponents' moves and optimize their positioning.

Most prior work on future game dynamics prediction has focused on *trajectory forecasting*, for player or ball movement [4, 22, 51], rather than anticipating actions such as passes or shots. Unlike action spotting, which detects actions using full video sequences, action anticipation requires reasoning about upcoming actions using only past observations. This paradigm shift introduces additional

challenges, requiring models to recognize patterns in past actions and infer likely outcomes.

This work introduces the first structured benchmark for action anticipation in football by presenting the SoccerNet Ball Action Anticipation dataset, an adaptation of previous versions of the SoccerNet dataset [7, 10, 17]. As an initial solution to solve the task, we propose the Football Action ANticipation TRAnsformer (FAANTRA) model, a baseline that adapts FUTR [23], a state-of-the-art action anticipation model, to predict ball-related actions within five- and ten-second future windows, using only past frames. In line with current trends in action spotting [28, 55], we employ an efficient video feature extractor that we train in an end-to-end approach. To evaluate performance, we propose new anticipation metrics, including mAP@δ for temporal evaluation and mAP@∞ for action occurrence within the anticipation window. Additionally, we conduct extensive ablation studies to analyze the impact of various task settings, input configurations, and model architectures. These studies highlight the challenges of the task while also underscoring the importance of key modeling choices.

Contributions. We summarize our contributions as follows: **(i)** We introduce the first structured benchmark for action anticipation in sports, formalizing the task with a novel SoccerNet Ball Action Anticipation dataset and pertinent metrics to evaluate that task. **(ii)** We propose FAANTRA, the first baseline model specifically designed for action anticipation in football broadcasts. **(iii)** We conduct extensive ablation studies, providing insights into the challenges of the anticipation task as well as effective model design.

2. Related work

Action spotting. Given a long, untrimmed video stream, the task of action spotting consists in identifying and precisely localizing actions of interest in time [17, 19, 41]. Automatically extracting actions is key for many sports applications, including generating game statistics [10], supporting video analysts in coaching or referees during games [26, 27], or customizing highlights based on the viewer's preferences [48, 49]. The SoccerNet challenges [6, 7, 18] have led to the development of numerous methods, which can be categorized into two training paradigms: *feature-based approaches* [5, 16, 43, 47, 64], which rely on pre-extracted features and train only the spotting head, and *end-to-end approaches* [3, 11, 28, 54], which simultaneously train both the backbone and the spotting head. Yet, all these methods require access to the complete video stream containing the actions to spot them, including past, present, and future contextual frames. In this work, we propose to tackle the task of action anticipation, in which the methods only have access to past context to predict future actions. Particularly, we follow current trends in action spotting by proposing an end-to-end approach to anticipation, leveraging an effi-

cient feature extractor from T-DEED [55], the state-of-the-art (SOTA) method for action spotting.

Action anticipation. This task involves predicting future actions based solely on the observation of past context. It can be divided into two main types: short-term and long-term action anticipation. In *short-term action anticipation* [14, 20, 56, 59, 62], models predict actions within a small fixed future window, typically ranging from 1 to 5 seconds, which defines the action class. This is usually done by dividing the window into intervals and predicting the action occurring at the end of each one. In contrast, *long-term action anticipation* [12, 23, 37, 57, 58, 63] aims to predict the sequence of future actions along with their durations, and may look up to several minutes ahead. Our proposed action anticipation task in football aligns with the short-term setting, anticipating actions within fixed five- or ten-second windows. However, rather than localizing activities at the end of predefined intervals, the objective is to localize the exact timestep of each anticipated action.

Most action anticipation methods rely on pre-extracted features to address the task [2, 15, 35, 36, 62], while only a few tackle it in an end-to-end manner [21, 32, 33, 60]. To process information within the observed time window, these methods typically employ architectures designed to capture the sequential nature of videos, such as Recurrent Neural Networks (RNNs) [12, 14, 15, 33, 46]. More recently, Transformers have become the preferred choice, as seen in methods like FUTR [23] and Anticipatr [37], both of which leverage a transformer encoder for this purpose. When generating anticipations, some methods [1, 2, 12, 39] adopt an auto-regressive approach, where predictions are made sequentially. Non auto-regressive approaches, on the other hand, typically fall into one of two main categories: anchor-based and query-based approaches. Anchor-based approaches [56, 58, 59] consider predefined temporal positions within the anticipation window, predicting the action occurring at each position. Most short-term anticipation methods fall into this category, as they focus on predicting actions within fixed temporal intervals. In contrast, query-based approaches [23, 37, 63] use a set of learnable representations that are not tied to specific temporal positions, instead predicting both the action and its precise localization. FAANTRA aligns with current trends by leveraging transformers while adopting a query-based approach, which our ablation studies show to be beneficial.

Action anticipation datasets are typically adapted from action localization or segmentation datasets, often focusing on daily activities, such as cooking [8, 30, 31, 44], or sourced from movie and TV-series datasets [9, 25]. However, no dataset currently addresses action anticipation in sports. Therefore, our proposed SoccerNet Ball Action Anticipation dataset is the first *video-based action* anticipation dataset focusing on sports, specifically football. Moreover,

the fast-paced and uncertain nature of football adds an extra challenge compared to other datasets, where the sequence of actions tends to follow a more scripted pattern.

Anticipation in sports. While research on anticipation in sports exists, none of the current studies approach it from a video-based action anticipation perspective. Instead, some methods [13, 22] rely on ball and player tracking data along with metadata and are constrained to anticipating only the next action class. In contrast, other works [4, 51, 52] focus on predicting the trajectories of the ball and players, which differs from our task of anticipating actions.

3. Action anticipation in sports

Task definition. Given a set of N observed, trimmed videos $\mathcal{V} = \{v^1, v^2, \ldots, v^n, \ldots, v^N\}$ with a supposed fixed length of T_c seconds, the task of **_action anticipation_** aims at predicting a set of future actions of interest $\mathcal{A}^n = \{a_1^n, a_2^n, \ldots, a_k^n, \ldots a_{|\mathcal{A}^n|}^n\}$ for each video v^n, where $|\mathcal{A}^n|$ is the total number of future actions of interest for video n, which may vary across different videos. More precisely, the goal is to only use the observed video v^n, also referred to as the *context window*, to identify and localize all actions that will occur within its immediate unobserved future time window $w^n \in \mathcal{W}$, also called the *anticipation window*, with a fixed length of T_a seconds. All actions a_k^n are defined by two features: (1) a class $c \in \{1, \ldots, C\}$, C being the total number of action classes of interest, and (2) a timestamp $t \in [0, T_a]$, indicating the exact moment the action will happen in the anticipation window.

Evaluation metrics. To benchmark our action anticipation task, we adapt the mean Average Precision (mAP) metric, proposed by Giancola et al. [17] for Action Spotting (AS), to evaluate performance within the non-observed anticipation window. Specifically, the mAP for a given temporal tolerance δ (*i.e.*, a prediction is considered correct if it falls within a window of $\delta/2$ seconds before or after the ground truth), denoted as mAP@δ, is computed by averaging the Average Precision (AP) values across different action classes [19]. AP summarizes precision-recall points into a single value by estimating the area under the precision-recall curve. To evaluate different aspects of model performance, we consider six variations of the mAP@δ metric with different δ tolerances: $\{1, 2, 3, 4, 5, \infty\}$. Smaller tolerances emphasize on precise localization of anticipated actions, while larger tolerances are more loose on minor localization errors. When $\delta = \infty$, the metric completely disregards localization and focuses only on the correct identification and counting of action classes. These variations provide a balanced evaluation of both localization accuracy and action identification. Finally, we report the average performance across all six variations.

Dataset	Year	Domain	Spontaneity	Hours	Classes
50Salads[44]	2013	Cooking	Medium	4.5	17
Breakfast[30]	2014	Cooking	Medium	77	48
THUMOS14[29]	2014	Web	Medium	20	20
Charades[42]	2016	ADL	Low	82.3	157
TVSeries[9]	2016	Movie	Medium	16	30
EGTEA Gaze+[31]	2018	Cooking	Medium	28	106
EpicKitchens-100[8]	2020	Cooking	High	100	3807
Ego4D[24]	2022	ADL	High	243	4756
Assembly101[40]	2022	ADL	Medium	513	1380
SN-BAA (ours)	2025	Football	Very High	11.4	10

Table 1. Comparison of video-based action anticipation datasets, including SN-BAA, with details on the year, domain, spontaneity level, total hours, and number of action classes, following Zhong et al. [61]. ADL stands for Activities of Daily Living.

4. SoccerNet Ball Action Anticipation dataset

We introduce the SoccerNet Ball Action Anticipation dataset (SN-BAA), by adapting the SoccerNet Ball Action Spotting dataset (SN-BAS) for our novel action anticipation task. SN-BAS comprises untrimmed videos from nine professional football matches, annotated with the precise localization of $C = 12$ different on-ball action types. The dataset is divided into four splits: four games for training, one for validation, two for testing, and two with hidden ground truth for challenge evaluation.

In adapting the dataset for the anticipation task, we preserve the original splits from SN-BAS. For the training and validation splits, we retain the original untrimmed videos, allowing different methods to process them freely based on their characteristics. To ensure standardized evaluation on the test and challenge splits, we clip the untrimmed videos into 30 seconds video clips, using a sliding window with a stride of T_a seconds. This enables methods to work with context windows T_c ranging from 0 to 30 seconds. Using a stride of T_a seconds ensures that all actions are evaluated. Regarding the action classes considered, we exclude goals and free-kicks from the SN-BAA dataset, reducing the set to $C = 10$ classes. As detailed in the dataset analysis in the supplementary material, these classes are rare, with only six and two observations, respectively, in the test split. Such low frequency leads to an unstable metric, where a single correct or incorrect anticipation can cause large fluctuations in performance. Therefore, to ensure a more robust evaluation, we removed these classes from SN-BAA.

Comparing the SN-BAA dataset with other common video-based action anticipation datasets in Tab. 1, we find that our dataset stands out in terms of spontaneity (*i.e.*, the unpredictability and suddenness of actions). The fast-paced and adversarial nature of football, where players attempt to hide their intentions from opponents, make the actions in our dataset less predictable and more challenging than in others. Additionally, SN-BAA is the first action anticipation dataset focused on sports, specifically in football. However,
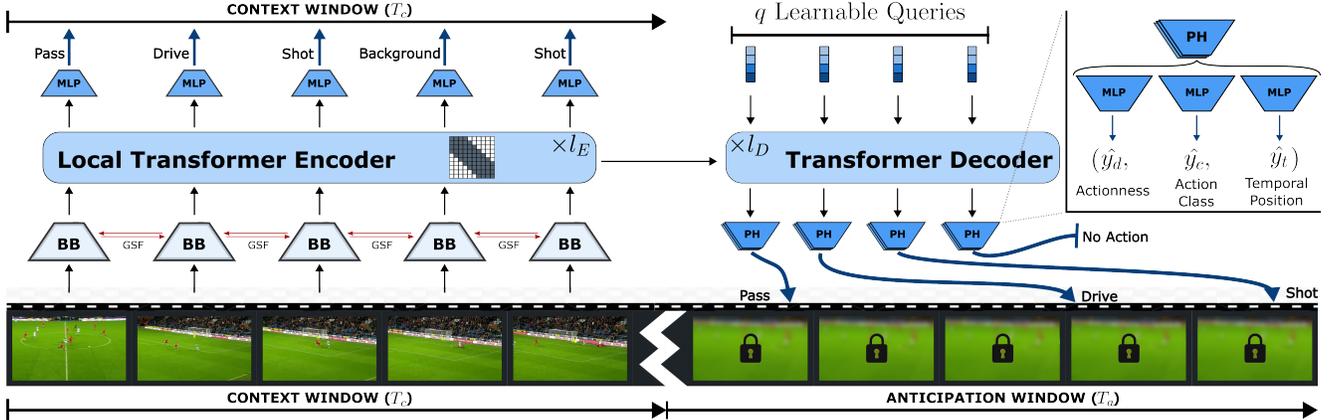
Figure 2. **FAANTRA Architecture Overview.** FAANTRA processes context video frames by extracting per-frame representations through a backbone (BB). These features are fed into a transformer encoder to capture temporal dependencies. A set of learnable queries, representing action predictions, are initialized in the transformer decoder and refined through multiple layers, leveraging information from the encoder. Each refined query is then processed by a prediction head (PH) to output three components representing the anticipated actions: action detection (*i.e.*, actionness), action class, and temporal position.

it has a moderate number of hours and classes compared to other datasets. Further analysis of the dataset statistics in the supplementary material reveals a total of 12433 actions across the entire dataset (*i.e.*, one action every 3.30 seconds). The action classes follow a long-tail distribution, with passes and drives being the most frequent, while actions such as shots or successful tackles occur far less often. A similar pattern is observed when examining the distribution of actions within an anticipation window of $T_a = 5$ seconds, where the dataset contains an average of 1.5 actions per window, with a maximum of 8 actions.

5. Football action anticipation transformer

Method overview. Our new method, namely Football Action ANticipation TRAnsformer (FAANTRA), builds on top of FUTR [23], a SOTA method for action anticipation. Specifically, our approach consists of three main components: a feature extractor, a transformer encoder-decoder, and a prediction head. Following current trends in Action Spotting (AS) [28], we employ end-to-end training with a feature extractor that directly processes frames from fixed-length videos spanning T_c seconds, corresponding to the context window. The number of frames within this window is denoted as F_c. In line with SOTA feature extractors for AS [55], we use a 2D efficient backbone, RegNetY [38]. This backbone extracts per-frame representations, which are further refined by a transformer encoder-decoder. The encoder, with l_E layers, processes the extracted features, while the decoder, with l_D layers and learnable queries to represent the anticipated actions, captures relevant information from the encoder to anticipate future actions. We further introduce locality biases in the encoder through local self-attention, as shown beneficial in the presented abla-

tions. Finally, each refined query is processed by the prediction head, which comprises three components: (i) a binary classifier to detect action presence, (ii) a classification module to identify action class, and (iii) a timestamp predictor to estimate the precise temporal position of the anticipated action. Further details of the method are presented in the following sections.

5.1. Feature extractor

Following SOTA in Action Spotting, we use RegNetY [38, 55] as our backbone, known for its efficiency. Additionally, we also incorporate Gate-Shift-Fuse (GSF) modules [45] in the latter half of the backbone, allowing the extracted features to integrate local spatial modelling before capturing longer temporal dependencies within the transformer encoder. A linear layer is added after the feature extractor to project the representations into d dimensions. As a result, the feature extractor processes the input frame sequence corresponding to the observed context window of F_c frames, $v^n \in \mathbb{R}^{F_c \times H \times W \times 3}$, where $H \times W$ denotes the spatial resolution and produces a sequence of F_c, per-frame, d-dimensional representations of dimension $F_c \times d$.

5.2. Transformer encoder-decoder

The transformer encoder-decoder module processes the features, which primarily contain spatial information along with local temporal information. The objective of the transformer is to capture longer-term temporal dependencies and extract the relevant cues to anticipate actions within the anticipation window. The encoder refines these features, while the decoder transfers the refined information to a set of learnable queries that represent the anticipated actions.

Transformer encoder. It consists of l_E transformer en-

coder layers, adapted from the original vanilla transformer encoder layers [50] to use local self-attention instead of global self-attention, as shown beneficial in Sec. 7.2. The frame representations from the feature extractor of dimension $F_c \times d$, pass through l_E encoder layers, each comprising a multi-head local self-attention module with h heads, restricting tokens to attend within a neighborhood of k tokens. Each layer also includes a feed-forward network and layer normalization. Additionally, learnable positional encodings are added to the frame representations at each transformer encoder layer.

Transformer decoder. It consists of l_D vanilla transformer decoder layers. Each layer applies standard multi-head self-attention and cross-attention, both with h heads, followed by a two-layer feed-forward network and layer normalization. The decoder processes a set of q learnable queries representing the anticipated actions. Within each layer, these queries first undergo global self-attention before capturing information from the encoder's output through the cross-attention mechanism. The resulting refined queries of dimension $q \times d$ are then fed to the prediction head.

5.3. Prediction head

The refined queries, representing q possibly anticipated actions, are fed into the prediction head, which consists of three components: (i) an action detection component, referred to as *actionness*, (ii) an action classification component, and (iii) a timestamp regressor. Each component processes the refined queries in parallel by linearly projecting them through a linear layer. The key differences lie in the output dimensions of 1, C, and 1, respectively, and the activation functions, which are sigmoid, softmax, and the identity function, respectively. Thus, the prediction head outputs $\hat{y} = (\hat{y}_d, \hat{y}_c, \hat{y}_t)$, where $\hat{y}_d \in \mathbb{R}^{q \times 1}$, $\hat{y}_c \in \mathbb{R}^{q \times C}$, and $\hat{y}_t \in \mathbb{R}^{q \times 1}$. These represent (i) the probability of a query anticipating an arbitrary action, (ii) the confidence for each class, expressed as a probability distribution, and (iii) the absolute temporal position of the action within the anticipation window.

5.4. Multi-task learning

The model is trained using a combination of two losses: one for anticipation and one for an auxiliary action segmentation task within the context window.

Anticipation loss. The anticipation loss consists of three components: action detection $\mathcal{L}_D$, action classification $\mathcal{L}_C$, and temporal position prediction $\mathcal{L}_T$. These losses are computed for each query, which are sequentially paired with ground-truth actions in the anticipation window based on their temporal order. When no more ground-truth actions are available, remaining queries are treated as non-actions. For action detection ($\mathcal{L}_D$), we use binary cross-entropy, assigning a positive label to queries paired with ground-truth

actions and a negative label otherwise. Then, the action classification loss ($\mathcal{L}_C$) is computed using cross-entropy, with the target being the action class of the paired ground-truth action. Finally, the temporal position loss ($\mathcal{L}_T$) is the mean squared error between the predicted and actual temporal positions of the paired action, with position values scaled by the anticipation window T_a and learned in exponential space, following FUTR [23]. For unpaired queries, only the action detection loss is applied. The anticipation loss is then computed as $\mathcal{L}_A = \lambda_D \mathcal{L}_D + \lambda_C \mathcal{L}_C + \lambda_T \mathcal{L}_T$, where λ_D, λ_C, and λ_T weights each loss component.

Segmentation loss. In addition to the anticipation task, the model is trained on an auxiliary action segmentation task, commonly used in anticipation settings [2, 23, 63]. This task requires the model to segment actions within the context window, providing direct supervision over frames where actions occur. This helps the model learn the semantics of different action classes, which improves predictions within the anticipation window, as discussed in Sec. 7.2. To do so, a linear layer with softmax activation is applied to the output of the transformer encoder, generating a probability vector of size $\mathbb{R}^{F_c \times (C+1)}$, representing the presence of each action class or background at each temporal position. The auxiliary segmentation loss, $\mathcal{L}_S$, is computed using cross-entropy with label dilation [28], where supervision is applied not only at the exact temporal positions of ground-truth actions but also within a dilation radius around each action, of $r = 4$ frames.

Multi-task loss. The final loss function is then computed as $\mathcal{L} = \mathcal{L}_A + \lambda_S \mathcal{L}_S$, where λ_S controls the weight of the auxiliary segmentation loss.

5.5. Inference

During inference, for each query, we multiply the per-class probabilities ($\hat{y}_c$) with the actionness score ($\hat{y}_d$) to represent the confidence of specific action classes occurring at the predicted temporal position ($\hat{y}_t$), resulting in the final set of anticipated actions.

6. Results

In this section, we detail the implementation and training details of our method, the evaluation protocol, and present experimental results.

6.1. Implementation and training details

We train our baseline model using a context window of $T_c = 5$ seconds at 6.25 fps (i.e., $F_c = 32$ frames) with a spatial resolution of 448×796 pixels. The anticipation window is set to either $T_a = 5$ or $T_a = 10$ seconds, as discussed in Sec. 7.1. Unless stated otherwise, our model consists of $l_E = 4$ encoder layers, $l_D = 2$ decoder layers, $h = 8$ attention heads, a local encoder with an attention span of $k = 15$ neighboring embeddings, and $q = 8$

	Model	Size	Data	mAP@δ ($T_a = 5$)							mAP@δ ($T_a = 10$)						
				$\delta = 1$	2	3	4	5	∞	Avg.	$\delta = 1$	2	3	4	5	∞	Avg.
ANTICIPATION	FAANTRA	200MF	SN-BAA	8.27	14.65	20.98	23.74	26.01	28.15	20.30	6.38	11.38	14.60	16.94	18.94	27.72	15.99
	FAANTRA	400MF	SN-BAA	7.83	13.41	19.49	22.58	24.50	26.65	19.08	6.03	11.37	14.30	16.28	18.17	27.80	15.66
	FAANTRA	200MF	SN-AS + SN-BAA	9.59	17.31	23.84	27.94	30.78	32.96	23.74	5.42	10.84	16.21	18.91	21.45	32.20	17.50
	FAANTRA	400MF	SN-AS + SN-BAA	**9.74**	**17.47**	**24.11**	**28.56**	**31.13**	**33.47**	**24.08**	**7.32**	**12.62**	**18.39**	**22.08**	**25.02**	**33.95**	**19.90**
SPOTTING	T-DEED	200MF	SN-BAA	49.70	51.91	53.40	53.62	54.44	56.06	53.19	50.66	54.17	56.04	56.44	57.28	58.25	55.47
	T-DEED	400MF	SN-BAA	52.70	55.68	57.20	57.39	57.84	59.28	56.69	-	-	-	-	-	-	-
	T-DEED	200MF	SN-AS + SN-BAA	56.72	60.89	61.95	62.54	62.95	64.02	61.51	**52.03**	**54.95**	**56.05**	**56.96**	**57.82**	**58.93**	**56.12**
	T-DEED	400MF	SN-AS + SN-BAA	**59.46**	**61.90**	**64.34**	**65.04**	**65.30**	**66.99**	**63.85**	-	-	-	-	-	-	-

Table 2. FAANTRA's main results on the proposed action anticipation task, evaluated using mAP@δ metrics from Sec. 3. Results are reported for $T_a = 5$ and $T_a = 10$ seconds, with models of different feature extractor sizes (in MegaFlops) trained on either SN-BAA alone or jointly with SN-AS. Results from the SOTA Action Spotting method, T-DEED, are also included as an upper bound, under the same variations. Due to limited GPU memory, T-DEED results for $T_a = 10$ seconds with a 400MF feature extractor are not reported.

learnable queries in the decoder. As feature extractor, we use either RegNetY-200MF or RegNetY-400MF with hidden dimensions of $d = 512$. For $T_a = 10$ experiments, the number of queries is doubled to $q = 16$. All models were trained on an NVIDIA RTX 6000 GPU.

For training, we process untrimmed videos by splitting them into clips of length T_c using a sliding window with a 90% overlap between consecutive clips, which are randomly sampled. Each clip is linked to annotations for actions occurring within its context window, as well as those in the following T_a seconds (*i.e.*, the anticipation window) for supervision. Models are trained for 30 epochs with the AdamW [34] optimizer, using a batch size of 4 clips, a base learning rate of $1e - 4$, 3 linear warmup epochs, and cosine learning rate decay. The loss weights are set to $\lambda_D = 1$, $\lambda_C = 1$ $\lambda_T = 10$, and $\lambda_S = 1$. In the cross-entropy loss, class weights for both auxiliary segmentation and anticipation tasks are set inversely proportional to their frequency. Common data augmentation techniques, including random horizontal flip, Gaussian blur, and color jitter, are applied.

6.2. Evaluation protocol

Following standard practices, we train on the train set, use the validation set to prevent overfitting, and evaluate on the test set. Model performance is measured using the mAP variations detailed in Sec. 3.

6.3. Main results

In Tab. 2, we present the results of our proposed baseline, FAANTRA, using two variants of the feature extractor (*i.e.*, RegNetY-200MF or RegNetY-400MF). We also report results from training on SN-BAA alone and jointly training on SN-BAA and the original SoccerNet Action Spotting (SN-AS) dataset, which includes 500 additional games with related action classes. To incorporate SN-AS, we process untrimmed videos similarly to SN-BAA and duplicate the segmentation and anticipation prediction heads, allowing us to leverage both datasets simultaneously, as done by the SoccerNet 2024 Ball Action Spotting challenge win-

FAANTRA Performance (mAP@δ)							
Action	$\delta = 1$	2	3	4	5	∞	Avg.
Pass	22.25	41.89	55.59	61.35	63.24	66.84	51.86
Drive	28.70	47.86	57.80	63.22	65.37	70.07	55.50
High Pass	2.30	5.32	8.33	15.92	18.86	19.60	11.72
Header	10.30	19.87	27.11	29.69	31.18	32.19	25.05
Out	10.70	18.26	26.86	29.35	30.82	32.38	24.74
Throw-in	2.68	9.68	15.78	28.96	37.47	45.23	23.30
Cross	7.84	13.22	22.51	25.31	28.98	30.26	21.35
Ball Player Block	2.52	4.69	9.20	11.23	13.23	14.09	9.16
Shot	4.61	7.27	9.37	11.68	12.93	14.65	10.08
Successful Tackle	5.43	6.62	8.59	8.86	9.26	9.39	8.02
All	9.74	17.47	24.11	28.56	31.13	33.47	24.08

Table 3. Per-class performance of FAANTRA using RegNetY-400MF, trained on SN-BAA and SN-AS, with results based on the metrics described in Sec. 3 for $T_a = 5$ seconds.

ner [6]. Additionally, we include results for the SOTA Ball Action Spotting method, T-DEED, using a context window that matches the anticipation window (*i.e.*, $v^n = w^n$). This serves as an upper bound, where the model has full access to the anticipation window.

We observe a consistently large performance gap between FAANTRA and the upper bound, particularly evident in low tolerance metrics, highlighting the difficulty of the task –especially its localization component– and emphasizing the need for further research on anticipation methods in the football domain. Additionally, we observe that a larger feature extractor does not always improve performance when training only on SN-BAA, but it does when incorporating SN-AS. This suggests that the larger backbone benefits from additional data, but is prone to overfitting when limited to SN-BAA for anticipation. Furthermore, adding SN-AS clearly boosts performance, reinforcing the value of this auxiliary dataset. As expected, anticipating actions further in the future is more challenging, as shown by the performance drop when increasing the anticipation window from $T_a = 5$ to $T_a = 10$ seconds.

Additionally, Tab. 3 presents the per-class results of the best-performing FAANTRA model, using RegNetY-

400MF, trained on both SN-BAA and SN-AS, and for $T_a = 5$ seconds. We observe that the most frequent actions, such as passes and drives, are better captured by the model. In contrast, actions like headers, ball out, throw-ins or crosses, which often have observable cues before they occur, demonstrate intermediate performance. On the other hand, highly spontaneous actions with limited examples, such as shots, ball player blocks, or successful tackles, are not well captured by the model.

7. Ablation studies

To assess each component's contribution, we conduct an ablation study by systematically removing or modifying key modules. We evaluate the impact on the test split using the metrics defined in Sec. 3, reporting results averaged over two different seeds. We distinguish between task ablations, which are linked to the task definition; general model ablations, which involve components common to all methods performing the task; and model-specific ablations, which focus on the specific components of our baseline.

7.1. Task ablations

In this section, we study the task difficulty with varying anticipation window lengths, T_a. As shown in Fig. 3, performance decreases across almost all metrics as the anticipation window increases. This aligns with the nature of the task, where predicting actions further in advance becomes more difficult. The only exception is mAP@∞, which remains stable throughout. This suggests that while the performance related to the proportion of action classes predicted remains unaffected by longer anticipation windows, the challenge lies in localizing those actions. Given the already low scores for smaller windows ($T_a = 5$, $T_a = 10$), we focus the task on these window sizes, as reflected in the main results (see Tab. 2). Future improvements in action anticipation for these smaller windows will pave the way for tackling the more challenging task of larger anticipation windows. Following ablations are performed using $T_a = 5$.

7.2. General model ablations

In this subsection, we analyze the impact of five key components common to any potential action anticipation method: the input frame spatial resolution, temporal resolution, temporal context length, auxiliary action segmentation task within the context window, and prediction head types.

Spatial resolution. As shown in Tab. 4 (a), reducing the spatial resolution by half results in a performance drop of 5 points on average. This emphasizes the importance of higher resolutions for effectively capturing relevant information within the context window. Since broadcast videos often cover large areas of the field, with key ball-related actions occupying only a small portion of the frame, increased
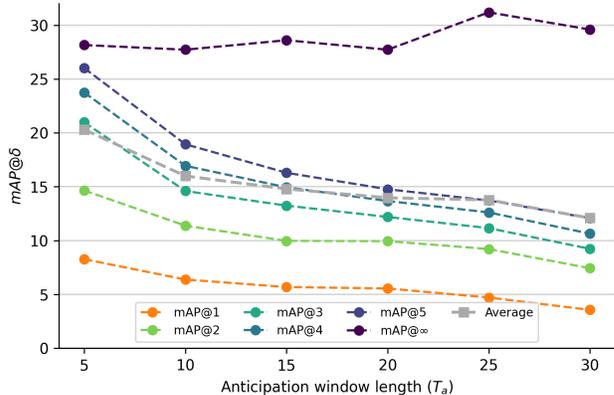

Figure 3. Anticipation window length analysis: Performance evaluation across different mAP@δ metrics for varying T_a anticipation windows.

Experiment		mAP@δ						
		$\delta = 1$	2	3	4	5	∞	Avg.
FAANTRA		8.27	14.65	20.98	23.74	26.01	28.15	20.30
(448 × 796, 6.25fps, w. AST, Q-Act)								
(a)	224 × 398	6.23	11.40	16.21	18.11	19.79	21.69	15.57
(b)[†]	3.125 fps	4.80	12.93	15.34	18.12	19.32	21.79	15.38
	6.25 fps	6.23	11.40	16.21	18.11	19.79	21.69	15.57
	12.5 fps	5.75	11.21	16.48	18.78	19.96	22.15	15.72
	25 fps	5.29	11.56	15.47	17.85	19.79	21.47	15.24
(c)	w/o AST	2.44	4.43	6.77	8.74	9.52	10.87	7.13
(d)	Q-EOS	6.60	12.28	17.89	19.83	21.63	23.20	16.90
	Q-Bckg	6.70	12.28	16.50	18.84	20.80	22.83	16.32
	Q-BCE	5.25	9.55	13.04	15.41	16.45	18.97	13.11
	Q-Hung(t)	7.38	10.28	13.01	14.97	16.52	21.36	13.92
	Q-Hung(a)	5.26	10.93	15.93	20.53	23.09	25.21	16.82
	Anchors	8.39	12.09	14.71	16.72	18.95	25.15	15.84

Table 4. Ablation studies on general model components from FAANTRA. We ablate the (a) input frame spatial resolution, (b) temporal resolution, (c) auxiliary segmentation task (AST), and (d) prediction head. [†] denotes ablations performed at lower spatial resolution (224 × 398) due to computational constraints.

resolution helps capture these important parts of the video more effectively but requires more memory to process.

Temporal resolution. In Tab. 4 (b), we observe that the optimal frame rate lies between 6.25 and 12.5 fps. Lower frame rates likely degrade performance due to missing information from large frame steps, while higher frame rates seem to introduce redundant information without benefiting the method, decreasing the temporal receptive field of the encoder-decoder and increasing computational cost.

Temporal context length. As shown in Fig. 4, performance initially improves as the context window length increases but plateaus or slightly declines beyond $T_c = 5$ seconds. This suggests that, for our proposed baseline, extending the context beyond 5 seconds offers minimal benefits while increasing computational cost. Given the fast-paced nature of football, information from more distant context appears to add little value for accurately anticipating actions.

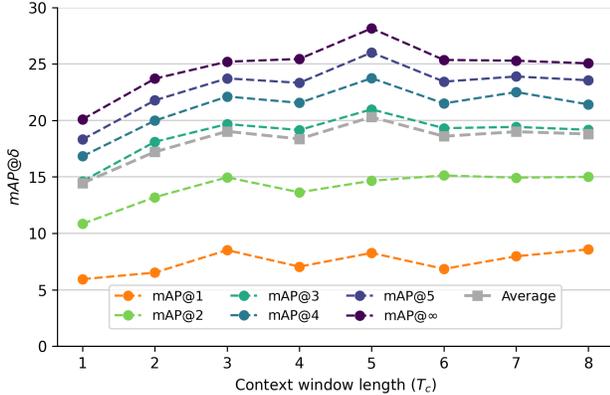**Auxiliary segmentation task.** Table 4 (c) highlights the

Figure 4. Context window length analysis: Performance evaluation across mAP@δ metrics for varying T_c context windows.

| Experiment | mAP@δ | | | | | | |
	$\delta = 1$	2	3	4	5	∞	Avg.
FAANTRA	8.27	14.65	20.98	23.74	26.01	28.15	20.30
(l_E=4, l_D=2, k=15, q=8)							
(a) $l_E = 2$	8.17	14.24	19.61	22.45	24.38	26.80	19.28
$\quad\;\; l_E = 6$	7.12	12.69	18.05	21.25	22.74	25.09	17.82
(b) $l_D = 4$	7.86	14.23	19.76	24.46	26.87	27.17	20.06
$\quad\;\; l_D = 6$	7.81	14.12	18.97	21.43	23.96	25.90	18.70
(c) $k = 7$	8.66	15.13	21.19	23.63	25.28	27.75	20.27
$\quad\;\; k = 23$	8.70	14.72	19.46	21.75	23.00	25.40	18.84
$\quad\;\; k = 31$	7.77	13.55	18.34	20.98	22.49	24.96	18.01
$\quad\;\;$ Global	6.33	12.28	18.42	21.18	23.05	24.96	17.70
(d) $q = 6$	8.18	13.86	18.78	21.55	23.44	26.62	18.74
$\quad\;\; q = 10$	8.85	14.72	19.92	23.16	24.73	26.95	19.72
$\quad\;\; q = 12$	7.27	13.20	18.07	20.63	22.05	24.48	17.62
$\quad\;\; q = 14$	7.32	13.39	18.68	21.51	23.74	25.98	18.44

Table 5. Ablation studies of FAANTRA's architectural components: (a) encoder layers l_E, (b) decoder layers l_D, (c) attention locality in the encoder k, and (d) number of learnable queries q.

7.3. Model-specific ablations

We ablate four architectural components specific to our baseline: (a) encoder layers, (b) decoder layers, (c) attention locality in the transformer encoder, and (d) number of learnable queries. As shown in Tab. 5, the optimal number of encoder and decoder layers is $l_E = 4$ and $l_D = 2$, although $l_D = 4$ results in comparable performance. The model performs better with smaller attention neighborhoods, with $k = 7$ and $k = 15$ yielding similar results. In contrast, more global attention shows no benefit, highlighting the advantage of starting with more localized attention that can expand to attend to more distant temporal positions in later layers. Finally, $q = 8$ queries appears to be optimal, which coincides with the maximum number of actions observed within a $T_a = 5$ second anticipation window.

8. Conclusion

This work introduced the first video-based action anticipation task in sports, presenting the SoccerNet Ball Action Anticipation dataset. Unlike other anticipation datasets, this task poses an added challenge due to the high spontaneity of football, driven by its adversarial nature. To evaluate the task, we adapted the $mAP@\delta$ metric from action spotting to the anticipation setting. Additionally, we proposed FAANTRA, the first baseline model for this task, and compared its performance to an action spotting upper bound, demonstrating both the feasibility and difficulty of the problem. Finally, we conducted extensive ablations to analyze the impact of different task settings, input configurations, and model architectures, highlighting the importance of high spatial resolution and an auxiliary action segmentation task within the context window to capture the semantics of action classes. By publicly releasing the dataset, baseline, and benchmark, we aim to promote reproducibility and guide future research in football action anticipation.

importance of the auxiliary segmentation task in our baseline. Removing this component from FAANTRA results in a substantial performance drop, emphasizing the need for segmentation of observed actions to guide the learning process and capture the semantics of different action classes, which aids in better anticipating future actions.

Prediction head. Table 4 (d) compares various alternative prediction heads to our baseline, Q-Act, which uses a query-based approach with components for actionness, action class, and temporal position. We evaluate five query-based alternatives and one anchor-based method: Q-EOS, which removes the actionness component and adds an End of Sequence (EoS) action class; Q-Bckg, which replaces EoS with a background class; Q-BCE, which adapts Q-Act by removing the actionness component and using sigmoid and binary cross-entropy for each class instead of softmax and cross-entropy; Q-Hung(t) which pairs ground-truth and predictions based on temporal position using the Hungarian algorithm instead of doing it sequentially; Q-Hung(a), which pairs based on action classes; and Anchors, as in [43], where each query is assigned with a temporal position. Additional details on these alternative prediction heads can be found in the supplementary material.

As shown in Tab. 4, Q-Act outperforms Q-EOS, Q-Bckg, and Q-BCE, highlighting the importance of the actionness component within the prediction head. This component simplifies the learning by predicting whether a query contains an action or not independently of its class. For Q-Hung(t) and Q-Hung(a), performance is better when pairing predictions and ground-truths based on action classes rather than temporal position. This makes sense given that the temporal location is more challenging to anticipate. However, this pairing does not outperform sequentially pairing predictions and ground truths. Lastly, using Anchors does not improve performance, likely due to the similar difficulty of action localization, as each query in the Anchors prediction head is itself based on a temporal position.

References

[1] Yazan Abu Farha and Juergen Gall. Uncertainty-aware anticipation of activities. In *IEEE/CVF Int. Conf. Comput. Vis. Work. (ICCV Work.)*, pages 1197–1204, Seoul, South Korea, 2019. 2

[2] Yazan Abu Farha, Qiuhong Ke, Bernt Schiele, and Juergen Gall. Long-term anticipation of activities with cycle consistency. In *DAGM German Conference on Pattern Recognition*, pages 159–173, 2021. 2, 5

[3] Ruslan Baikulov. Solution for SoccerNet ball action spotting challenge 2023. https://github.com/lRomul/ball-action-spotting, 2023. 2

[4] Kai-Shiang Chang, Wei-Yao Wang, and Wen-Chih Peng. Where will players move next? dynamic graphs and hierarchical fusion for movement forecasting in badminton. In *AAAI*, pages 6998–7005, Washington, D.C., USA, 2023. 1, 3

[5] 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, 2020. 2

[6] Anthony Cioppa, Silvio Giancola, Vladimir Somers, Victor Joos, Floriane Magera, Jan Held, Seyed Abolfazl Ghasemzadeh, Xin Zhou, Karolina Seweryn, Mateusz Kowalczyk, Zuzanna Mróz, Szymon Łukasik, Michał Hałoń, Hassan Mkhallati, Adrien Deliège, Carlos Hinojosa, Karen Sanchez, Amir M. Mansourian, Pierre Miralles, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Ghanem Bernard, Marc Van Droogenbroeck, Adam Gorski, Albert Clapés, Andrei Boiarov, Anton Afanasiev, Artur Xarles, Atom Scott, ByoungKwon Lim, Calvin Yeung, Cristian Gonzalez, Dominic Rüfenacht, Enzo Pacilio, Fabian Deuser, Faisal Sami Altawijri, Francisco Cachón, HanKyul Kim, Haobo Wang, Hyeonmin Choe, Hyunwoo J Kim, Il-Min Kim, Jae-Mo Kang, Jamshid Tursunboev, Jian Yang, Jihwan Hong, Jimin Lee, Jing Zhang, Junseok Lee, Kexin Zhang, Konrad Habel, Licheng Jiao, Linyi Li, Marc Gutiérrez-Pérez, Marcelo Ortega, Menglong Li, Nikita Lopatto, Milosz an Kasatkin, Norbert Nemtsev, Nikolay an Oswald, Oleg Udin, Pavel Kononov, Pei Geng, Saad Ghazai Alotaibi, Sehyung Kim, Sergei Ulasen, Sergio Escalera, Shanshan Zhang, Shuyuan Yang, Sunghwan Moon, Thomas B. Moeslund, Vasyl Shandyba, Vladimir Golovkin, Wei Dai, Won-Taek Chung, Xinyu Liu, Yongqiang Zhu, Youngseo Kim, Yuan Li, Yuting Yang, Yuxuan Xiao, and Zhihao Cheng, Zehua an Li. SoccerNet 2024 challenges results. *arXiv*, 2024. 2, 6

[7] 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 Huang, and Ziyu Meng. SoccerNet 2023 challenges results. *Sports Eng.*, 27(2):1–18, 2024. 2

[8] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for EPIC-KITCHENS-100. *Int. J. Comput. Vis.*, 130(1):33–55, 2021. 2, 3

[9] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. Online action detection. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 269–284. 2016. 2, 3

[10] 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/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 4503–4514, Nashville, TN, USA, 2021. 2

[11] Julien Denize, Mykola Liashuha, Jaonary Rabarisoa, Astrid Orcesi, and Romain Hérault. COMEDIAN: Self-supervised learning and knowledge distillation for action spotting using transformers. In *IEEE/CVF Winter Conf. Appl. Comput. Vis. Work. (WACVW)*, pages 518–528, Waikoloa, HI, USA, 2024. 2

[12] Yazan Abu Farha, Alexander Richard, and Juergen Gall. When will you do what? anticipating temporal occurrences of activities. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 5343–5352, Salt Lake City, UT, USA, 2018. 2

[13] Panna Felsen, Pulkit Agrawal, and Jitendra Malik. What will happen next? forecasting player moves in sports videos.

In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 3362–3371, Venice, Italy, 2017. 3

[14] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling LSTMs for action anticipation from first-person video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(11): 4021–4036, 2021. 2

[15] Harshala Gammulle, Simon Denman, Sridha Sridharan, and Clinton Fookes. Forecasting future action sequences with neural memory networks. In *Br. Mach. Vis. Conf. (BMVC)*, pages 1–12, Cardiff, United Kingdom, 2019. 2

[16] Silvio Giancola and Bernard Ghanem. Temporally-aware feature pooling for action spotting in soccer broadcasts. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 4485–4494, Nashville, TN, USA, 2021. 2

[17] 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, 2018. 2, 3

[18] 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., 2022. 2

[19] Silvio Giancola, Anthony Cioppa, Bernard Ghanem, and Marc Van Droogenbroeck. Deep learning for action spotting in association football videos. *arXiv*, abs/2410.01304, 2024. 1, 2, 3

[20] Harshayu Girase, Nakul Agarwal, Chiho Choi, and Karttikeya Mangalam. Latency matters: Real-time action forecasting transformer. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 18759–18769, Vancouver, Can., 2023. 2

[21] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pages 13485–13495, Montreal, QC, Canada, 2021. 2

[22] Ryota Goka, Yuya Moroto, Keisuke Maeda, Takahiro Ogawa, and Miki Haseyama. What to do and where to go next? action prediction in soccer using multimodal co-attention transformer. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 75–79, Melbourne, Victoria, Aust., 2024. 1, 3

[23] Dayoung Gong, Joonseok Lee, Manjin Kim, Seong Jong Ha, and Minsu Cho. Future transformer for long-term action anticipation. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 3042–3051, New Orleans, LA, USA, 2022. 2, 4, 5

[24] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4D: Around the world in 3,000 hours of egocentric video. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 18973–18990, New Orleans, LA, USA, 2022. 3

[25] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. AVA: A video dataset of spatio-temporally localized atomic visual actions. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 6047–6056, Salt Lake City, UT, USA, 2018. 2

[26] 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., 2023. 2

[27] 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/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3267–3279, Seattle, WA, USA, 2024. 2

[28] James Hong, Haotian Zhang, Michaël Gharbi, Matthew

Fisher, and Kayvon Fatahalian. Spotting temporally precise, fine-grained events in video. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 33–51, Tel Aviv, Israël, 2022. 2, 4, 5

[29] 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". *Comput. Vis. Image Underst.*, 155:1–23, 2017. 3

[30] Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 780–787, Columbus, OH, USA, 2014. 2, 3

[31] Yin Li, Miao Liu, and James M. Rehg. In the eye of beholder: Joint learning of gaze and actions in first person video. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 639–655, 2018. 2, 3

[32] 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 *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4794–4804, New Orleans, LA, USA, 2022. 2

[33] Junwei Liang, Lu Jiang, Juan Carlos Niebles, Alexander Hauptmann, and Li Fei-Fei. Peeking into the future: Predicting future person activities and locations in videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 2960–2963, Long Beach, CA, USA, 2019. 2

[34] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Int. Conf. Learn. Represent.*, New Orleans, LA, USA, 2019. 6

[35] Esteve Valls Mascaro, Hyemin Ahn, and Dongheui Lee. Intention-conditioned long-term human egocentric action anticipation. In *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pages 6037–6046, Waikoloa, HI, USA, 2023. 2

[36] Tushar Nagarajan, Yanghao Li, Christoph Feichtenhofer, and Kristen Grauman. Ego-topo: Environment affordances from egocentric video. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 160–169, Seattle, WA, USA, 2020. 2

[37] Megha Nawhal, Akash Abdu Jyothi, and Greg Mori. Rethinking learning approaches for long-term action anticipation. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 558–576, 2022. 2

[38] Ilija Radosavovic, Raj Prateek Kosaraju, Ross Girshick, Kaiming He, and Piotr Dollar. Designing network design spaces. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 10425–10433, Seattle, WA, USA, 2020. 4

[39] Fadime Sener, Dipika Singhania, and Angela Yao. Temporal aggregate representations for long-range video understanding. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 154–171, 2020. 2

[40] Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 21064–21074, New Orleans, LA, USA, 2022. 3

[41] Karolina Seweryn, Anna Wróblewska, and Szymon Łukasik. Survey of action recognition, spotting and spatio-temporal localization in soccer – current trends and research perspectives. *arXiv*, abs/2309.12067, 2023. 2

[42] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 510–526, 2016. 3

[43] João V. B. Soares, Avijit Shah, and Topojoy Biswas. Temporally precise action spotting in soccer videos using dense detection anchors. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 2796–2800, Bordeaux, France, 2022. 2, 8

[44] Sebastian Stein and Stephen J. McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *ACM Int. Jt. Conf. Pervasive Ubiquitous Comput.*, pages 729–738, Zurich, Switzerland, 2013. 2, 3

[45] Swathikiran Sudhakaran, Sergio Escalera, and Oswald Lanz. Gate-shift-fuse for video action recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9):10913–10928, 2023. 4

[46] Chen Sun, Abhinav Shrivastava, Carl Vondrick, Rahul Sukthankar, Kevin Murphy, and Cordelia Schmid. Relational action forecasting. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 273–283, Long Beach, CA, USA, 2019. 2

[47] 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, 2021. 2

[48] Joakim Valand, Haris Kadragic, Steven Hicks, Vajira Thambawita, Cise Midoglu, Tomas Kupka, Dag Johansen, Michael Riegler, and Pål Halvorsen. AI-based video clipping of soccer events. *Mach. Learn. & Knowl. Extr.*, 3(4):1–19, 2021. 2

[49] Joakim Valand, Haris Kadragic, Steven Hicks, Vajira Thambawita, Cise Midoglu, Tomas Kupka, Dag Johansen, Michael Riegler, and Pål Halvorsen. Automated clipping of soccer events using machine learning. In *Int. Symp. Multimedia (ISM)*, pages 210–214, Naple, Italy, 2021. 2

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, pages 6000–6010, Long Beach, CA, USA, 2017. 5

[51] Xinyu Wei, Patrick Lucey, Stuart Morgan, and Sridha Sridharan. Predicting shot locations in tennis using spatiotemporal data. In *Digit. Image Comput.: Tech. Appl.*, pages 1–8, Hobart, TAS, Australia, 2013. 1, 3

[52] Xinyu Wei, Patrick Lucey, Stephen Vidas, Stuart Morgan, and Sridha Sridharan. Forecasting events using an augmented hidden conditional random field. In *Asian Conf. Comput. Vis. (ACCV)*, pages 569–582, 2015. 3

[53] Fei Wu, Qingzhong Wang, Jiang Bian, Ning Ding, Feixiang Lu, Jun Cheng, Dejing Dou, and Haoyi Xiong. A survey on video action recognition in sports: Datasets, methods and

applications. *IEEE Trans. Multimedia*, 25:7943–7966, 2023. 1

[54] Artur Xarles, Sergio Escalera, Thomas B. Moeslund, and Albert Clapés. ASTRA: An Action Spotting TRAnsformer for soccer videos. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, page 93–102, Ottawa, Ontario, Can., 2023. 2

[55] Artur Xarles, Sergio Escalera, Thomas B. Moeslund, and Albert Clapés. T-DEED: Temporal-discriminability enhancer encoder-decoder for precise event spotting in sports videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3410–3419, Seattle, WA, USA, 2024. 2, 4

[56] Mingze Xu, Yuanjun Xiong, Hao Chen, Xinyu Li, Wei Xia, Zhuowen Tu, and Stefano Soatto. Long short-term transformer for online action detection. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, pages 1086–1099. 2021. 2

[57] Ce Zhang, Changcheng Fu, Shijie Wang, Nakul Agarwal, Kwonjoon Lee, Chiho Choi, and Chen Sun. Object-centric video representation for long-term action anticipation. In *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pages 6737–6747, Waikoloa, HI, USA, 2024. 2

[58] Qi Zhao, Shijie Wang, Ce Zhang, Changcheng Fu, Minh Quan Do, Nakul Agarwal, Kwonjoon Lee, and Chen Sun. AntGPT: Can large language models help long-term action anticipation from videos? In *Int. Conf. Learn. Represent.*, Vienna, Austria, 2024. 2

[59] Yue Zhao and Philipp Krähenbühl. Real-time online video detection with temporal smoothing transformers. In *Eur. Conf. Comput. Vis. (ECCV)*, pages 485–502. 2022. 2

[60] Yi Zhong and Wei-Shi Zheng. Unsupervised learning for forecasting action representations. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 1073–1077, Athens, Greece, 2018. 2

[61] Zeyun Zhong, Manuel Martin, Michael Voit, Juergen Gall, and Jürgen Beyerer. A survey on deep learning techniques for action anticipation. *arXiv*, abs/2309.17257, 2023. 3

[62] Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and Jurgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation. In *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pages 6057–6066, Waikoloa, HI, USA, 2023. 2

[63] Zeyun Zhong, Chengzhi Wu, Manuel Martin, Michael Voit, Juergen Gall, and Jürgen Beyerer. DiffAnt: Diffusion models for action anticipation. *arXiv*, abs/2311.15991, 2023. 2, 5

[64] Xin Zhou, Le Kang, Zhiyu Cheng, Bo He, and Jingyu Xin. Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection. *arXiv*, abs/2106.14447, 2021. 2