Noname manuscript No.
(will be inserted by the editor)

A Hitchhiker's Guide to Action Understanding: Advances, Challenges, and Outlooks

Alexandros Stergiou · Ronald Poppe

the date of receipt and acceptance should be inserted later

Abstract We have witnessed impressive advances in action understanding research, both in terms of performance as well as the diversification of tasks. Powered by datasets of expanded sizes and increased computation availability, current systems can provide fine- and coarse-grained descriptions of video scenes, extract segments corresponding to queries, synthesize unobserved parts of the video, and predict the context of videos. This survey provides a comprehensive review of recent advances in uni-and-multi-modal action understanding. We focus on the prevalent challenges, overview widely adopted datasets, and survey seminal works. We broadly distinguish three main scopes of tasks based on the temporal extent of their inputs. We discuss recognition tasks on the actions observed in full, prediction tasks on ongoing partially observed actions, and forecasting tasks for the subsequent unobserved action. This division allows us to identify specific challenges and focus on tasks relating to specific times during the execution of an action. Finally, we provide an outward look into future directions to address current shortcomings.

Keywords Action Understanding · t2 · t3

1 Introduction

Rapid technological advancements are embedded into multiple aspects of our everyday life. Digital processing of information has enabled great levels of automation enhancing our quality of life. These advancements allow us to capture, share, and consume footage of our or other's everyday experiences. The definition of algorithms able to *understand* these experiences and the actions performed has been of particular interest to the computer vision community since its early days.

In developmental psychology, action understanding has been explored across several distinct psychological processes (Thompson et al 2019). Such aspects include; **The capacity to understand the action performed**. This relates to the capability to differentiate between analogous actions (Gallese et al 1996; Jeannerod 1994) and the conceptualization of *how* an action is performed (Spunt et al 2011).

Determining the goal of the action. Works (Calvo-Merino et al 2005; Kohler et al 2002; Rizzolatti et al 2001) have studied action understanding in relation to immediate goals. This includes both the motor functions responsible for the execution of an action and the sensory perception of actions performed by others.

Determining the actor's intention. In contrast to goals that can be understood immediately after the action's completion, intentions are the high-level motivations for performing an action (Kilner 2011). Such intentions have been defined as the sequential grouping of the individual actions (Fogassi et al 2005) and their abstract associated target (Uithol et al 2011).

Inspired by the cognitive aspects of action understanding, we define three broad *temporal scopes* and group seminal machine vision action understanding tasks. We visualize the action progression in Figure 1 with a currently performed action followed by a subsequent action. Tasks that require the first action to be *observed in full* are broadly termed as **recognition** tasks and infer information such as the action categories or high-

A. Stergiou
Faculty of Electrical Engineering, Mathematics and Computer Science at the University of Twente, Drienerlolaan 5, 7522 NB Enschede, The Netherlands

R.Poppe
Department of Information and Computing Sciences at Utrecht University, Princetonplein 5, 3584 CC Utrecht, The Netherlands

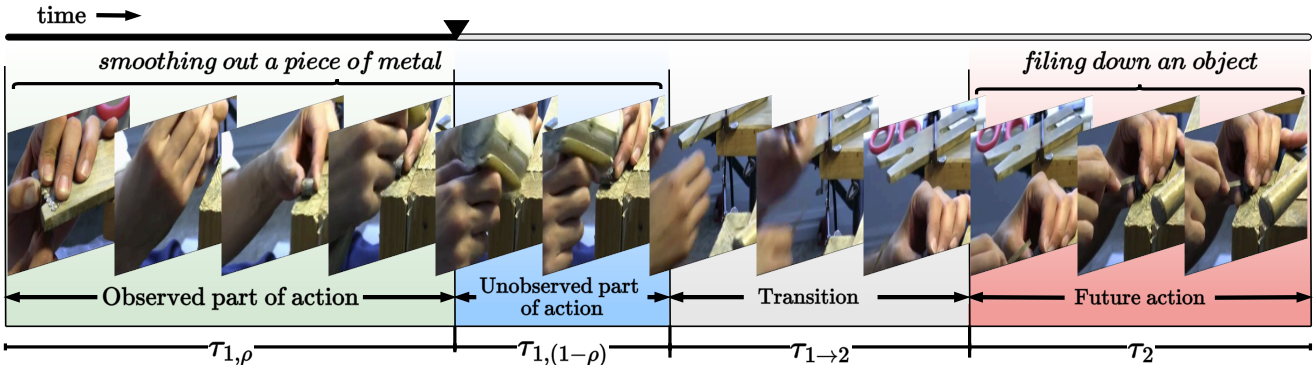


Fig. 1. **Action understanding tasks**. The progress of the video is shown by the top bar. From the currently performed, action of total duration τ_1, only the $\tau_{1,\rho} < \tau_1$ part is currently observable. After a transition period $0 \leq \tau_{1\to 2}$ another action is performed with duration τ_2. **Action recognition** tasks are based on full observations of the action at τ_1. **Action prediction** uses only part $\tau_{1,\rho}$ of ongoing action. **Action forecasting** uses current action at τ_1 to predict future actions. Example selected from (Wang et al 2019b).

level semantics. Prediction about the ongoing actions are made from the only *partial observations* without the action being completed, we define such tasks as **prediction**. In **forecasting** tasks, reasoning from the currently observed action is used for the actions *not yet observed*. For each temporal scope, we discuss the previous surveys that overview advancements in tasks and application domains.

Recognition. Early surveys on action recognition have primarily focused on works based on motion modeling. Aggarwal et al (1994) defined a taxonomy of motions based on rigidness and subsequently (Aggarwal et al 1998) discuss further subdivisions of motion categories based on prior knowledge of the objects' shape. Cedras and Shah (1995) and later Moeslund and Granum (2001) discussed temporal modeling approaches in the context of classification and tracking. As overviewed by Buxton (2003) tracking approaches have also been used in more complex tasks such as behavior analysis or non-verbal human interactions. Human motion modeling and estimation divisions were later discussed by Poppe (2007). More explicit scopes such as action classification, and localization (Weinland et al 2011), behavior understanding (Chaaraoui et al 2012), or surveillance applications (Vishwakarma and Agrawal 2013) emerged with later recognition advancements. They explored task-specific methods and more semantically complex topics, primarily in relation to classification. Turaga et al (2008) and Poppe (2010) both discussed approaches addressing atomic actions and group activities. Herath et al (2017) provided an initial summary of approaches using learned-features for action recognition. Following surveys overviewed the adoption and adaptation of deep learning approaches for topics such as; depth-based motion recognition (Wang et al 2018c), activity recognition (Beddiar et al 2020), human-human interactions (Stergiou and Poppe 2019), and pose estimation (Zheng et al 2020). Sun et al (2022b) reviewed approaches across multiple modalities combining motion features, audio, and vision. More recently, Selva et al (2023) discussed attention-based approaches for video tasks while Schiappa et al (2023) focused on self-supervised approaches.

Prediction. The recent advancements in action recognition have also enabled the definition of more challenging predictive tasks from partial observations. Rasouli (2020) discussed four main domains of predictive models including video, action, trajectory, and motion prediction. Kong and Fu (2022) described recent advancements for action recognition and prediction focusing on their applications to domains such as robot vision, surveillance, and driver behavior prediction. The surveys of Dhiman and Vishwakarma (2019) and later Ramachandra et al (2020) overviewed predictive methods for anomaly detection.

Forecasting. Similarly, forecasting tasks have also recently been at the forefront of action understanding research. Rodin et al (2021) discussed prevalent approaches for future action anticipation in egocentric videos. Zhong et al (2023b) overviewed both short and long-term anticipation approaches. Hu et al (2022c) provided a review of online and anticipation methods. More recently, Plizzari et al (2024) discussed challenges in egocentric videos and discussed future outlooks for multiple scopes including forecasting tasks.

Despite the extensive coverage of topics, prior surveys focus on specific aspects of action understanding. As summarized in Table 1, an overview over critical works of the past decades that groups multiple aspects

of action understanding is missing from the current literature.

In this survey we address the diversity of action understanding tasks by their use of the temporal dimension. We overview the general methods for modeling actions in videos over the years in 2. We discuss datasets and benchmarks used by prior works in Section 3. We then first distinguish the tasks that require actions to be observed in full to infer semantics in Section 4. Predictive tasks that aim to predict ongoing actions in which only parts of the action are observed are overviewed in Section 5. We overview tasks and anticipative approaches next in Section 6. For each of these directions we outline main challenges and provide our insights as to what the future of action understanding may look like in Section 7. We conclude with a summary of this survey's findings and general thoughts in Section 8.

2 Modeling actions in videos

Video is composed of spatial information that relates to visual aspects of objects, background information, or the visual context of scenes. It also includes temporal information corresponding to changes of these visual aspects over time. In this section we overview two general approaches for encoding space-time information from videos without explicitly relating them to tasks. The first set of approaches discussed in Section 2.1 model spatial features and temporal information separately. The second create joint spatio-temporal representations that are overviewed in Section 2.2.

2.1 Separating visual and temporal information

Tracking and template matching. Early works (Bobick and Davis 2001) have considered template-matching approaches for determining spatial and temporal locations where motions occur. Template-based methods have also been explored over local patches (Shechtman and Irani 2005), through correlation filters (Rodriguez et al 2008), or based on voxel similarity (Ke et al 2007). A different direction of research has considered the discovery of temporal patterns by tracking visual features over time (Cipolla and Blake 1990; Isard and Blake 1998; Rohr 1994).

Local descriptors. A key characteristic of actions is appearance changes over time. Multiple methods have proposed approaches to associate changes in local features with actions. Mikolajczyk and Uemura (Mikolajczyk and Uemura 2008) clustered local features to tree representations and related them to action categories. Similarly, pose-based primitives (Thurau and Hlavác 2008), temporal bins (Nowozin et al 2007), pictorial structures (Tran et al 2012), and graphical structures of the actions (Ni et al 2014) have been used with local descriptor features. Other approaches (Gupta et al 2009; Yao and Fei-Fei 2010) cast action recognition as a structural connectivity task by recognizing parts of objects and understanding actions by pose. Following methods have also extended this to individual regions (Ikizler-Cinbis and Sclaroff 2010), poselet clusters (Pishchulin et al 2013), decision trees (Rahmani et al 2014), and covariance matrices (Kviatkovsky et al 2014).

Spatial convolutions. CNNs have become widely used feature extractors for a variety of vision tasks. An initial effort (Karpathy et al 2014) to fuse temporal information with CNN's static features included the combination of frame embeddings either in the first few layers or final layers of the architecture. Others explored the factorization of frame embeddings (Sun et al 2015), frame ranking (Fernando et al 2015), pooling (Fernando et al 2016), salient region focus (Girdhar and Ramanan 2017; Zong et al 2021), and relation reasoning between neighboring frames (Zhou et al 2018a). Le et al (2011) spatially convolved videos over dyadic combinations of the spatial and temporal dimensions. Seminal efforts focused on single volumes to represent motion (Bilen et al 2016; Chung and Zisserman 2016; Iosifidis et al 2012) or learned the correlation and exclusion of action classes (Hoai and Zisserman 2015). Tran et al (2018) proposed convolutional blocks based on spatial and temporal kernels creating more efficient video models. As densely-sampled frames may include redundancies, Lin et al (2019) proposed shifting features at subsequent frames to model actions in both online and offline settings. (Sudhakaran et al 2020)

Temporal recursion. A parallel line of research works have processed frame embeddings of spatial CNNs with recurrent units (Ballas et al 2015; Dwibedi et al 2018; Yue-Hei Ng et al 2015; Ullah et al 2017) to temporally model information. Efforts Donahue et al (2015); Srivastava et al (2015) have encoded frame features and learned to changes in appearance over time with LSTMs. Similarly, for multi-actor action recognition, Wang et al (2017b) used three individual pathways with LSTMs for person action, group action, and scene recognition.

Two-stream models. An alternative approach considers the inclusion of a motion-specific stream in the model pipeline. Two-stream models (Simonyan and Zisserman 2014) model motion explicitly through an optical flow stream. Further extensions (Feichtenhofer et al 2016) included the fusion of flow and spatial streams at intermediate layers. Other approaches towards sharing information across appearance and motion streams

Table 1. **Action understanding surveys through the years**. Columns highlight the temporal scope of surveys; with overviews of recognition (Rec.), prediction (Pred.), or anticipation (Ant.) works. Three broad objectives are tracked across surveys; multi-modality (MM), self-supervision (Self-Sup.), and multi-view (MV). The two tasks of human interactions (HI) to objects or other humans, and long video understanding (LVU)., are also reported per survey. Scopes/objectives/tasks addressed partially by surveys are denoted with (partial) and the main focus is denoted with ✔.

Author(s)	Year	#Papers	Temporal Scope			Objectives			Tasks	
			Rec.	Pred.	Ant.	MM	Self-Sup.	MV	HI	LVU
Aggarwal et al (1994)	1994	69	(partially)							
Cedras and Shah (1995)	1995	76	(partially)							
Aggarwal et al (1998)	1998	104	(partially)							
Aggarwal and Cai (1999)	1999	51	(partially)							
Moeslund and Granum (2001)	2001	155	(partially)							
Buxton (2003)	2003	88	(partially)						✔	
Moeslund et al (2006)	2006	424	✔			(partially)				
Yilmaz et al (2006)	2006	160	(partially)	(partially)					(partially)	
Poppe (2007)	2007	125	✔			(partially)				
Turaga et al (2008)	2008	144	✔			(partially)			✔	
Poppe (2010)	2010	180	✔							
Weinland et al (2011)	2011	153	✔			(partially)		(partially)	✔	
Chaaraoui et al (2012)	2012	123	✔			(partially)		✔	✔	
Metaxas and Zhang (2013)	2013	188						(partially)	✔	
Vishwakarma and Agrawal (2013)	2013	231	✔			(partially)				
Herath et al (2017)	2017	161	✔							
Wang et al (2018c)	2018	182	✔	(partially)		(partially)		(partially)	(partially)	
Dhiman and Vishwakarma (2019)	2019	208			✔	(partially)			(partially)	
Hussain et al (2019)	2019	141	✔			✔		(partially)	✔	
Stergiou and Poppe (2019)	2019	178	✔						✔	
Yao et al (2019)	2019	106	✔							
Zhang et al (2019a)	2019	127	✔							
Beddiar et al (2020)	2020	237	✔		(partially)			(partially)	✔	
Ramachandra et al (2020)	2020	109		✔						
Zheng et al (2020)	2020	317				✔				
Rasouli (2020)	2020	333		✔						
Pareek and Thakkar (2021)	2021	218	✔			(partially)				
Rodin et al (2021)	2021	156			✔	(partially)		✔		✔
Song et al (2021)	2021	157	✔							
Sun et al (2022b)	2022	503	✔			(partially)	✔			
Kong and Fu (2022)	2022	337	✔	(partially)				✔	(partially)	
Hu et al (2022c)	2022	168	✔		✔				✔	(partially)
Oprea et al (2022)	2022	211	✔	✔						
Schiappa et al (2023)	2023	216	✔				✔	✔		✔
Selva et al (2023)	2023	209	✔							
Wang et al (2023a)	2023	229	✔			(partially)		(partially)		
Zhong et al (2023b)	2023	207			✔	✔	✔	✔		✔
Ding et al (2023)	2023	168	✔					✔		(partially)
Plizzari et al (2024)	2024	367	✔		✔	✔		✔	✔	✔
Stergiou and Poppe (**this survey**)	2024	**843**	✔	✔	✔	✔	✔	✔	✔	✔

used cross-stream connections (Feichtenhofer et al 2017), concatenated appearance and motion volumes (Jain et al 2015), or recurrent layers (Singh et al 2016). Wang et al (2016b) proposed to segment videos into individual snippets, process them in parallel, and fuse class scores from each snippet. Works (Wang et al 2017c) have also explored encoding of visual (RGB) and motion (OF) at multiple levels of abstractions. Improvements in inference speeds of two-stream models have also been reported with the addition of motion vectors (Zhang et al 2016) or key volume mining (Zhu et al 2016). Although such approaches have included a new research direction in modeling videos, the representation of motion with precomputed features limits the capabilities of learned backbones. Sevilla-Lara et al (2019) empirically showed that one of the main limitations in optical flow representations is capturing accurate movements near the edges of objects.

2.2 Jointly encoding space time

In contrast to modeling video information separately to the vision characteristics and the temporal information, works have also considered joint spatiotemporal volumes.

Part-based representations. SpatioTemporal Interest Points (STIPs) (Laptev and Lindeberg 2003) extended spatial detection methods (Förstner and Gülch 1987; Harris et al 1988) to space and time with local activity endpoints. STIPs features were later used as histogram codewords (Schuldt et al 2004). Similarly, (Liu and Shah 2008; Oikonomopoulos et al 2005) explored salient points based on peaks of activity variations. Approaches have also studied action-relevant temporal locations across viewpoints (Yilmaz and Shah 2006) and view-invariant trajectories (Sheikh et al 2005). Dollár et al (2005) proposed modeling periodic motions through sparse distributions of points of interest. This feature

extractor prompted subsequent works (Niebles et al 2008) with the actions classified through a codebook of features.

Holistic stochastic representations. Approaches have also explored the recognition of actions based on global information. Efros et al (2003) created volumes representing different parts of the body regressing towards previously seen action fragments. Subsequent works have explored representations of objects over attributes such as shape (Gorelick et al 2006; Jia and Yeung 2008), movements (Sun et al 2009), and interest points (Wong and Cipolla 2007). Approaches also study the use of multiple features and temporal scales (Amer and Todorovic 2012; Liu et al 2008; Zelnik-Manor and Irani 2001; Yang et al 2020a). The representation of actions has also been explored in other spatiotemporal volumes. Blank et al (2005) proposed the concatenation of 2D silhouettes for the space-time shapes corresponding to actions. Sadanand and Corso (Sadanand and Corso 2012) proposed a bank of volumetrically pooled features containing high-level representations of the actions which are then classifier by an SVM.

3D CNNs. Orthogonal to hand-crafted features, works (Baccouche et al 2011; Ji et al 2012; Taylor et al 2010; Tran et al 2015) have extended convolutions to encode space and time jointly. Subsequent works have also extended established image models to video (Hara et al 2018). The efficiency of video models has also been explored in subsequent works that process videos by either creating distinct spatiotemporal volumes across channels (Chen et al 2018c), using tiled 3D kernels (Hegde et al 2018), channel-separated convolutions (Jiang et al 2019b; Luo and Yuille 2019; Tran et al 2019), temporally residual connections (Qiu et al 2017), global feature fusion (Qiu et al 2019), resolution reductions (Chen et al 2019; Stergiou and Poppe 2021b), or relating appearance to spatiotemporal embeddings (Wang et al 2018b; Zhou et al 2018d). Carreira and Zisserman (Carreira and Zisserman 2017) integrated 3D convolutions to two-stream models for modeling motion both implicitly in the RGB stream and explicitly in the optical flow stream. The resulting I3D model has been widely adopted as a baseline by subsequent works. Instead of the conversion of strong image models to video, works have also proposed architectures specifically for action recognition (Feichtenhofer 2020; Kondratyuk et al 2021; Liu et al 2022g). Works have also utilized visual context from the scenes of actions, by either scene-type objectives (Choi et al 2019), decoupling scene and motion features (Wang et al 2021a), or by fusing motion and scene information (Stergiou and Poppe 2021a). Feichtenhofer et al (2019) proposed a dual pathway video model; with a slow pathway operating over low frame rates for spatial semantics, and a fast pathway with a high frame rate for motion. Similarly, Wang et al (2020a) included a contrastive objective for learning the pace in videos. Xu et al (2019a) explored temporal reasoning from clip order prediction as an additional task to improve action recognition. Works (Ji et al 2020; Hussein et al 2019; Varol et al 2017) have also extended 3D CNNs to longer sequences by segmenting videos with multiple temporal patches.

Spatiotemporal attention. Sharma et al (2015) used visual attention to localize action regions from CNN features with temporal information then modeled by recurrent layers. Du et al (2017) identified spatial features from multiple frames which are then temporally attended based on their relevance to the action. Chen et al (2018b) aggregated and propagated global information by attending over convolution features. Similarly, Wang et al (2018d) introduced non-local operations with bi-directional attention blocks over convolutions. Another early application of attention (Girdhar et al 2019) was based on regional proposals and the creation of feature banks (Wu et al 2019a) in longer videos. Later, the introduction of vision transformers (Dosovitskiy et al 2020) as an approach for encoding visual information through region-based tokenization, has also led to its adaptation for video inputs. Works on video transformers explored different attention configurations for spatiotemporal volumes (Arnab et al 2021a; Bertasius et al 2021). Others explored token selection (Bulat et al 2021; Ryoo et al 2021; Zha et al 2021), and inclusion of contextual information (Kim et al 2021c). Liu et al (2022h) proposed attention computed over shifted non-overlapping windows. Feature hierarchies and latent resolution reductions also led to models with improvements in efficiency (Fan et al 2021; Li et al 2022f) and memory use (Wu et al 2022b). Following architectures such as MViT (Yan et al 2022), Hiera (Ryali et al 2023), and UniFormer (Li et al 2022c), have further improved both performance and capacities of video models. More recently, self-supervision has been widely adopted as a pre-training approach with either contrastive-learning (Chen et al 2020c) or token masking (He et al 2022b). Xing et al (2023) increased the complexity of the contrastive objective with pseudo labels and token mixing from different inputs. Other masking approaches have considered the extension of masked autoencoders to video data (Feichtenhofer et al 2022; Wei et al 2022). Subsequent works have explored adaptive token masking (Bandara et al 2023), double masking on the encoder and decoder (Wang et al 2023c), fusion of tokens (Kim et al 2024a), or teacher-student masked autoencoders (Wang et al 2023e).

Video-language models. Recently, semantics of language representations learned by Large Language Models (LLMs) (Brown et al 2020; Touvron et al 2023) have also been used as a supervisory signal for vision tasks (Li et al 2023a; Liu et al 2024a; Radford et al 2021). Initial efforts (Zellers et al 2021) matched frame-level encodings to corresponding LLM embeddings of captions. Approaches have further optimized image-based encodings over frames by pooling spatial tokens (Yu et al 2022a), cross-modal skip connections (Xu et al 2023a), cross attending over modalities (Alayrac et al 2022), and jointly attending visual and text embeddings (Maaz et al 2023). Static features provide only a limited view of videos. High-level visual concepts in videos are based on both space and time. Thus, works have also used space and time vision encoders (Piergiovanni et al 2024) with further extensions (Wang et al 2024b; Zhao et al 2024a) using two-step self-supervised pre-training with video to text alignment and video masking.

3 Video datasets of human actions

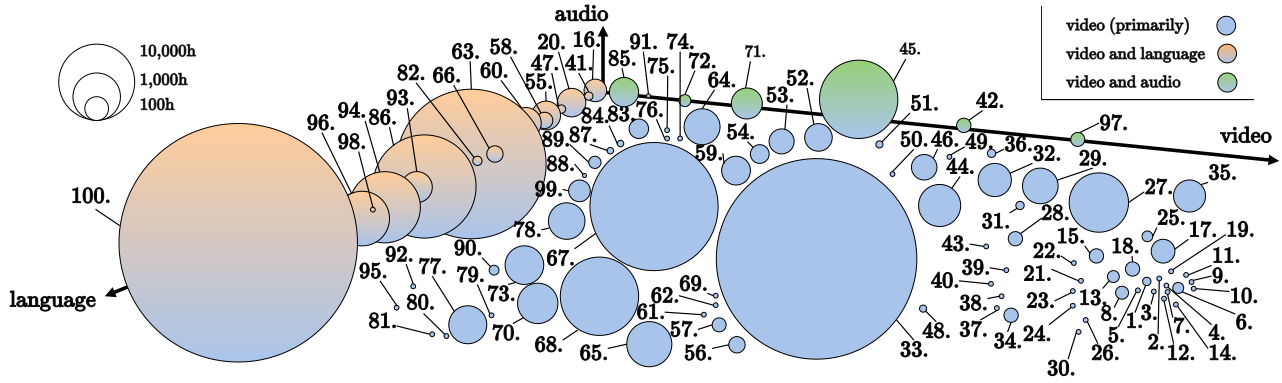
A significant effort has been made to collect diverse videos to comprise baselines for action understanding tasks. We overview two general directions. The first set of datasets discussed in Section 3.1 includes general-purpose datasets used for baselining models over multiple tasks or as large datasets for model pre-training. Section 3.2 details datasets collected for evaluating models over specific modalities or tasks.

3.1 General datasets

Over the past two decades, video datasets have scaled up the number of available videos and presented new and more robust baselines. We chronologically present widely-adopted benchmarks in Table 2. Initial efforts (Schuldt et al 2004; Gorelick et al 2007) have primarily focused on atomic actions such as categorizing *walking*, *running*, *hand waving* etc. of low motion magnitudes. Later efforts before web-based crawling of videos in large-scale, the majority of datasets were comprised of videos from either TV shows and movies (Laptev and Pérez 2007; Laptev et al 2008; Marszalek et al 2009; Patron-Perez et al 2010; Kuehne et al 2011) or sports (Rodriguez et al 2008; Liu et al 2009; Reddy and Shah 2013; Niebles et al 2010). A step towards establishing large-scale datasets for the video domain was made with the introduction of Kinetics (Carreira and Zisserman 2017), Sports-1M (Karpathy et al 2014), and YouTube-8M (Abu-El-Haija et al 2016), that included videos sourced from the web over diverse sets of actions. These datasets paved the way as general benchmarks for pertaining models that can be adapted for action-type-specific smaller datasets such as UCF-101 (Soomro et al 2012) and ActivityNet (Caba Heilbron et al 2015). In tandem, domains such as egocentric vision, human-object interactions, and hierarchical action understanding have also gained popularity prompting the creation of domain-specific datasets. EGTEA Gaze+ (Li et al 2015), EPIC KITCHENS (Damen et al 2022), and later EGO4D (Grauman et al 2022) have been the main datasets and benchmarks for egocen-

Table 2. **Action recognition datasets and benchmarks**. Groups are formed based on the year of release denoted with Y. Number of classes, video instances, and actors are denoted with # Cls., # Inst., and # Act. The average duration per annotation is shown in the AD column. Short descriptions per dataset are discussed in the Context column.

Y	Dataset	# Cls.	# Inst.	# Act.	AD	Context
2004-2007	KTH (Schuldt et al 2004)	6	2K	25	∼2.5s	Grayscaled videos of motions
	Weizmann (Gorelick et al 2007)	10	90	8	∼12s	Low-res. atomic motions
	Coffee & Cigarettes (Laptev and Pérez 2007)	2	245	5	5s	Smoking/drinking in movies
	CASIA Action (Wang et al 2007)	15	1446	24	N/A	Outdoor activities
2008-2014	UCF Sports (Rodriguez et al 2008)	9	150	<100	∼5s	Sports videos
	Hollywood (Laptev et al 2008)	8	475	<100	∼16s	Actions in movies
	UT-interaction (Ryoo and Aggarwal 2009)	6	90	60	∼17s	Dyadic human interactions
	CMU-MMAC (la Torre Frade et al 2008)	5	182	43	∼7m	Multiview recipe preparations
	UCF-11 (Liu et al 2009)	11	1K	100+	∼5s	Actions in YouTube videos
	Hollywood2 (Marszalek et al 2009)	12	3K	100+	∼12s	Actions from movies
	TV-HI (Patron-Perez et al 2010)	4	300	100+	∼3s	Interactions in TV shows
	UCF-50 (Reddy and Shah 2013)	50	5K	100+	∼15s	Web-sourced videos
	Olympic Sports (Niebles et al 2010)	16	800	100+	∼3s	Actions in sports
	HMDB-51 (Kuehne et al 2011)	51	7K	100+	∼3s	Actions from movies
	CCV (Jiang et al 2011)	20	9K	100+	∼80s	Web-sourced videos
	UCF-101 (Soomro et al 2012)	101	13K	100+	∼15s	Action with hierarchies
	CAD-60 (Sung et al 2012)	12	60	<30	∼45s	Atomic actions in RGB-D
	MPII (Rohrbach et al 2012)	65	5.6K	100+	∼11m	Web-source actions
	ADL (Pirsiavash and Ramanan 2012)	32	436	20	∼1.3s	Videos of daily activities
	50 Salads (Stein and McKenna 2013)	17	899	25	∼37s	Salad making videos
	J-HMDB (Jhuang et al 2013)	21	928	100+	∼1.2s	Videos with joints positions
	CAD-120 (Koppula et al 2013)	12	120	<60	∼45s	Extension of CAD-60
	Penn Action (Zhang et al 2013)	15	2.3K	100+	∼2s	Web-sourced atomic actions
	Sports-1M (Karpathy et al 2014)	487	1M	1,000+	∼9s	Sports actions/activities
2015-2018	EGTEA Gaze+ (Li et al 2015)	106	15K	32	∼28s	Egocentric actions w/ gaze
	ActivityNet-100 (Caba Heilbron et al 2015)	100	5K	100+	∼2m	Untrimmed web videos
	Watch-n-Patch (Wu et al 2015)	21	2K	7	∼30s	Daily activities in RGB-D
	NTU-RGB-60 (Shahroudy et al 2016)	60	57K	40	∼2s.	Multi-sensory actions
	ActivityNet-200 (Caba Heilbron et al 2015)	200	15K	100+	∼2m	ActivityNet-100 extension
	YouTube-8M (Abu-El-Haija et al 2016)	N/A	8M	N/A	N/A	Multi-labelled videos
	Charades (Sigurdsson et al 2016)	157	67K	267	∼30s	Daily activities videos
	ShakeFive2 (Van Gemeren et al 2016)	5	153	33	∼7s	Interactions with pose data
	DALY (Weinzaepfel et al 2016)	10	510	100+	4m	Untrimmed YouTube videos
	OA (Li and Fritz 2016)	48	480	<100	5s	Ongoing actions
	CONVERSE (Edwards et al 2016)	10	N/A	N/A	N/A	Human interactions
	TV-Series (De Geest et al 2016)	30	6,2K	100+	∼2s	Actions from TV series
	Volleyball (Ibrahim et al 2016)	6	1.4K	<100	<1s	Group actions in volleyball
	MSR-VTT (Xu et al 2016)	200K	7.1K	1,000+	∼20s	Video captions
	Okutama Action (Barekatain et al 2017)	12	4.7K	∼400	∼60s	Aerial views of action
	K-400 (Kay et al 2017)	400	306K	1,000+	∼10s	Web-sourced short actions
	Smthng-Smthng v1 (Goyal et al 2017)	174	109K	100+	∼4s	Human actions with objects
	MultiTHUMOS (Yeung et al 2018)	65	39K	100+	∼3s	Densely labeled actions
	Diving-48 (Li et al 2018b)	48	18K	N/A	∼3s	Diving sequences
	EK-55 (Damen et al 2018)	2,747	40K	35	∼3s	Egocentric actions in kitchens
	K-600 (Carreira et al 2018)	600	495K	100+	∼10s	Extension of K-400
	VLOG (Fouhey et al 2018)	30	122K	10.7K	∼10s	Actions in lifestyle VLOGs
	AVA (Gu et al 2018)	80	430	100+	15m	Localized atomic actions
2019-now	NTU-RGB-120 (Shahroudy et al 2016)	120	114K	106	∼2s.	Multi-sensory actions
	Charades-Ego (Sigurdsson et al 2018)	156	7.8K	100+	∼29s	Daily indoor activities
	Smthng-Smthng v2 (Goyal et al 2017)	174	221K	100+	∼4s	Human actions with objects
	K-700 (Carreira et al 2019)	700	650K	1,000+	∼10s	Extension of K-600
	Moments in Time (Monfort et al 2019)	339	1M	1,000+	3s	Short dynamic scenes
	HACS (Clips) (Zhao et al 2019)	200	1.5M	1,000+	2s	Action over fixed durations
	IG65M (Ghadiyaram et al 2019)	N/A	65M	N/A	N/A	Actions in Instagram videos
	Toyota Smarthome (Dai et al 2022a)	31	16K	18	21m	Senior home activities
	AViD (Piergiovanni and Ryoo 2020)	887	450K	1,000+	∼9s	Anonymized videos
	HVU (Diba et al 2020)	3K	572K	1,000+	∼10s	Hierarchy of semantics
	Action-Genome (Ji et al 2020)	453	10K	100+	∼1s	Daily home activities
	K-700 (2020) (Smaira et al 2020)	700	647K	1,000+	∼10s	Update of K-700
	FineGym (Shao et al 2020)	530	33K	100+	∼10m	Gymnastics videos
	RareAct (Miech et al 2020a)	122	7.6K	100+	10s	Unusual actions
	HAA500 (Chung et al 2021)	500	10K	1,000+	∼2s	Atomic actions
	MultSports (Li et al 2021b)	4	3.2K	100+	∼21s	Localized sports actions
	MOMA (Luo et al 2021)	136	12K	100+	∼ 10s	Hierarchical actions
	WebVid-2M (Bain et al 2021)	N/A	2M	1,000+	∼4s	Video-image pairs
	HOMAGE (Rai et al 2021)	453	5.7K	40	∼2s	Extension of (Ji et al 2020)
	EK-100 (Damen et al 2022)	4,053	90K	37	∼3s	Egocentric actions
	FineAction (Liu et al 2022f)	106	103K	1,000+	∼7s	Hierarchies for TAL
	EGO4D (Grauman et al 2022)	1000+	9.6K	931	∼48s	Diverse egocentric videos
	Assembly-101 (Sener et al 2022)	1.3K	4.3K	53	∼2s	Procedural activities
	Ego-Exo-4D (Grauman et al 2024)	689	5,035	740	∼5m	Multi-modal multi-view videos



1. KTH (Schuldt et al 2004)
2. Weizmann (Gorelick et al 2007)
3. Coffee & Cigarettes (Laptev and Pérez 2007)
4. CASIA Action (Wang et al 2007)
5. UCF Sports (Rodriguez et al 2008)
6. Hollywood (Laptev et al 2008)
7. UT-Interaction (Ryoo and Aggarwal 2009)
8. CMU-MMAC (la Torre Frade et al 2008)
9. UCF-11 (Liu et al 2009)
10. Hollywood2 (Marszalek et al 2009)
11. TV-HI (Patron-Perez et al 2010)
12. Humaneva Sigal et al (2010)
13. UCF-50 (Reddy and Shah 2013)
14. Olympic Sports (Niebles et al 2010)
15. HMDB-51 (Kuehne et al 2011)
16. MSVD (Chen and Dolan 2011)
17. CCV (Jiang et al 2011)
18. UCF-101 (Soomro et al 2012)
19. CAD-60 (Sung et al 2012)
20. MPII (Rohrbach et al 2012)
21. ADL (Pirsiavash and Ramanan 2012)
22. 50 Salads (Stein and McKenna 2013)
23. AVENUE (Lu et al 2013)
24. J-HMDB (Jhuang et al 2013)
25. CAD-120 (Koppula et al 2013)
26. Penn Action (Zhang et al 2013)
27. Sports-1M (Karpathy et al 2014)
28. EGTEA Gaze+ (Li et al 2015)
29. ActivityNet-100 (Caba Heilbron et al 2015)
30. Watch-n-Patch (Wu et al 2015)
31. NTU-RGB-60 (Shahroudy et al 2016)
32. ActivityNet-200 (Caba Heilbron et al 2015)
33. YouTube-8M (Abu-El-Haija et al 2016)
34. Charades (Sigurdsson et al 2016)
35. ShakeFive2 (Van Gemeren et al 2016)
36. DALY (Weinzaepfel et al 2016)
37. OA (Li and Fritz 2016)
38. CONVERSE (Edwards et al 2016)
39. TV-servies (De Geest et al 2016)
40. Volleyball (Ibrahim et al 2016)
41. MSR-VTT (Xu et al 2016)
42. Greatest Hits (Owens et al 2016)
43. Okutama Action (Barekatain et al 2017)
44. K-400 (Kay et al 2017)
45. AudioSet (Gemmeke et al 2017)
46. Smthng-Smthng (v1/v2) (Goyal et al 2017)
47. TGIF (Jang et al 2017)
48. CMU Panoptic (Joo et al 2017)
49. MultiTHUMOS (Yeung et al 2018)
50. RESOUND (Li et al 2018b)
51. EK-55 (Damen et al 2018)
52. K-600 (Carreira et al 2018)
53. VLOG (Fouhey et al 2018)
54. AVA (Gu et al 2018)
55. TVQA (Lei et al 2018)
56. UCF-Crime (Sultani et al 2018)
57. Charades-Ego (Sigurdsson et al 2018)
58. YouCook2 (Zhou et al 2018b)
59. K-700 (Carreira et al 2019)
60. COIN (Tang et al 2019)
61. AIST (Tsuchida et al 2019)
62. Drive&act (Martin et al 2019)
63. HowTo100m (Miech et al 2019)
64. Moments in Time (Monfort et al 2019)
65. HACS (Zhao et al 2019)
66. VATEX (Wang et al 2019b)
67. IG65M (Ghadiyaram et al 2019)
68. Toyota Smarthome (Dai et al 2022a)
69. OOPS (Epstein et al 2020)
70. AViD (Piergiovanni and Ryoo 2020)
71. VGGSound (Chen et al 2020a)
72. LLP (Tian et al 2020)
73. HVU (Diba et al 2020)
74. DMP Ortega et al (2020)
75. Action Genome (Ji et al 2020)
76. Countix (Dwibedi et al 2020)
77. K-700 (2020) (Smaira et al 2020)
78. FineGym (Shao et al 2020)
79. RareAct (Miech et al 2020a)
80. Broxton et al (2020)
81. HAA500 (Chung et al 2021)
82. IKEA ASM (Ben-Shabat et al 2021)
83. MultiSports (Li et al 2021b)
84. MOMA (Luo et al 2021)
85. Spoken Moments (Monfort et al 2021)
86. WebVid-2M (Bain et al 2021)
87. Home action genome (Rai et al 2021)
88. DanceTrack (Sun et al 2022a)
89. EK-101 (Damen et al 2022)
90. FineAction (Liu et al 2022f)
91. AVSBench (Zhou et al 2022)
92. Li et al (2022d)
93. Assembly101 (Sener et al 2022)
94. Ego4D (Grauman et al 2022)
95. SportsMOT (Cui et al 2023)
96. Ego-Exo-4D (Grauman et al 2024)
97. EPIC-Sounds (Huh et al 2023)
98. MVBench (Li et al 2024a)
99. OVR (Dwibedi et al 2024)
100. Yang et al (2024a)

Fig. 2. **Comparisons of total dataset durations by primary modality**. The size of the circle corresponds to the (approximate) summed duration of all videos in the datasets. Recent datasets (i.e. > 80) not only have multi-fold longer total running times compared to previous datasets but also include additional primary modalities such as language or audio.

tric vision. Similarly, Something-Something (Goyal et al 2017) and Charades (Sigurdsson et al 2016) have been used as benchmarks for object-based actions with a greater focus on temporal information. Datasets for semantic hierarchies include Diving-48 (Li et al 2018b) and FineGym (Shao et al 2020). More recently datasets have also been created for subtasks that are adjacent to action recognition including; instruction learning (Alayrac et al 2016; Bansal et al 2022; Ben-Shabat et al 2021; Ohkawa et al 2023; Sener et al 2022; Tang et al 2019), action phase alignment (Sermanet et al 2017), repeating action counting (Dwibedi et al 2020, 2024; Hu et al 2022a; Runia et al 2018; Zhang et al 2020), action completion prediction (Epstein et al 2020), driver behavior (Martin et al 2019; Ortega et al 2020), anomaly detection (Acsintoae et al 2022; Liu et al 2018b; Lu et al 2013; Sultani et al 2018; Wu et al 2020a), hand-object interactions (Chao et al 2021; Garcia-Hernando et al 2018; Hampali et al 2020; Kwon et al 2021; Moon et al 2020; Mueller et al 2017), and object state changes in actions (Souček et al 2022).

3.2 Task- and modality-specific datasets

Apart from general-purpose datasets, benchmarks have also been proposed for specific aspects of action understanding. We overview three groups of tasks and benchmarks based on the holistic understanding of scenes beyond standard monocular videos and the use of supplementary modalities to video such as language and audio.

Multiview. The human visual system can perceive the world around us in great detail. Usually, this detail comes from head movements that change the viewer's perspective and improve a scene's holistic understanding. Initial efforts compiled multi-view videos from a small number of subjects (Sigal et al 2010) or through synthetic data (Ionescu et al 2013). Capturing high-quality multi-view videos is highly dependent of the hardware and setup. CMU panoptic (Joo et al 2017) captures group interactions with a geodesic dome with 480 VGA cameras. Interactions included social settings, games, dancing, and musical performances. Similarly, ZJU-Mocap (Peng et al 2021) comprises dynamic videos of human motions from 20 cameras. More recently, the Immersive Light Field dataset (Broxton et al 2020) collected light field videos with 6 degrees of freedom with a camera rig consisting of 46 action cameras. Multi-

view datasets have also been collected for varying target tasks 3D video synthesis of human actions and interactions in indoor (Li et al 2022d) and outdoor (Lin et al 2021b; Yoon et al 2020) settings, dance sequence reconstruction (Tsuchida et al 2019), and the dyncamic synthesis of indoor spaces that actions take place (Tschernezki et al 2024).

Video-language. In recent years language has been integrated into vision tasks as a natural extension to represent high-level semantics. Commonly, learning to map textual concepts and visual representations in a shared embedding space has been a widely adopted strategy by many video tasks TODO refs. Initial video-language datasets (Chen and Dolan 2011; Xu et al 2016) were based on short video snippets and short textual descriptions of actions performed over the video. Recent efforts have also provided multilingual descriptions (Wang et al 2019b). Video question-answering is another popular language-based task that has been explored by a large number of works. (Jang et al 2017; Lei et al 2018; Li et al 2024a; Oncescu et al 2021; Xiao et al 2021). For long videos, The sequentiality of instructions has been of great interest with the introduction of benchmarks such as HowTo100M (Miech et al 2019), YouCook2 (Zhou et al 2018b). Benchmarks have also been proposed for other long-form tasks such as moment retrieval (Song et al 2024; Yang et al 2024a), and long-term reasoning (Mangalam et al 2023).

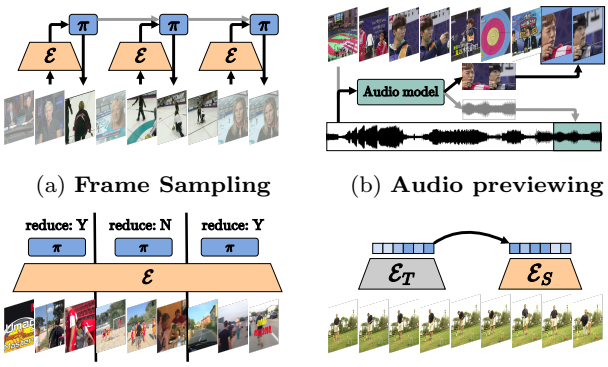
Audio and vision. Human perception has often relied on both vision and audio for perceiving actions, especially in conditions where appearance may lead to ambiguous predictions. Audioset (Gemmeke et al 2017) is the largest audio-visual dataset containing 2.1M clips across a long-tail distribution of 527 classes. VGG-Sound (Chen et al 2020a) is another common benchmark with 200K videos uniformly distributed across 300 classes. Datasets have also been collected for specific tasks such as audio-visual semantic segmentation (Zhou et al 2022), audio-visual video parsing (Tian et al 2020), materials sound and action classification (Huh et al 2023; Owens et al 2016), and video captioning (Monfort et al 2021).

4 From recognition to research tasks

(Wang et al 2023d)s

4.1 Temporal-based tasks

(Albanie et al 2020) (Stergiou and Deligiannis 2023)
Challenges.



(a) **Frame Sampling** (b) **Audio previewing**

(c) **Video input permuting** (d) **Knowledge transfer**

Fig. 3. **Redundancy reduction methods** include; the selection of salient frames for downstream tasks (a), the use of supplementary modalities such as audio to preview relevant regions to sample from (b), permitting the input by reducing the computation overhead of non-relevant frames and segments (c), and using embeddings for a teacher model to optimize representations (d).

Reducing redundancies in recognition. Based on a 2024 study (Sandvine 2024), video accounts for over 5 exabytes of the world's daily internet traffic.

As the efficient processing of videos can be a computational burden, works have focused on reducing redundancies. One of the most common approaches for reducing redundancies is *frame sampling*. Works on frame sampling rely on policy networks over videos selecting frames based on the action's complexity (Ghodrati et al 2021; Yeung et al 2016), correspondence with the video's context (Wu et al 2019c), or changes in the target class' probability (Korbar et al 2019). Wang et al (2021c) used a recurrent network to localize action-relevant regions with further extensions including early stopping (Wang et al 2022b) and features from the entire video to determine the action-relevant patch coordinates (Wang et al 2022c). Xia et al (2022b) classified frames as salient and non-salient by pseudo labels obtained by embedding distance to class centroids. Other approaches have used reward functions based on predictions from the selected frames (Wu et al 2020c), combined frame-level and video-level predictions (Gowda et al 2021), optimized towards balancing accuracy and number of frames used (Wu et al 2019b), or removed tokens in transformer architectures (Wu et al 2024).

A more recent adjacent set of approaches has been based on frame sampling by *previewing audio*. Gao et al (2020) used both frame and audio features with a recurrent network to predict the next informative moment in the video. Similarly, Nugroho et al (2023) used soft pre-

dictions to localize salient audio segments from which the corresponding parts of the video can be selected.

Although sampling can be a beneficial technique in short videos, as the context of videos increases, using a small number of frames can result in information loss. Another line of research thus studies redundancies reduction by *video input permutations*. These works either change frame resolutions based on classifier confidence (Meng et al 2020) or quantize frames at different precision (Abati et al 2023; Sun et al 2021). Zhang et al (2022c) used a two-branch approach for light computations larger context and heavier computations in a smaller context similar to (Feichtenhofer et al 2019).

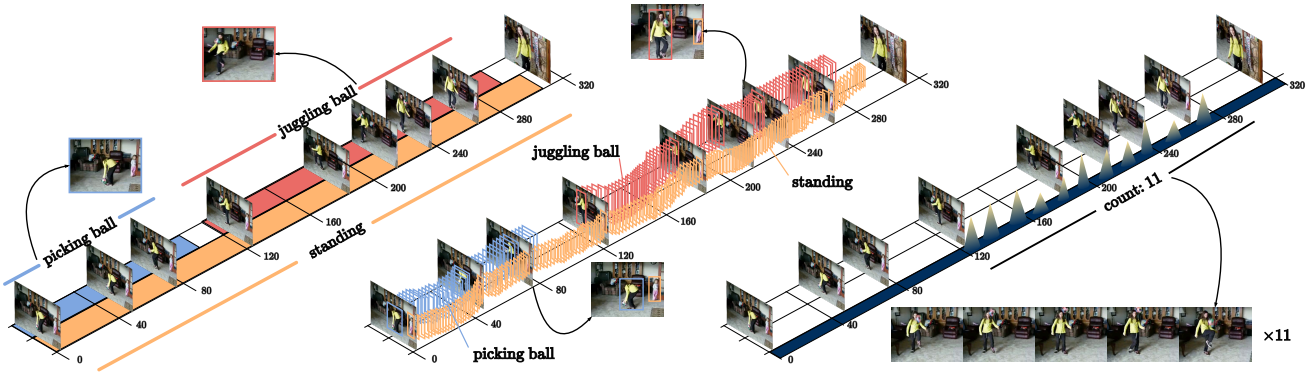
Recent efforts have also used *knowledge distillation* to improve the training efficiency of video model pipelines. Works have focused on learning to match reduced resolution features from the student to full resolution teacher features (Ma et al 2022), or cross-attended between teacher and student features during training (Kim et al 2021b). Distillation approaches have also used teacher models from additional modalities. Lei et al (2021b) bound language embeddings to sparsely sampled clips from long videos. Xia et al (2022a) used embeddings from textual event-object relations to discover salient frames. Tan et al (2023b) proposed a reconstruction approach for interpolating egocentric video features using embeddings from partial frames and the camera motion for the unobserved frames.

Temporal localization. A well-established video task is discovering and classifying temporal segments and the actions performed. Temporal Action Localization (TAL) aims to infer the classification label alongside the start and ends of the corresponding locations in untrimmed videos. Early attempts have used Improved Dense Trajectories (Wang and Schmid 2013) and Fisher Vector (Oneata et al 2013) to model the temporal dynamics of scenes. Shou et al (2016) was one of the first to approach TAL by defining a joint action proposal and classification objective to optimize a regional CNN. This joint optimization has been adapted in subsequent works to use spatial and temporal-only networks (Lin et al 2018; Paul et al 2018; Wang et al 2017a), regional proposal selection (Chao et al 2018; Xu et al 2017b), or intra-proposal relationships with graph convolutions (Zeng et al 2019). Shou et al (2017) predicted granularities at a frame level by transposing the temporal resolution of pre-trained video encoders. More recent approaches for TAL can be categorized into three broad categories.

Most similar to the aforementioned methods, **one-stage** approaches, localize and classify actions in a single step using hierarchical embeddings from feature pyramids (Lin et al 2021a; Liu and Wang 2020; Shi et al 2023; Zhang et al 2022b) or discovering segments by relating relevant videos (Shou et al 2018; Yang et al 2020b). Recently, Yan et al (2023) has also introduced the use of vision-language encoders for encoding and using the scene's semantic context for TAL. As the definition of specific start and end frames for actions can often be ambiguous, a set of approaches have relaxed their objectives to weigh the training loss by the importance of each frame within the action segment (Shao et al 2023) or learning distributions of possible start and end times (Moltisanti et al 2019).

In a different line of research, a group of methods use additional steps to regress temporal boundaries. One line of **two-stage** approaches disentangles the optimization into separate parts. Zhai et al (2020) combined proposals from spatial- and temporal-only streams into a fused final prediction. Chen et al (2022a) used long- and short-range temporal information to refine the confidence of the generated action proposals. Huang et al (2019) decoupled classification and localization to two objectives during training with two separate models that use cross-modal connections for exchanging information. Approaches have also aimed to maximize the embedding difference between representations of frames from action segments and non-relevant frames either with positive and negative instances (Luo et al 2020; Zhang et al 2021a) or by a scoring function (Rizve et al 2023). Methods have also improved the features of backbone encoders with more TAL-relevant pretext tasks (Zhang et al 2022a). Most two-step methods however process videos holistically (Alwassel et al 2021; He et al 2022a; Liu et al 2021c; Qing et al 2021). Alwassel et al (2021) encode local features over a sliding window. Use aggregated features from all local encoders to find the regions of action while classifying every segment. Graphs have also been adopted for TAL with Bai et al (2020) using generated candidate proposals for the graph's start/end edges and its connected nodes. Zhao et al (2021) created a graph of hierarchical features with windows over multiple temporal resolutions. More recent approaches (Nag et al 2023) have also formulated proposal prediction as a denoising task with noisy action proposals as input to a diffusion model conditioned on the video.

A recent set of methods has been based on adapting **DEtection TRansformers DETR** (Carion et al 2020) to TAL. DETR-based approaches rely on transformer encoder-decoders to create regional proposals that can be in turn optimized with bipartite matching. Tan et al (2021) adapted DETR with a matching scheme of multiple positive action proposals to address sparsity in temporal annotations. Works that aimed towards optimization have included dense residual con-



(a) TAL (b) STAD (c) VRC

Fig. 4. **Objective visualization for Temporal Action Localization (TAL), Spatio-Temporal Action Detection (STAD), Video Repetition Counting (VRC)**. TAL (a) discovers the start and end times of individual actions. In contrast, STAD (b) is more complex as it requires temporally and spatially localizing actions with bounding boxes for actors and objects over time. Distinctively, VRC (c) is not based on action labels and instead requires counting repetitions of actions or motions in an open-set setting. Video source from (Kay et al 2017).

nections (Zhao et al 2023) or caching short-term features (Cheng and Bertasius 2022; Hong et al 2022a). Liu et al (2024b) have scaled model capacity upwards by training intermediate adaptors to propagate information to the decoder from intermediate frozen encoder layers. Approaches have also explored training recipes with sparsely updating model layers (Cheng et al 2022), vision-language pre-training distillation (Ju et al 2023), proposal hierarchies (Wu et al 2023a), and end-to-end TAL encoder-decoder optimization (Liu et al 2022e).

Spatiotemporal detection. Adding to TAL, SpatioTemporal Action Detection (STAD) is more complex as it localizes the temporal extend of actions and detects action-relevant actors and objects over frames. Maintaining consistency between the per-frame detections and the temporal action proposals in the main challenge of STAD methods. Similar to the research directions for localization, two categories can be used to overview relevant literature.

Bulding upon the advancements of image-based object detectors (Girshick et al 2014; Girshick 2015), the majority of approaches perform STAD in **two stages** by first detecting objects and then temporally localizing action by tracking object candidates (Jain et al 2014; Weinzaepfel et al 2015), ROI-pooling rgb and flow features (Peng and Schmid 2016), refining proposals iteratively (Soomro et al 2015), aligning source and target domain features (Agarwal et al 2020), or using the general action level in the video as context (Mettes et al 2016). Li et al (2018a) build upon prior two-stage detection with the incorporation of recurrent proposals to include temporal context. Other refining approaches (Singh et al 2017) used the arrow of time with different portions of the video detected at each step. The introduction of benchmarks with videos of higher temporal resolutions (Gu et al 2018) also led approaches towards including additional context over longer frame sequences with future banks (Feng et al 2021b; Pan et al 2021; Tang et al 2020a; Wang and Gupta 2018; Wu et al 2019a, 2022b) or additional objects (Arnab et al 2021b; Hou et al 2017; Zhang et al 2019c) which has showed further improvements in detection capabilities of models. Additional information such as keyframe saliency maps (Li et al 2020b; Ulutan et al 2020), hands and poses (Faure et al 2023), actor-object relations (Sun et al 2018), self-supervision (Wang et al 2023c) have also been explored. Alwassel et al (2018) analyzed the advancements of two-stage approaches beyond performance metrics showing that improvements are centered towards handling temporal context. However, features used are pre-computed from large backbones requiring auxiliary task-specific models.

Drawing inspiration from **single-stage** object detection methods (Carion et al 2020; Redmon et al 2016; Liu et al 2016), STAD single-stage approaches use a unified framework for both localization and detection (Chen et al 2021; Girdhar et al 2019; Zhu et al 2024). Ntinou et al (2024) extended the bipartite matching loss from Carion et al (2020) to spatio-temporal tokens. Other approaches also used adaptive feature sampling (Wu et al 2023c), conditioning modeled visual features based on motion (Zhao and Snoek 2019), or have contrasted

different views in training (Kumar and Rawat 2022). Directly predicting tubelets has also been adopted as as a recent direction by recent approaches (Gritsenko et al 2024; Kalogeiton et al 2017; Song et al 2019; Yang et al 2019; Zhao et al 2022). Kalogeiton et al (2017) stacked embeddings from a backbone used in a sliding window to regress classes and tubelets over the entire video. Zhao et al (2022) used an encoder-decoder to generate tubelet queries and cross-attended them to visual features. Gritsenko et al (2024) generated a candidate tubelet based on a condensed query representation used to cross-attend with features at each frames. Beyon STAD, tubelets have also been used are a self-similarity pre-training objective (Thoker et al 2023) to enforce correspondence of videos from different domains but with similar local motions.

Repetition counting. Relating to TAL and STAD, Video Repetition Counting (VRC) approaches aim to count the number of action repetitions. In contrast to TAL/STAD however, VRC is an open-set task and does not require temporally localizing actions. Early works relied on the periodicity of signals (Thangali and Sclaroff 2005) decomposing the repetitions of signals with a Fourier analysis (Albu et al 2008; Briassouli and Ahuja 2007; Azy and Ahuja 2008; Cutler and Davis 2000; Pogalin et al 2008). Signal-based works have used the flow directions over time (Runia et al 2018). A number of approaches have defined VRC as a classification task. Lu and Ferrier (2004) used dynamic parameters based on the Frobenius norm to classify changes corresponding to action end times. Zhang et al (2021c) fused audio and video representations while Zhang et al (2020) used multiple cycles to refine the repetition count prediction. Li et al (2024b) extracted action query features and classified the queries for repeating actions. In contrast to defining a classification task with a predefined number of possible repetitions, Dwibedi et al (2020) adopted a temporal self-similarity matrix (Ben-Abdelkader et al 2004; Junejo et al 2010; Körner and Denzler 2013) to discover the periodicity of repetitions. Subsequent methods have used embedding similarity matrices with multiple scales (Bacharidis and Argyros 2023; Hu et al 2022a), triplet contrastive losses (Destro and Gygli 2024), and graph representations (Panagiotakis et al 2018). As representations can be highly similar for adjacent frames, several recent methods have instead aimed to limit the discovery of correspondences in repetition instances to poses (Ferreira et al 2021; Yao et al 2023), specific frames (Li and Xu 2024; Zhao et al 2024b), visual exemplars (Sinha et al 2024), or language descriptions (Dwibedi et al 2024).

Future outlooks.

4.2 Language semantics in videos

(Song et al 2024) (Yu et al 2017) (Anderson et al 2018) (Xu et al 2021) (Ashutosh et al 2023) (Kahatapitiya et al 2024) (Li et al 2023c) (Jiang et al 2023) (Bain et al 2021) (Fu et al 2021) (Han et al 2022b) (Ko et al 2022) (Lei et al 2021b) (Li et al 2022a) (Li et al 2020a) (Miech et al 2020b) (Seo et al 2022) (Seo et al 2021) (Sun et al 2019a) (Wang et al 2023b) (Wang et al 2022a) (Zellers et al 2021) (Lei et al 2021a) (Lin et al 2022) (Wang et al 2022d) (Zhu and Yang 2020)

(Zhao et al 2024a) (Cheng et al 2024) (Wang et al 2024a) (Kuo et al 2023)

Video Question Answering (VideoQA) (Xue et al 2023) (Min et al 2024) (Yang et al 2022c) (Yang et al 2021) (Yang et al 2022a) (Xiao et al 2021) (Xu et al 2017a) (Zhang et al 2023a) (Yu et al 2023b) (Xiao et al 2024) (Gao et al 2023) (Xiao et al 2022) (Xiao et al 2023) (Li et al 2022e) (Li et al 2023d) (Jiang and Han 2020) (Park et al 2021a) (Guo et al 2021) (Jang et al 2017) (Fan et al 2019) (Gao et al 2018) (Huang et al 2020a) (Li et al 2019) (Liu et al 2021b) (Dang et al 2021) (Cherian et al 2022) (Geng et al 2021) (Ye et al 2017) (Zeng et al 2017)

Video Captioning ✔ (Alayrac et al 2022) (Krishna et al 2017) (Yang et al 2023a) (Chen et al 2024) (Ren et al 2024) (Zhou et al 2024) (Islam et al 2024) (Mavroudi et al 2023) (Iashin and Rahtu 2020a) (Iashin and Rahtu 2020b) (Wang et al 2018a) (Wang et al 2020c) (Chen and Jiang 2021) (Deng et al 2021) (Mun et al 2019) (Rahman et al 2019) (Shen et al 2017) (Shi et al 2019) (Wang et al 2021b) (Zhou et al 2018c) (Han et al 2023b) (Han et al 2023a) (Han et al 2024) (Seo et al 2022) captioning multitask (Chadha et al 2021) (Li et al 2018c)

Video Retrieval ✔ (Gordo and Larlus 2017) (Wang et al 2016a) (Xu et al 2015a) (Torabi et al 2016) (Dong et al 2018) (Otani et al 2016) (Kim et al 2024b) (Wray et al 2019) (Wray et al 2021) (Ge et al 2022b) (Xue et al 2022) (Gabeur et al 2020) (Mithun et al 2018) (Liu et al 2019) temporal grounding (Anne Hendricks et al 2017) (Gao et al 2017a) (Regneri et al 2013) (Qian et al 2024) (Gu et al 2024a) (Yang et al 2022b) (Escorcia et al 2019) (Flanagan et al 2023) (Cao et al 2021) (Chen et al 2018a) (Ge et al 2019) (Jiang et al 2019a) (Liu et al 2021a) (Liu et al 2018a) (Qu et al 2020) (Wang et al 2020b) (Xu et al 2019b) (Chen et al 2020b) (Hao et al 2022) (Liu et al 2022a) (Nan et al 2021) (Yuan et al 2019) (Zhang et al 2021b)

4.3 Audio-visual and multimodal recognition

The recognition of actions or activities has been primarily studied in the vision domain. Modeling distin-

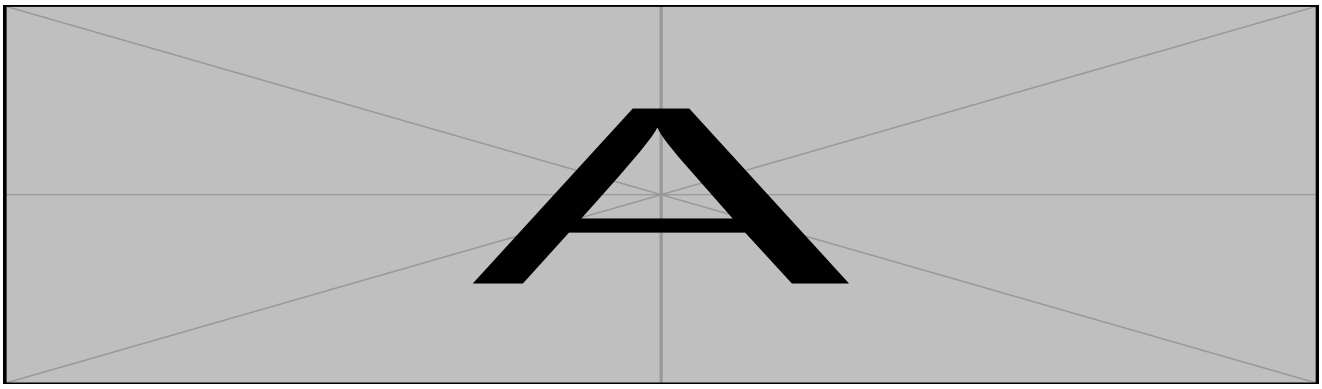


Fig. 5. TODO: Table for grouped language approaches.

guishable appearance- or motion-based characteristics corresponding to actions has been the primary motivation for vision-based action understanding methods. Instead, the auditorily recognition of actions focuses on the sounds emitted by objects, their interactions, or actions. This can include distinct challenges as the sounds emitted by different objects or actions can be similar. Time-frequency spectrograms have been a popular format for the representation of audio events in videos. Initial audio-based models have been built following image-based object recognition (Gong et al 2021) or video classification (Kazakos et al 2021) CNNs. Attending over audio patches has been either integrated with convolutions (Gulati et al 2020; Kong et al 2020) or employed image-pre-trained models (Koutini et al 2022). Approaches have also explored the effects of patch masking (Baade et al 2022; Huang et al 2022), focusing on salient sounds (Stergiou and Damen 2023a), adapting (Liu et al 2022b) or compressing (Feng et al 2024) spectrogram resolutions. More recently, the use of audio has gained attention in multi-modal learning settings as it can provide supplementary information to both visual features and language context.

Audio-visual models. In developmental psychology, relationships between visual and auditory understanding of the environment are developed at a young age (Morrongiello et al 1998). This has motivated early efforts in active speaker recognition (Chen and Rao 1998; Matthews et al 2002) and person identification (Aleksic and Katsaggelos 2006) to study vision and audio cues in tandem. As video and audio signal types differ significantly, works have used two-step models to infer predictions. Two-step approaches extract video and audio embeddings first and then either fuse modality-specific predictions (Fayek and Kumar 2020), embeddings from multiple paths from multiple modalities (Xiao et al 2020), or jointly attending vision and audio features for the final prediction (Gong et al 2022b). More recently, architectures have tokenized and attended audio and vision jointly with multi-modal learnable tokens (Nagrani et al 2021), cross-modal attention (Jaegle et al 2021), and modality-gating (Xue and Marculescu 2023). Accounting for models trained on uni-modal tasks, Lin et al (2023) also proposed using cross-modal adaptors to combine architectures in multi-modal tasks. Exploring the relevant audio and visual features with self-supervision has also been a learning paradigm of significant research interest. The use of a common embedding spaced can be useful for discovering correspondences in both in-modal and cross-modal retrieval (Arandjelovic and Zisserman 2018; Wu and Yang 2021), multi-modal clustering (Hu et al 2019), and sound source separation (Hu et al 2022b; Mo and Morgado 2023; Zhao et al 2018). Token reconstruction through masking has also been a popular self-supervision pre-training task with a variety of training schemes including; concatenating masked tokens (Gong et al 2023), multi-view masking per modality (Huang et al 2023), fusing a mixture of per modality masked tokens (Guo et al 2024b), combining modality-specific masked and unmasked embeddings (Georgescu et al 2023), or using multiple masking ratios with siamese networks (Lin and Bertasius 2024). Variation in the relevance of either visual and auditory signals depending on the instance. A promising direction for integrating this into optimization is Gradient blending (Wang et al 2020d) which recalibrates the loss per modality. Other research works have explored multi-audio to single-visual scene correspondence with contrastive learning. This has been done by utilizing joint semantic similarity in both modalities (Morgado et al 2021), using active sampling to diversify the pool of negative samples (Ma et al 2021), or by counterfactual audio and video pairs to enforce a relationship between multiple audio to single visual scenes (Singh et al 2024). Enforcing a similarity constraint between audio

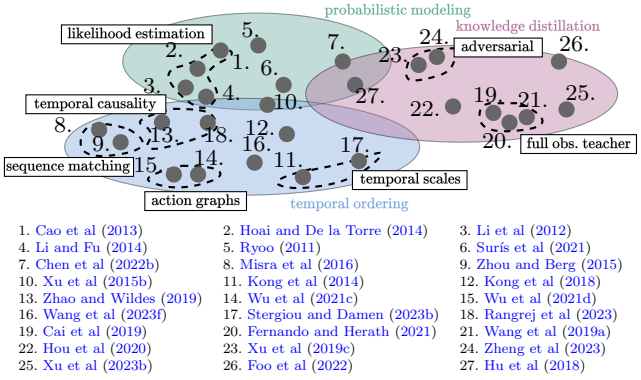


Fig. 6. **Early Action Prediction methods** clustered by research approach. The three main clusters are denoted with blue, teal, and purple. Smaller groups are shown with dashed lines. Positioning of works represents an abstract proximity of the research idea to seminal works.

and vision steams can also be used to train incremental tasks (Pian et al 2023).

Multi-modal models (Akbari et al 2021) (Kaiser et al 2017) (Radevski et al 2023) (Srivastava and Sharma 2024a) (Zhang et al 2024a) (Munro and Damen 2020) (Recasens et al 2023) (Dai et al 2022b) (Zellers et al 2022) (Srivastava and Sharma 2024b)

4.4 Interaction recognition

Dyadic human-human interactions (Nguyen et al 2024) (Ong et al 2023)

Group interactions

Human-object interactions

5 Predictions in ongoing actions

(Huang et al 2018)

5.1 Early action prediction

Early Action Prediction (EAP) assumes an *ongoing* action is being performed with predictions made based on the observable part of the action $\tau_{1,\rho}$.

Challenges.

Probabilistic modeling. A large portion of the EAP literature has originally been based on probabilistic modeling of action classification from partial observations (Cao et al 2013; Hoai and De la Torre 2014; Li et al 2012; Li and Fu 2014; Ryoo 2011). Ryoo (2011) used a bag of words based on feature distributions. This division into segments has been relevant in subsequent probabilistic approaches that used sparse coding (Cao et al 2013), max-margin (Hoai and De la Torre 2014), or scoring functions (Li et al 2012; Li and Fu 2014) to infer the action likelihood. More recent probabilistic approaches have included the use of hyperbolic representations (Surís et al 2021) for hierarchical predictions of actions. Future prediction ambiguousness has also been explored as the generation and subsequent selection of multiple future representations (Chen et al 2022b).

Temporal ordering. A different line of works has explored EAP based on the temporal evolution of the action. The arrow of time (Pickup et al 2014) can provide a strong signal to associate the procedural understanding of actions with high-level categorical semantics (Misra et al 2016; Zhou and Berg 2015). Xu et al (2015b) formulated EAP with an auto-completion objective matching candidate futures to a partial action observation query. The predictability of partial observations can be difficult in instances where there are visual similarities in the performance of actions. To address this approaches have either used multiple temporal scales (Kong et al 2014), created key-value memories of representations (Kong et al 2018), or propagated the residuals of features' residuals over time (Zhao and Wildes 2019). More recently approaches have also used temporal graph representations (Wu et al 2021c,d), contrastive learning over partial observations of the same action (Wang et al 2023f), aggregated attention over temporal scales (Stergiou and Damen 2023b), or attending over relevant space-time regions (Rangrej et al 2023).

Knowledge distillation from full observations. Transferring class knowledge (Park et al 2019) from models trained on the full videos can be an effective technique for refining predictions from partial observations. Cai et al (2019) and later Fernando and Herath (2021), and Wang et al (2019a), used learned representations of the full observations as the target representations for partial observations. Further methods (Hou et al 2020) have refined this with sequentiality of motions to learn soft targets and regress model predictions. In a similar effort, (Xu et al 2019c) and (Zheng et al 2023) integrated an adversarial objective for generating representations for the non-observable parts. Similarly, Xu et al (2023b) learned to reconstruct representations of full observations with a masked autoencoder (He et al 2022b). Other works have fine-tuned expert heads for each action category (Foo et al 2022) or learned by focusing on videos with distinct visual features (Hu et al 2018).

Future outlook.

5.2 Frame-level prediction

Adjacent to EAP, Video Frame Prediction (VFP) aims to reconstruct future frames of ongoing actions using a partially observed action $\tau_{1,\rho}$. Although the high-level semantics such as action levels are not learned as in EAP, VFP still requires to relate the sequentially of motions and likely intended action, to the reconstruction of subsequent frames.

Challenges

Sequential adversarial predictions. A significant portion of VFP works have been based on sequential frame generation (Castrejon et al 2019; Chaabane et al 2020; Chang et al 2021, 2022; Chen et al 2017; Guen and Thome 2020; Hwang et al 2019; Jin et al 2020; Liang et al 2017; Villegas et al 2018; Wang et al 2018e; Wu et al 2021b). These approaches use recursion to generate representations or predictions autoregressively. A line of methods (Chen et al 2017; Jin et al 2017) focused on the correspondence of objects between frames to guide the generation of the next frames. Castrejon et al (2019) used similar adversarial guidance by fusing context information from previous frames. Additional supervisory signals included motion flow (Liang et al 2017), partial differential equations (Guen and Thome 2020), and embeddings over multiple temporal resolutions (Gao et al 2022). Another line of approaches (Chang et al 2021; Villegas et al 2018; Wang et al 2018e) has also included long-term memory connections to discover causalities from frames over greater temporal resolutions. Park et al (2021b) aimed at incorporating time dynamics for VFP with the inclusion of ordinary differentiable equations (ODE). Davtyan et al (2023) used ODE with the previous frame as the initial condition and integrate the vector field from Flow Matching (Lipman et al 2022) to predict the next frame.

Parallel multi-frame synthesis. In contrast to the sequential reconstruction of future frames, approaches have also generated multiple future frames in a single step. One of the first efforts for multi-frame prediction (Liu et al 2017b) used a multi-frame per-pixel optical flow vector with further adaptations including multiple scales (Hu et al 2023). Attention-based architectures have also been used for parallelization of frame prediction with works that introduce encodings of context for frame prediction (Ye and Bilodeau 2023), attend over temporal patches (Tan et al 2023a), condition the generation based on short-term representation variations (Hu et al 2023; Smith et al 2024), multiple motion and appearance scales (Zhong et al 2023a), and reduce inference speeds (Ye and Bilodeau 2022; Tang et al 2024). An extension to spatio-temporal attention, Nie et al (2024) used a triplet module to attend across all dimensions of the video representations sequentially.

Probabilistic generation. Another group of approaches have studied the reconstruction of future frames with probabilistic approaches. Babaeizadeh et al (2018) and Denton and Fergus (2018) used a probabilistic variational model on the stochasticity of the video to generate frame predictions. Wang et al (2020e) models the perceptual uncertainty in future frames with a Bayesian framework to weigh future prediction candidates. Diffusion-based models (Dhariwal and Nichol 2021; Ho et al 2020; Rombach et al 2022) have been applied to a multitude of generative tasks with their adaptation to VFP by a set of approaches (Gu et al 2024b; Höppe et al 2024; Shrivastava and Shrivastava 2024; Voleti et al 2022; Ye and Bilodeau 2024; Zhang et al 2024b). They gradually transform a complex distribution into unstructured noise and learn to recover the original distribution from noise at generation.

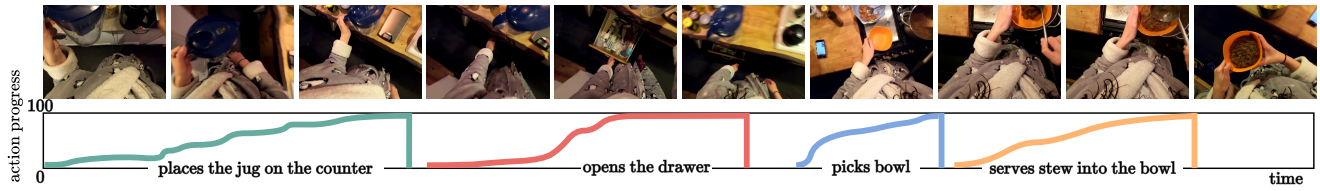
Future outlook.

5.3 State changes

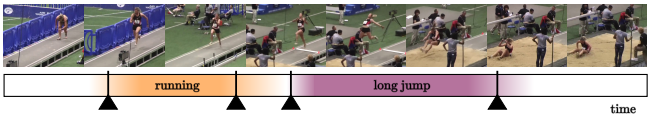
Another set of action prediction tasks includes modeling the state changes in the environment, actions, objects, and execution speeds. It can also involve inferring the reasoning for these changes. An overview of the tasks' objectives is visualized in Figure 7.

Challenges

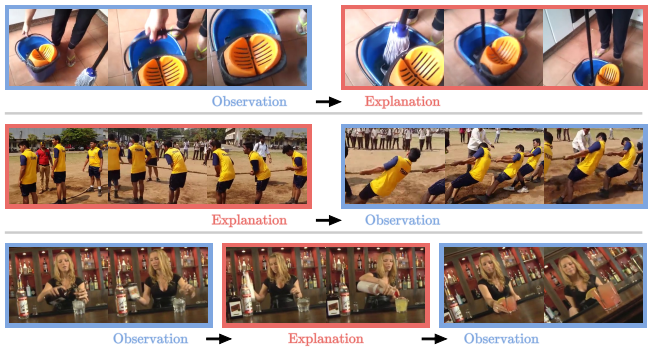
Action progress. Actions can be understood by procedural sets of motions performed towards an intended goal. Vaina and Jaulent (Vaina and Jaulent 1991) have suggested that understanding the state and progress of the action at different times can provide a holistic understanding of the intent and objective. In machine vision, an initial approach for Action Progress Prediction (APP) (Fathi and Rehg 2013) used local descriptions to model per-frame state changes. (Kataoka et al 2016) used a descriptor to discover transitional actions within activity sequences. Xiong et al (2017) introduced a score function to distinguish actions based on learned distinctive parts. Becattini (Becattini et al 2020) used actor and scene context information as an additional supervisory signal for APP. Price et al (2022) expressed the progress of multiple actions through threads of activities that in long procedural videos can also overlap. Shen and Elhamifar (Shen and Elhamifar 2024) causally attended videos to define a task graph for APP over each action. More recently generative approaches (Souček et al 2024) using conditional control (Zhang et al 2023c) and procedural knowledge (Ashutosh et al 2024; Zhou et al 2023) have also been used to generate keyframes of changes.



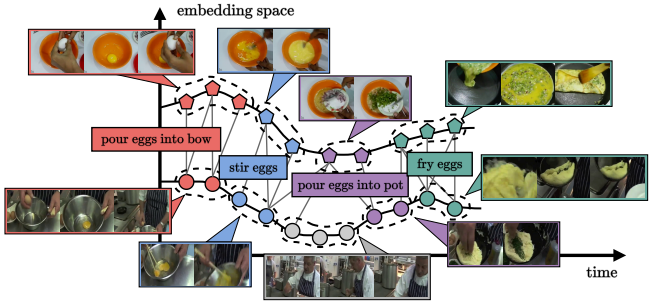
(a) **Action Process Prediction (APP)**. Given a video stream of a procedural task, estimate the progress of each ongoing action by inferring the time it will take to complete the action performed. Video sourced from (Grauman et al 2024)



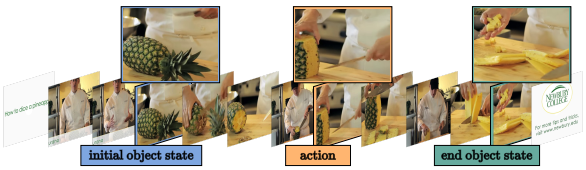
(b) **Event Boundary Detection (EBD)**. Detect the start and end times of ongoing events in video streams. Video sourced from (Carreira and Zisserman 2017)



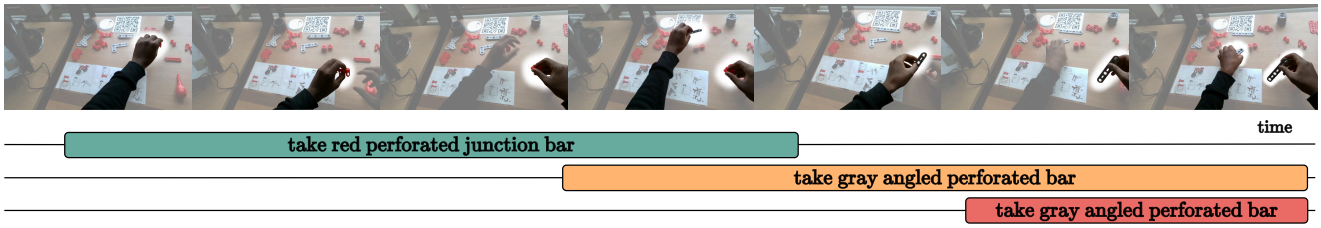
(d) **Visual Abductive Reasoning (VAR)**. Given the observable part of the video in blue. Infer a likely explanation in red for what follows before, after, or during the observation. The task requires a high-level understanding of the action or activity performed. Video sourced from (Liang et al 2022)



(d) **Video Alignment (VA)**. Find correspondences across video instances with the same action performed and align them so the execution of the action is synchronized. Video sourced from (Tang et al 2019)



(e) **Object State Change Detection (OSCD)**. State modifying actions such as *cutting*, progressively change the visual appearance of objects from an initial state in blue to a final post-action execution state in teal. OSCD localizes the times that these changes occur. Video sourced from (Souček et al 2022).



(f) **Active Object Detection (AOD)**. Given a video in which a person interacts with multiple objects, AOD aims to detect the object the person is currently using. Video sourced from (Ragusa et al 2021).

Fig. 7. **Tasks relating to object and action state change**. Each of the presented sets of tasks can involve further specific objectives that correspond and address different aspects.

Another line of research works (Heidarivincheh et al 2018, 2016) has also aimed to localize the moments that actions are completed to specific frames. The speed of completion or state changes in actions has also been studied in the context of skill determination (Doughty et al 2018) or their semantic correspondence to textual adverbs (Doughty et al 2020; Doughty and Snoek 2022; Moltisanti et al 2023). Scoring approaches (Tang et al 2020b) have been used to study the procedural execution of action in the context of quality assessment. Adjacent tasks such as video captioning and action classification have also been integrated in multi-task settings (Parmar and Morris 2019).

Event boundary detection. Different from the adjacent well-studied task of action localization, Event Boundary Detection (EBD) (Shou et al 2021) aims at localizing event changes in videos *regardless of the action classes*. Aakur and Sarkar (Aakur and Sarkar 2019) proposed a self-suprvised objective in which their model is initially trained to reconstruct subsequently observed features. (Shou et al 2021) used a self-similarity metric to determine event boundaries by relating encoded frame features. Further EBD approaches (Mounir et al 2024) have also studied hierarchies of the video events.

Video alignment. As the performance of individual parts of actions can vary, Video Alignment (VA) methods aim to temporally match key moments in the execution of the same action across videos. Initial efforts, motivated by temporal coherence (Goroshin et al 2015; Fernando et al 2017; Zhang et al 2023b), have studied VA through approaches based on Canonical Correlation Analysis (CCA) (Andrew et al 2013) or contrastively creating joint representations from multiple viewpoints (Sermanet et al 2018). Dynamic Time Wraping (Sakoe and Chiba 1978) is an algorithm that aligns variable length signals which has also been adopted for VA (Chang et al 2019; Hadji et al 2021; Dvornik et al 2021). A more recent self-supervision objective (Dwibedi et al 2018) for VA is to train a video model to project per-frame embeddings in pairs of target videos by matching embeddings of one video to the nearest neighbor embeddings of the other. This approach was extended in subsequent works by including context from the entire video (Haresh et al 2021), creating anchor frames to align redundant frames (Liu et al 2022d), leveraging embeddings from text (Epstein et al 2021), and regularizing the correspondence to repetitions of the same action (Donahue and Elhamifar 2024).

Abductive reasoning. A key element in action understanding is the intended goal. High-level reasoning of events has initially been considered in hierarchies with approaches using rule-based (Hakeem and Shah 2004) or conditionality of action occurrence (Albanese et al 2010) at each hierarchy. Pei et al (2011) detected atomic actions with graph representations to decompose complex events. Visual Abductive Reasoning (VAR) (Liang et al 2022) is the vision-language task that uses characteristics of partial observations as a premise and requires formulating an explanation. VAR Models have been based on modeling intention with contrastively learning visual and language context (Li et al 2023b), modeling timelines for news story understanding (Liu et al 2023), or forecasting actions with multi-modal inputs (Zhu et al 2023). Evaluation of VAR models has also been studied in counterfactual vision-language pairs (Park et al 2022) similar to text-only tasks (Ippolito et al 2019; Huang et al 2020b).

Object state change. Actions such as 'whisking eggs', 'filling cup', or 'assembling legos' can often alter the appearance or state of objects. Object State Change Detection (OSCD) approaches associate the visual changes with changes in the states of objects in the scene. Efforts (Alayrac et al 2017; Liu et al 2017a; Zhuo et al 2019) have initially focused on state modifications that do not involve significant appearance changes, e.g; 'open/close door' or 'fill/empty cup'. Hong et al (2021) proposed a reasoning-based approach defining a triplet of complexities for single- and multi-step transformations and multi-step transformations with additional viewpoint changes. Other reasoning-based approaches include the use of language (Xue et al 2024) and visual exemplars of start and end states (Souček et al 2022). OSCD has also been studied in combination with other tasks including cross-state object segmentation (Yu et al 2023a), cross-action relevance (Alayrac et al 2024), or inspired by state-disentanglement for images (Gouidis et al 2023; Nagarajan and Grauman 2018; Saini et al 2022), generating start and end states by given context and scene prompts (Saini et al 2023).

Active object. Actions can include multiple objects during their execution. Active Object Detection (AOD) specifies the objects relevant to the currently performed atomic action. This task has recently gained interest as scenes can often be cluttered (Ragusa et al 2021) or a varying number of objects can be used for a single action (Miech et al 2019). Nagarajan et al (2019) specifically focused on localizing the human-object interaction areas to define focal points of importance during the execution of actions. (Fu et al 2022) defined a voting module over potential bounding boxes corresponding to the active object. Kim et al (2021a) used a parallelized model for separately detecting instances and in turn detecting hand-object interaction. Yang and Liu (Yang and Liu 2024) used scene context from text to define plausible interactions with target objects for AOD.

Future outlooks.

5.4 Anomaly detection

Close-set. Anomalies can be discovered by *close-set* tasks that aim to model both normal and abnormal sequences. Sultani et al (2018) used Multiple Instance Ranking (MIL) (Dietterich et al 1997) to define positive groups that include videos with at least a single abnormal segment and negative groups of normal videos. The objective in turn compared the maximum anomaly score between the assigned positive and negative groups. Subsequent efforts (Dubey et al 2019;

Zhang et al 2019b; Zhu and Newsam 2019; Feng et al 2021a; Tian et al 2021; AlMarri et al 2024) have built upon MIL with learned features (Dubey et al 2019), or pseudo labels (Feng et al 2021a). With MIL being influenced by the dominant negative instances, (Zhang et al 2019b) proposed inner-group sampling, (Pu et al 2023; Zhu and Newsam 2019) used temporal weighting, and (Tian et al 2021) aim to maximize the separability between normal and anomalous representations. The use of multiple temporal pretext tasks (AlMarri et al 2024; Georgescu et al 2021) and temporal scales (Li et al 2022b) has also been explored. (Chen et al 2023) used a contrastive objective between representations of normal and abnormal videos. Clustering approaches focus on sparsity modeling (Lu et al 2013), enforcing high distribution variance in abnormalities (Li et al 2021a), combining dense/spare clusters for normal/abnormal segments (Zaheer et al 2020a), and using pseudo labels for anomalous segements (Zaheer et al 2020b). Another set of methods (Zhong et al 2019; Purwanto et al 2021) has included graph networks to sequentially detect abnormal segments. More recent methods have distinguished between states with the use of additional modalities such as audio (Wu et al 2020a) and language context (Yang et al 2024b; Zanella et al 2024).

Open-set. As the close-set solutions can only model abnormalities in labeled data, models cannot effectively generalize to distributions different than those they are trained in. This has been studied by Zhao et al (2011) and later Luo et al (2017) as a sparse-coding (Lee et al 2006) problem in which the model is trained to reconstruct normal sequences. Abnormalities can then be inferred through the reconstruction loss. Temporal regularity can also be modeled with autoencoders as a reconstruction task (Hasan et al 2016). Further AE-based approaches also used pseudo representations to improve abnormal sequence scarcity in the embedding space (Astrid et al 2021b,a) or used prototypes of normal sequences (Park et al 2020). Other works have constrained the representation space of normal sequence through optimizing piecewise linear decision boundaries (Wang and Cherian 2019). Two-steam AE frameworks (Cho et al 2022; Nguyen and Meunier 2019) have also been used to separately reconstruct appearance and motion characteristics of normal sequences. Generative approaches (Micorek et al 2024) have also focused on the latent space with Gaussian mixture model and inferred an anomaly score across all noise levels.

As part of the video anomaly detection requires (Fioresi et al 2023)

trajectory-based video anomaly detection methods (Markovitz et al 2020; Morais et al 2019; Flaborea et al 2023; Stergiou et al 2024).

6 Future forecasting

6.1 Action anticipation

Action Anticipation (AA) requires recognizing current action(s) performed at τ_1 to forecast a *proceeding action* τ_2. In contrast to the partial observations for EAP, anticipation tasks only rely on the expected sequence with which actions can be performed. Early works (Kitani et al 2012; Kuehne et al 2014; Koppula and Saxena 2015) have used graphs to model this sequentiality of actions over time. However, given the limited long-range dependency of graph-based approaches, works have explored either the progress in the execution of actions (Abu Farha et al 2018; Furnari and Farinella 2019; Ke et al 2019), the motion transition intensity between actions (Huang and Kitani 2014), gaze and hand information (Shen et al 2018), or used future-action objective with multiple predictions (Furnari et al 2018; Zatsarynna et al 2024). Despite the diversity of the approaches, some challenges remain present throughout the proposed methods.

Challenges

Embedding similarity maximization. The representations of future actions can be used as a target for learned embeddings. A large number of methods in the literature have thus used future embedding reconstruction tasks to in turn infer future action labels. Gao et al (2017b) used a recurrent decoder to regress future embeddings with an additional policy for class predictions over time. Interaction between objects and actors (Sun et al 2019b; Luc et al 2018) has also been explored by early attempts. Subsequent methods aimed to either maximize the similarity between future and current embeddings through memory banks (Liu and Lam 2022), optimize latents representing intended goals (Roy and Fernando 2022), prototypes (Diko et al 2024), or adversarial representations (Gammulle et al 2019). Other generative approaches use pose information as priors (Villegas et al 2017) or focus on the extrapolation of activity trajectories (Chi et al 2023). Autoregressive approaches have recently shown great promise using either contrastive objectives (Wu et al 2020b), causal attention (Girdhar and Grauman 2021), or audio-visual inputs (Zhong et al 2023c). As future predictions depend on the usefulness of current observations, works have also integrated uncertainty terms in their predictions. Vondrick et al (2016a) regressed towards multiple plausible future embeddings, Abdelsalam et al (2023) grounding the sequentiality of visual embeddings to language, while Guo et al (2024a) defined probabilistic transformer outputs through a top-k prediction loss similar to (Furnari et al 2018).

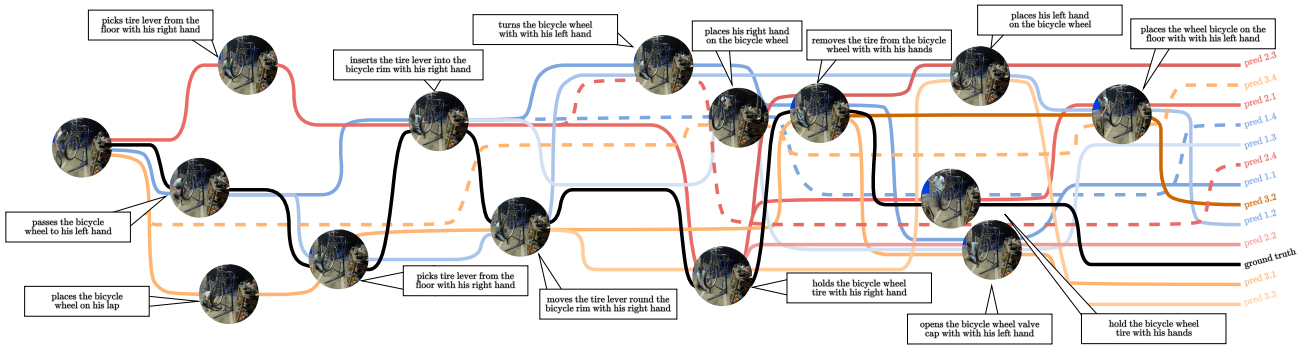


Fig. 8. **Visualization of forecasted future actions**. Starting from the observed action, anticipation approaches infer the sequence of probable next actions. Increases in the number of future actions to anticipate also relate to the number of possible scenarios. Predictions are shown in a narrative chart format similar to (Randall 2009). Example selected from (Grauman et al 2024).

Long-term anticipation. The anticipation of the future can also be extended to forecasting multiple upcoming actions over a longer temporal duration. Bokhari and Kitani (2017) used a q-learning framework with reward functions for the activity label, and locations where actions are performed. Nawhal et al (2022) used a two-stage approach to first infer potential labels and use their logits alongside visual features to predict future action segments. Similarly, Gong et al (2022a) used learnable latents for the future embeddings and cross-attend (Jaegle et al 2021; Lee et al 2019) them with observed video embeddings. Generative approaches have learned future embeddings based on pre-defined temporal states (Piergiovanni et al 2020), logit sequences (Zhao and Wildes 2020), using cyclic consistency (Abu Farha et al 2021), or through learning the expected variance in future representations (Mascaró et al 2023; Patsch et al 2024). Recently, Mittal et al (2024) used general language and visual queries to infer prediction through LLMs.

Next active object. A recently introduced set of anticipation tasks also study object-centric future forecasting. Next active object anticipation aims to forecast the objects that will be used in future actions with approaches using predictions on the salient regions (Dessalene et al 2021), hand position generated representations (Jiang et al 2021), or autoregressively attending object and visual information (Thakur et al 2024). Other methods may also forecast human object interaction regions (Liu et al 2020, 2022c; Roy et al 2024), object relations (Roy and Fernando 2021; Zatsarynna et al 2021), or time-to-contact estimates (Mur-Labadia et al 2024).

Future outlooks.

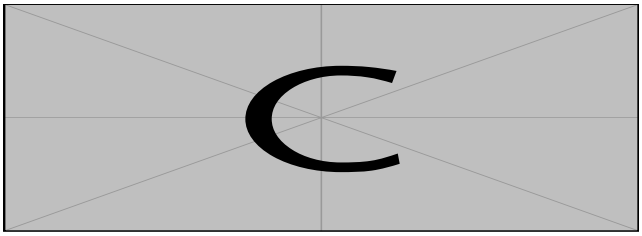


Fig. 9. TODO: Visual examples of generative models.

6.2 Video Generation ✔

Generative models short sequences GANs (Clark et al 2019) (Saito et al 2017) (Luc et al 2020) (Vondrick et al 2016b) (Menapace et al 2021) (Yu et al 2022b)

GANs for long sequences (Ge et al 2022a) (Brooks et al 2022) (Shen et al 2023) (Skorokhodov et al 2022)

Attention-based generative models (Han et al 2022a) (Hu et al 2022d)

Diffusion (Ge et al 2022a) (Blattmann et al 2023) (Nikankin et al 2023) (Yang et al 2023b) (Ho et al 2022b) (He et al 2022c) (Wu et al 2023b) (Harvey et al 2022) (Yu et al 2021) (Liu et al 2024c) (Hong et al 2022b) (Yu et al 2023c) (Zeng et al 2024) (Yu et al 2024)

Text-conditioned diffusion (Yan et al 2021) (Gupta et al 2023) (Zhuang et al 2024)

Future generation with diffusion (Fu et al 2023)

autoregressive (Ho et al 2022a) (Villegas et al 2022) (Singer et al 2023) (Wu et al 2021a) (Wu et al 2022a)

(Chang et al 2024)

7 Challenges and peeking through the future

8 Discussion

References

Aakur SN, Sarkar S (2019) A perceptual prediction framework for self supervised event segmentation. In: CVPR

Abati D, Ben Yahia H, Nagel M, Habibian A (2023) Resq: Residual quantization for video perception. In: ICCV

Abdelsalam MA, Rangrej SB, Hadji I, Dvornik N, Derpanis KG, Fazly A (2023) Gepsan: Generative procedure step anticipation in cooking videos. In: ICCV

Abu-El-Haija S, Kothari N, Lee J, Natsev P, Toderici G, Varadarajan B, Vijayanarasimhan S (2016) Youtube-8m: A large-scale video classification benchmark. arxiv

Abu Farha Y, Richard A, Gall J (2018) When will you do what?-anticipating temporal occurrences of activities. In: CVPR

Abu Farha Y, Ke Q, Schiele B, Gall J (2021) Long-term anticipation of activities with cycle consistency. In: DAGM GCPR

Acsintoae A, Florescu A, Georgescu MI, Mare T, Sumedrea P, Ionescu RT, Khan FS, Shah M (2022) Ubnormal: New benchmark for supervised open-set video anomaly detection. In: CVPR

Agarwal N, Chen YT, Dariush B, Yang MH (2020) Unsupervised domain adaptation for spatio-temporal action localization. In: BMVC

Aggarwal JK, Cai Q (1999) Human motion analysis: A review. CVIU

Aggarwal JK, Cai Q, Liao W, Sabata B (1994) Articulated and elastic non-rigid motion: A review. In: Workshop on Motion of Non-rigid and Articulated Objects

Aggarwal JK, Cai Q, Liao W, Sabata B (1998) Nonrigid motion analysis: Articulated and elastic motion. CVIU

Akbari H, Yuan L, Qian R, Chuang WH, Chang SF, Cui Y, Gong B (2021) Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. NeurIPS

Alayrac JB, Bojanowski P, Agrawal N, Sivic J, Laptev I, Lacoste-Julien S (2016) Unsupervised learning from narrated instruction videos. In: CVPR

Alayrac JB, Laptev I, Sivic J, Lacoste-Julien S (2017) Joint discovery of object states and manipulation actions. In: ICCV

Alayrac JB, Donahue J, Luc P, Miech A, Barr I, Hasson Y, Lenc K, Mensch A, Millican K, Reynolds M, et al (2022) Flamingo: a visual language model for few-shot learning. NeurIPS

Alayrac JB, Miech A, Laptev I, Sivic J, et al (2024) Multitask learning of object states and state-modifying actions from web videos. IEEE TPAMI

Albanese M, Chellappa R, Cuntoor N, Moscato V, Picariello A, Subrahmanian V, Udrea O (2010) Pads: A probabilistic activity detection framework for video data. IEEE TPAMI

Albanie S, Liu Y, Nagrani A, Miech A, Coto E, Laptev I, Sukthankar R, Ghanem B, Zisserman A, Gabeur V, et al (2020) The end-of-end-to-end: A video understanding pentathlon challenge (2020). arXiv

Albu AB, Bergevin R, Quirion S (2008) Generic Temporal Segmentation of Cyclic Human Motion. PR

Aleksic PS, Katsaggelos AK (2006) Audio-visual biometrics. Proceedings of the IEEE

AlMarri S, Zaheer MZ, Nandakumar K (2024) A multi-head approach with shuffled segments for weakly-supervised video anomaly detection. In: WACVw

Alwassel H, Heilbron FC, Escorcia V, Ghanem B (2018) Diagnosing error in temporal action detectors. In: ECCV

Alwassel H, Giancola S, Ghanem B (2021) Tsp: Temporally-sensitive pretraining of video encoders for localization tasks. In: ICCV

Amer MR, Todorovic S (2012) Sum-product networks for modeling activities with stochastic structure. In: CVPR

Anderson P, Wu Q, Teney D, Bruce J, Johnson M, Sünderhauf N, Reid I, Gould S, Van Den Hengel A (2018) Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In: CVPR

Andrew G, Arora R, Bilmes J, Livescu K (2013) Deep canonical correlation analysis. In: ICML

Anne Hendricks L, Wang O, Shechtman E, Sivic J, Darrell T, Russell B (2017) Localizing moments in video with natural language. In: ICCV

Arandjelovic R, Zisserman A (2018) Objects that sound. In: ECCV

Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C (2021a) Vivit: A video vision transformer. In: ICCV

Arnab A, Sun C, Schmid C (2021b) Unified graph structured models for video understanding. In: CVPR

Ashutosh K, Girdhar R, Torresani L, Grauman K (2023) Hiervl: Learning hierarchical video-language embeddings. In: CVPR

Ashutosh K, Ramakrishnan SK, Afouras T, Grauman K (2024) Video-mined task graphs for keystep recognition in instructional videos. NeurIPS

Astrid M, Zaheer MZ, Lee JY, Lee SI (2021a) Learning not to reconstruct anomalies. In: BMVC

Astrid M, Zaheer MZ, Lee SI (2021b) Synthetic temporal anomaly guided end-to-end video anomaly detection. In: ICCVw

Azy O, Ahuja N (2008) Segmentation of Periodically Moving Objects. In: ICPR

Baade A, Peng P, Harwath D (2022) Mae-ast: Masked autoencoding audio spectrogram transformer. In: Interspeech

Babaeizadeh M, Finn C, Erhan D, Campbell RH, Levine S (2018) Stochastic variational video prediction. In: ICLR

Baccouche M, Mamalet F, Wolf C, Garcia C, Baskurt A (2011) Sequential deep learning for human action recognition. In: HBU

Bacharidis K, Argyros A (2023) Repetition-aware Image Sequence Sampling for Recognizing Repetitive Human Actions. In: ICCVw

Bai Y, Wang Y, Tong Y, Yang Y, Liu Q, Liu J (2020) Boundary content graph neural network for temporal action proposal generation. In: ECCV

Bain M, Nagrani A, Varol G, Zisserman A (2021) Frozen in time: A joint video and image encoder for end-to-end retrieval. In: ICCV

Ballas N, Yao L, Pal C, Courville A (2015) Delving deeper into convolutional networks for learning video representations. In: ICLR

Bandara WGC, Patel N, Gholami A, Nikkhah M, Agrawal M, Patel VM (2023) Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In: CVPR

Bansal S, Arora C, Jawahar C (2022) My view is the best view: Procedure learning from egocentric videos. In: ECCV

Barekatain M, Martí M, Shih HF, Murray S, Nakayama K, Matsuo Y, Prendinger H (2017) Okutama-action: An aerial view video dataset for concurrent human action detection. In: ICCVw

Becattini F, Uricchio T, Seidenari L, Ballan L, Bimbo AD (2020) Am i done? predicting action progress in videos. TOMM

Beddiar DR, Nini B, Sabokrou M, Hadid A (2020) Vision-based human activity recognition: a survey. MTA

Ben-Shabat Y, Yu X, Saleh F, Campbell D, Rodriguez-Opazo C, Li H, Gould S (2021) The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose. In: WACV

BenAbdelkader C, Cutler RG, Davis LS (2004) Gait recognition using image self-similarity. EURASIP

Bertasius G, Wang H, Torresani L (2021) Is space-time attention all you need for video understanding? In: ICML

Bilen H, Fernando B, Gavves E, Vedaldi A, Gould S (2016) Dynamic image networks for action recognition. In: CVPR

Blank M, Gorelick L, Shechtman E, Irani M, Basri R (2005) Actions as space-time shapes. In: ICCV

Blattmann A, Rombach R, Ling H, Dockhorn T, Kim SW, Fidler S, Kreis K (2023) Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR

Bobick AF, Davis JW (2001) The recognition of human movement using temporal templates. IEEE TPAMI

Bokhari SZ, Kitani KM (2017) Long-term activity forecasting using first-person vision. In: ACCV

Briassouli A, Ahuja N (2007) Extraction and Analysis of Multiple Periodic Motions in Video Sequences. IEEE TPAMI)

Brooks T, Hellsten J, Aittala M, Wang TC, Aila T, Lehtinen J, Liu MY, Efros A, Karras T (2022) Generating long videos of dynamic scenes. NeurIPS

Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, Neelakantan A, Shyam P, Sastry G, Askell A, et al (2020) Language models are few-shot learners. NeurIPS

Broxton M, Flynn J, Overbeck R, Erickson D, Hedman P, Duvall M, Dourgarian J, Busch J, Whalen M, Debevec P (2020) Immersive light field video with a layered mesh representation. ACM TOG

Bulat A, Perez Rua JM, Sudhakaran S, Martinez B, Tzimiropoulos G (2021) Space-time mixing attention for video transformer. NeurIPS

Buxton H (2003) Learning and understanding dynamic scene activity: a review. IVC

Caba Heilbron F, Escorcia V, Ghanem B, Carlos Niebles J (2015) Activitynet: A large-scale video benchmark for human activity understanding. In: CVPR

Cai Y, Li H, Hu JF, Zheng WS (2019) Action knowledge transfer for action prediction with partial videos. In: AAAI

Calvo-Merino B, Glaser DE, Grèzes J, Passingham RE, Haggard P (2005) Action observation and acquired motor skills: an fmri study with expert dancers. Cerebral cortex

Cao M, Chen L, Shou MZ, Zhang C, Zou Y (2021) On pursuit of designing multi-modal transformer for video grounding. In: EMNLP

Cao Y, Barrett D, Barbu A, Narayanaswamy S, Yu H, Michaux A, Lin Y, Dickinson S, Mark Siskind J, Wang S (2013) Recognize human activities from partially observed videos. In: CVPR

Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: ECCV

Carreira J, Zisserman A (2017) Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR

Carreira J, Noland E, Banki-Horvath A, Hillier C, Zisserman A (2018) A short note about kinetics-600. arxiv

Carreira J, Noland E, Hillier C, Zisserman A (2019) A short note on the kinetics-700 human action dataset. arxiv

Castrejon L, Ballas N, Courville A (2019) Improved conditional vrnns for video prediction. In: CVPR

Cedras C, Shah M (1995) Motion-based recognition a survey. IVC

Chaabane M, Trabelsi A, Blanchard N, Beveridge R (2020) Looking ahead: Anticipating pedestrians crossing with future frames prediction. In: WACV

Chaaraoui AA, Climent-Pérez P, Flórez-Revuelta F (2012) A review on vision techniques applied to human behaviour analysis for ambient-assisted living. ESWA

Chadha A, Arora G, Kaloty N (2021) iperceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. In: WACV

Chang CY, Huang DA, Sui Y, Fei-Fei L, Niebles JC (2019) D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In: CVPR

Chang M, Prakash A, Gupta S (2024) Look ma, no hands! agent-environment factorization of egocentric videos. NeurIPS

Chang Z, Zhang X, Wang S, Ma S, Ye Y, Xinguang X, Gao W (2021) Mau: A motion-aware unit for video prediction and beyond. In: NeurIPS

Chang Z, Zhang X, Wang S, Ma S, Gao W (2022) Strpm: A spatiotemporal residual predictive model for high-resolution video prediction. In: CVPR

Chao YW, Vijayanarasimhan S, Seybold B, Ross DA, Deng J, Sukthankar R (2018) Rethinking the faster r-cnn architecture for temporal action localization. In: CVPR

Chao YW, Yang W, Xiang Y, Molchanov P, Handa A, Tremblay J, Narang YS, Van Wyk K, Iqbal U, Birchfield S, et al (2021) Dexycb: A benchmark for capturing hand grasping of objects. In: CVPR

Chen D, Dolan WB (2011) Collecting highly parallel data for paraphrase evaluation. In: ACL

Chen G, Zheng YD, Wang L, Lu T (2022a) Dcan: improving temporal action detection via dual context aggregation. In: AAAI

Chen H, Xie W, Vedaldi A, Zisserman A (2020a) Vggsound: A large-scale audio-visual dataset. In: ICASSP

Chen J, Chen X, Ma L, Jie Z, Chua TS (2018a) Temporally grounding natural sentence in video. In: EMNLP

Chen L, Lu C, Tang S, Xiao J, Zhang D, Tan C, Li X (2020b) Rethinking the bottom-up framework for query-based video localization. In: AAAI

Chen L, Lu J, Song Z, Zhou J (2022b) Ambiguousness-aware state evolution for action prediction. IEEE TCSVT

Chen S, Jiang YG (2021) Towards bridging event captioner and sentence localizer for weakly supervised dense event captioning. In: CVPR

Chen S, Sun P, Xie E, Ge C, Wu J, Ma L, Shen J, Luo P (2021) Watch only once: An end-to-end video action detection framework. In: ICCV

Chen T, Rao RR (1998) Audio-visual integration in multimodal communication. Proceedings of the IEEE

Chen T, Kornblith S, Norouzi M, Hinton G (2020c) A simple framework for contrastive learning of visual representations. In: ICML

Chen TS, Siarohin A, Menapace W, Deyneka E, Chao Hw, Jeon BE, Fang Y, Lee HY, Ren J, Yang MH, et al (2024)

Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In: CVPR
Chen X, Wang W, Wang J, Li W (2017) Learning object-centric transformation for video prediction. In: MM
Chen Y, Kalantidis Y, Li J, Yan S, Feng J (2018b) Aˆ 2-nets: Double attention networks. NeurIPS
Chen Y, Kalantidis Y, Li J, Yan S, Feng J (2018c) Multi-fiber networks for video recognition. In: ECCV
Chen Y, Fan H, Xu B, Yan Z, Kalantidis Y, Rohrbach M, Yan S, Feng J (2019) Drop an octave: Reducing spatial redundancy in convolutional neural networks with octave convolution. In: CVPR
Chen Y, Liu Z, Zhang B, Fok W, Qi X, Wu YC (2023) Mgfn: Magnitude-contrastive glance-and-focus network for weakly-supervised video anomaly detection. In: AAAI
Cheng F, Bertasius G (2022) Tallformer: Temporal action localization with a long-memory transformer. In: ECCV
Cheng F, Xu M, Xiong Y, Chen H, Li X, Li W, Xia W (2022) Stochastic backpropagation: A memory efficient strategy for training video models. In: CVPR
Cheng S, Guo Z, Wu J, Fang K, Li P, Liu H, Liu Y (2024) Egothink: Evaluating first-person perspective thinking capability of vision-language models. In: CVPR
Cherian A, Hori C, Marks TK, Le Roux J (2022) (2.5+ 1) d spatio-temporal scene graphs for video question answering. In: AAAI
Chi Hg, Lee K, Agarwal N, Xu Y, Ramani K, Choi C (2023) Adamsformer for spatial action localization in the future. In: CVPR
Cho M, Kim T, Kim WJ, Cho S, Lee S (2022) Unsupervised video anomaly detection via normalizing flows with implicit latent features. PR
Choi J, Gao C, Messou JC, Huang JB (2019) Why can't i dance in the mall? learning to mitigate scene bias in action recognition. NeurIPS
Chung J, Zisserman A (2016) Signs in time: Encoding human motion as a temporal image. In: ECCVw
Chung J, Wuu Ch, Yang Hr, Tai YW, Tang CK (2021) Haa500: Human-centric atomic action dataset with curated videos. In: ICCV
Cipolla R, Blake A (1990) The dynamic analysis of apparent contours. In: ICCV
Clark A, Donahue J, Simonyan K (2019) Adversarial video generation on complex datasets. arxiv
Cui Y, Zeng C, Zhao X, Yang Y, Wu G, Wang L (2023) Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In: ICCV
Cutler R, Davis LS (2000) Robust Real-Time Periodic Motion Detection, Analysis, and Applications. IEEE TPAMI
Dai R, Das S, Sharma S, Minciullo L, Garattoni L, Bremond F, Francesca G (2022a) Toyota smarthome untrimmed: Real-world untrimmed videos for activity detection. IEEE TPAMI
Dai Y, Tang D, Liu L, Tan M, Zhou C, Wang J, Feng Z, Zhang F, Hu X, Shi S (2022b) One model, multiple modalities: A sparsely activated approach for text, sound, image, video and code. arxiv
Damen D, Doughty H, Farinella GM, Fidler S, Furnari A, Kazakos E, Moltisanti D, Munro J, Perrett T, Price W, et al (2018) Scaling egocentric vision: The epic-kitchens dataset. In: ECCV
Damen D, Doughty H, Farinella GM, Furnari A, Kazakos E, Ma J, Moltisanti D, Munro J, Perrett T, Price W, et al (2022) Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. IJCV
Dang LH, Le TM, Le V, Tran T (2021) Hierarchical object-oriented spatio-temporal reasoning for video question answering. In: IJCAI
Davtyan A, Sameni S, Favaro P (2023) Efficient video prediction via sparsely conditioned flow matching. In: ICCV
De Geest R, Gavves E, Ghodrati A, Li Z, Snoek C, Tuytelaars T (2016) Online action detection. In: ECCV
Deng C, Chen S, Chen D, He Y, Wu Q (2021) Sketch, ground, and refine: Top-down dense video captioning. In: CVPR
Denton E, Fergus R (2018) Stochastic video generation with a learned prior. In: ICML
Dessalene E, Devaraj C, Maynord M, Fermüller C, Aloimonos Y (2021) Forecasting action through contact representations from first person video. IEEE TPAMI
Destro M, Gygli M (2024) CycleCL: Self-supervised Learning for Periodic Videos. In: WACV
Dhariwal P, Nichol A (2021) Diffusion models beat gans on image synthesis. NeurIPS
Dhiman C, Vishwakarma DK (2019) A review of state-of-the-art techniques for abnormal human activity recognition. EAAI
Diba A, Fayyaz M, Sharma V, Paluri M, Gall J, Stiefelhagen R, Van Gool L (2020) Large scale holistic video understanding. In: ECCV
Dietterich TG, Lathrop RH, Lozano-Pérez T (1997) Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence
Diko A, Avola D, Prenkaj B, Fontana F, Cinque L (2024) Semantically guided representation learning for action anticipation. In: ECCV
Ding G, Sener F, Yao A (2023) Temporal action segmentation: An analysis of modern techniques. IEEE TPAMI
Dollár P, Rabaud V, Cottrell G, Belongie S (2005) Behavior recognition via sparse spatio-temporal features. In: VS-PETS
Donahue G, Elhamifar E (2024) Learning to predict activity progress by self-supervised video alignment. In: CVPR
Donahue J, Anne Hendricks L, Guadarrama S, Rohrbach M, Venugopalan S, Saenko K, Darrell T (2015) Long-term recurrent convolutional networks for visual recognition and description. In: CVPR
Dong J, Li X, Snoek CG (2018) Predicting visual features from text for image and video caption retrieval. TM
Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, et al (2020) An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR
Doughty H, Snoek CG (2022) How do you do it? fine-grained action understanding with pseudo-adverbs. In: CVPR
Doughty H, Damen D, Mayol-Cuevas W (2018) Who's better? who's best? pairwise deep ranking for skill determination. In: CVPR
Doughty H, Laptev I, Mayol-Cuevas W, Damen D (2020) Action modifiers: Learning from adverbs in instructional videos. In: CVPR
Du W, Wang Y, Qiao Y (2017) Recurrent spatial-temporal attention network for action recognition in videos. IEEE T-IP
Dubey S, Boragule A, Jeon M (2019) 3d resnet with ranking loss function for abnormal activity detection in videos. In: ICCAIS
Dvornik M, Hadji I, Derpanis KG, Garg A, Jepson A (2021) Drop-dtw: Aligning common signal between sequences while dropping outliers. NeurIPS
Dwibedi D, Sermanet P, Tompson J (2018) Temporal reasoning in videos using convolutional gated recurrent units.

In: CVPRw
Dwibedi D, Aytar Y, Tompson J, Sermanet P, Zisserman A (2020) Counting out time: Class agnostic video repetition counting in the wild. In: CVPR
Dwibedi D, Aytar Y, Tompson J, Zisserman A (2024) Ovr: A dataset for open vocabulary temporal repetition counting in videos. arXiv
Edwards M, Deng J, Xie X (2016) From pose to activity: Surveying datasets and introducing converse. CVIU
Efros A, Berg A, Mori G, Malik J (2003) Recognizing action at a distance. In: ICCV
Epstein D, Chen B, Vondrick C (2020) Oops! predicting unintentional action in video. In: CVPR
Epstein D, Wu J, Schmid C, Sun C (2021) Learning temporal dynamics from cycles in narrated video. In: ICCV
Escorcia V, Soldan M, Sivic J, Ghanem B, Russell B (2019) Temporal localization of moments in video collections with natural language. arxiv
Fan C, Zhang X, Zhang S, Wang W, Zhang C, Huang H (2019) Heterogeneous memory enhanced multimodal attention model for video question answering. In: CVPR
Fan H, Xiong B, Mangalam K, Li Y, Yan Z, Malik J, Feichtenhofer C (2021) Multiscale vision transformers. In: ICCV
Fathi A, Rehg JM (2013) Modeling actions through state changes. In: CVPR
Faure GJ, Chen MH, Lai SH (2023) Holistic interaction transformer network for action detection. In: WACV
Fayek HM, Kumar A (2020) Large scale audiovisual learning of sounds with weakly labeled data. In: IJCAI
Feichtenhofer C (2020) X3d: Expanding architectures for efficient video recognition. In: CVPR
Feichtenhofer C, Pinz A, Zisserman A (2016) Convolutional two-stream network fusion for video action recognition. In: CVPR
Feichtenhofer C, Pinz A, Wildes RP (2017) Spatiotemporal multiplier networks for video action recognition. In: CVPR
Feichtenhofer C, Fan H, Malik J, He K (2019) Slowfast networks for video recognition. In: ICCV
Feichtenhofer C, Li Y, He K, et al (2022) Masked autoencoders as spatiotemporal learners. NeurIPS
Feng J, Erol MH, Chung JS, Senocak A (2024) From coarse to fine: Efficient training for audio spectrogram transformers. In: ICASSP
Feng JC, Hong FT, Zheng WS (2021a) Mist: Multiple instance self-training framework for video anomaly detection. In: CVPR
Feng Y, Jiang J, Huang Z, Qing Z, Wang X, Zhang S, Tang M, Gao Y (2021b) Relation modeling in spatio-temporal action localization. In: CVPRw
Fernando B, Herath S (2021) Anticipating human actions by correlating past with the future with jaccard similarity measures. In: CVPR
Fernando B, Gavves E, Oramas JM, Ghodrati A, Tuytelaars T (2015) Modeling video evolution for action recognition. In: CVPR
Fernando B, Gavves E, Oramas J, Ghodrati A, Tuytelaars T (2016) Rank pooling for action recognition. IEEE TPAMI
Fernando B, Bilen H, Gavves E, Gould S (2017) Self-supervised video representation learning with odd-one-out networks. In: CVPR
Ferreira B, Ferreira PM, Pinheiro G, Figueiredo N, Carvalho F, Menezes P, Batista J (2021) Deep Learning Approaches for Workout Repetition Counting and Validation. PRL
Fioresi J, Dave IR, Shah M (2023) Ted-spad: Temporal distinctiveness for self-supervised privacy-preservation for video anomaly detection. In: ICCV
Flaborea A, Collorone L, Di Melendugno GMD, D'Arrigo S, Prenkaj B, Galasso F (2023) Multimodal motion conditioned diffusion model for skeleton-based video anomaly detection. In: ICCV
Flanagan K, Damen D, Wray M (2023) Learning temporal sentence grounding from narrated egovideos. In: BMVC
Fogassi L, Ferrari PF, Gesierich B, Rozzi S, Chersi F, Rizzolatti G (2005) Parietal lobe: from action organization to intention understanding. Science
Foo LG, Li T, Rahmani H, Ke Q, Liu J (2022) Era: Expert retrieval and assembly for early action prediction. In: ECCV
Förstner W, Gülch E (1987) A fast operator for detection and precise location of distinct points, corners and centres of circular features. In: ICFPPD
Fouhey DF, Kuo Wc, Efros AA, Malik J (2018) From lifestyle vlogs to everyday interactions. In: CVPR
Fu Q, Liu X, Kitani KM (2022) Sequential decision-making for active object detection from hand. In: CVPR
Fu TJ, Li L, Gan Z, Lin K, Wang WY, Wang L, Liu Z (2021) Violet: End-to-end video-language transformers with masked visual-token modeling. arxiv
Fu TJ, Yu L, Zhang N, Fu CY, Su JC, Wang WY, Bell S (2023) Tell me what happened: Unifying text-guided video completion via multimodal masked video generation. In: CVPR
Furnari A, Farinella GM (2019) What would you expect? anticipating egocentric actions with rolling-unrolling lstms and modality attention. In: ICCV
Furnari A, Battiato S, Maria Farinella G (2018) Leveraging uncertainty to rethink loss functions and evaluation measures for egocentric action anticipation. In: ECCVw
Gabeur V, Sun C, Alahari K, Schmid C (2020) Multi-modal transformer for video retrieval. In: ECCV
Gallese V, Fadiga L, Fogassi L, Rizzolatti G (1996) Action recognition in the premotor cortex. Brain
Gammulle H, Denman S, Sridharan S, Fookes C (2019) Predicting the future: A jointly learnt model for action anticipation. In: ICCV
Gao D, Zhou L, Ji L, Zhu L, Yang Y, Shou MZ (2023) Mist: Multi-modal iterative spatial-temporal transformer for long-form video question answering. In: CVPR
Gao J, Sun C, Yang Z, Nevatia R (2017a) Tall: Temporal activity localization via language query. In: ICCV
Gao J, Yang Z, Nevatia R (2017b) Red: Reinforced encoder-decoder networks for action anticipation. arxiv
Gao J, Ge R, Chen K, Nevatia R (2018) Motion-appearance co-memory networks for video question answering. In: CVPR
Gao R, Oh TH, Grauman K, Torresani L (2020) Listen to look: Action recognition by previewing audio. In: CVPR
Gao Z, Tan C, Wu L, Li SZ (2022) Simvp: Simpler yet better video prediction. In: CVPR
Garcia-Hernando G, Yuan S, Baek S, Kim TK (2018) First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: CVPR
Ge R, Gao J, Chen K, Nevatia R (2019) Mac: Mining activity concepts for language-based temporal localization. In: WACV
Ge S, Hayes T, Yang H, Yin X, Pang G, Jacobs D, Huang JB, Parikh D (2022a) Long video generation with time-agnostic vqgan and time-sensitive transformer. In: ECCV

Ge Y, Ge Y, Liu X, Li D, Shan Y, Qie X, Luo P (2022b) Bridging video-text retrieval with multiple choice questions. In: CVPR

Gemmeke JF, Ellis DP, Freedman D, Jansen A, Lawrence W, Moore RC, Plakal M, Ritter M (2017) Audio set: An ontology and human-labeled dataset for audio events. In: ICASSP

Geng S, Gao P, Chatterjee M, Hori C, Le Roux J, Zhang Y, Li H, Cherian A (2021) Dynamic graph representation learning for video dialog via multi-modal shuffled transformers. In: AAAI

Georgescu MI, Barbalau A, Ionescu RT, Khan FS, Popescu M, Shah M (2021) Anomaly detection in video via self-supervised and multi-task learning. In: CVPR

Georgescu MI, Fonseca E, Ionescu RT, Lucic M, Schmid C, Arnab A (2023) Audiovisual masked autoencoders. In: ICCV

Ghadiyaram D, Tran D, Mahajan D (2019) Large-scale weakly-supervised pre-training for video action recognition. In: CVPR

Ghodrati A, Bejnordi BE, Habibian A (2021) Frameexit: Conditional early exiting for efficient video recognition. In: CVPR

Girdhar R, Grauman K (2021) Anticipative video transformer. In: ICCV

Girdhar R, Ramanan D (2017) Attentional pooling for action recognition. NeurIPS

Girdhar R, Carreira J, Doersch C, Zisserman A (2019) Video action transformer network. In: CVPR

Girshick R (2015) Fast r-cnn. In: ICCV

Girshick R, Donahue J, Darrell T, Malik J (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: CVPR

Gong D, Lee J, Kim M, Ha SJ, Cho M (2022a) Future transformer for long-term action anticipation. In: CVPR

Gong Y, Chung YA, Glass J (2021) Psla: Improving audio tagging with pretraining, sampling, labeling, and aggregation. IEEE/ACM TASLP

Gong Y, Liu AH, Rouditchenko A, Glass J (2022b) Uavm: Towards unifying audio and visual models. IEEE SPL

Gong Y, Rouditchenko A, Liu AH, Harwath D, Karlinsky L, Kuehne H, Glass J (2023) Contrastive audio-visual masked autoencoder. In: ICLR

Gordo A, Larlus D (2017) Beyond instance-level image retrieval: Leveraging captions to learn a global visual representation for semantic retrieval. In: CVPR

Gorelick L, Galun M, Sharon E, Basri R, Brandt A (2006) Shape representation and classification using the poisson equation. IEEE TPAMI

Gorelick L, Blank M, Shechtman E, Irani M, Basri R (2007) Actions as space-time shapes. IEEE TPAMI

Goroshin R, Bruna J, Tompson J, Eigen D, LeCun Y (2015) Unsupervised learning of spatiotemporally coherent metrics. In: ICCV

Gouidis F, Patkos T, Argyros A, Plexousakis D (2023) Leveraging knowledge graphs for zero-shot object-agnostic state classification. arxiv

Gowda SN, Rohrbach M, Sevilla-Lara L (2021) Smart frame selection for action recognition. In: AAAI

Goyal R, Ebrahimi Kahou S, Michalski V, Materzynska J, Westphal S, Kim H, Haenel V, Fruend I, Yianilos P, Mueller-Freitag M, et al (2017) The" something something" video database for learning and evaluating visual common sense. In: ICCV

Grauman K, Westbury A, Byrne E, Chavis Z, Furnari A, Girdhar R, Hamburger J, Jiang H, Liu M, Liu X, et al (2022) Ego4d: Around the world in 3,000 hours of egocentric video. In: CVPR

Grauman K, Westbury A, Torresani L, Kitani K, Malik J, Afouras T, Ashutosh K, Baiyya V, Bansal S, Boote B, et al (2024) Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: CVPR

Gritsenko AA, Xiong X, Djolonga J, Dehghani M, Sun C, Lucic M, Schmid C, Arnab A (2024) End-to-end spatio-temporal action localisation with video transformers. In: CVPR

Gu C, Sun C, Ross DA, Vondrick C, Pantofaru C, Li Y, Vijayanarasimhan S, Toderici G, Ricco S, Sukthankar R, et al (2018) Ava: A video dataset of spatio-temporally localized atomic visual actions. In: CVPR

Gu X, Fan H, Huang Y, Luo T, Zhang L (2024a) Context-guided spatio-temporal video grounding. In: CVPR

Gu X, Wen C, Ye W, Song J, Gao Y (2024b) Seer: Language instructed video prediction with latent diffusion models. In: ICLR

Guen VL, Thome N (2020) Disentangling physical dynamics from unknown factors for unsupervised video prediction. In: CVPR

Gulati A, Qin J, Chiu CC, Parmar N, Zhang Y, Yu J, Han W, Wang S, Zhang Z, Wu Y, et al (2020) Conformer: Convolution-augmented transformer for speech recognition. Interspeech

Guo H, Agarwal N, Lo SY, Lee K, Ji Q (2024a) Uncertainty-aware action decoupling transformer for action anticipation. In: CVPR

Guo Y, Sun S, Ma S, Zheng K, Bao X, Ma S, Zou W, Zheng Y (2024b) Crossmae: Cross-modality masked autoencoders for region-aware audio-visual pre-training. In: CVPR

Guo Z, Zhao J, Jiao L, Liu X, Li L (2021) Multi-scale progressive attention network for video question answering. In: ACL

Gupta A, Kembhavi A, Davis LS (2009) Observing human-object interactions: Using spatial and functional compatibility for recognition. IEEE TPAMI

Gupta A, Yu L, Sohn K, Gu X, Hahn M, Fei-Fei L, Essa I, Jiang L, Lezama J (2023) Photorealistic video generation with diffusion models. arxiv

Hadji I, Derpanis KG, Jepson AD (2021) Representation learning via global temporal alignment and cycle-consistency. In: CVPR

Hakeem A, Shah M (2004) Ontology and taxonomy collaborated framework for meeting classification. In: ICPR

Hampali S, Rad M, Oberweger M, Lepetit V (2020) Honnotate: A method for 3d annotation of hand and object poses. In: CVPR

Han L, Ren J, Lee HY, Barbieri F, Olszewski K, Minaee S, Metaxas D, Tulyakov S (2022a) Show me what and tell me how: Video synthesis via multimodal conditioning. In: CVPR

Han T, Xie W, Zisserman A (2022b) Temporal alignment networks for long-term video. In: CVPR

Han T, Bain M, Nagrani A, Varol G, Xie W, Zisserman A (2023a) Autoad ii: The sequel-who, when, and what in movie audio description. In: ICCV

Han T, Bain M, Nagrani A, Varol G, Xie W, Zisserman A (2023b) Autoad: Movie description in context. In: CVPR

Han T, Bain M, Nagrani A, Varol G, Xie W, Zisserman A (2024) Autoad iii: The prequel-back to the pixels. In: CVPR

Hao J, Sun H, Ren P, Wang J, Qi Q, Liao J (2022) Query-aware video encoder for video moment retrieval. Neurocomputing

Hara K, Kataoka H, Satoh Y (2018) Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In: CVPR

Haresh S, Kumar S, Coskun H, Syed SN, Konin A, Zia Z, Tran QH (2021) Learning by aligning videos in time. In: CVPR

Harris C, Stephens M, et al (1988) A combined corner and edge detector. In: AVC

Harvey W, Naderiparizi S, Masrani V, Weilbach C, Wood F (2022) Flexible diffusion modeling of long videos. NeurIPS

Hasan M, Choi J, Neumann J, Roy-Chowdhury AK, Davis LS (2016) Learning temporal regularity in video sequences. In: CVPR

He B, Yang X, Kang L, Cheng Z, Zhou X, Shrivastava A (2022a) Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization. In: CVPR

He K, Chen X, Xie S, Li Y, Dollár P, Girshick R (2022b) Masked autoencoders are scalable vision learners. In: CVPR

He Y, Yang T, Zhang Y, Shan Y, Chen Q (2022c) Latent video diffusion models for high-fidelity video generation with arbitrary lengths. arxiv

Hegde K, Agrawal R, Yao Y, Fletcher CW (2018) Morph: Flexible acceleration for 3d cnn-based video understanding. In: MICRO

Heidarivincheh F, Mirmehdi M, Damen D (2016) Beyond action recognition: Action completion in rgb-d data. In: BMVC

Heidarivincheh F, Mirmehdi M, Damen D (2018) Action completion: A temporal model for moment detection. In: BMVC

Herath S, Harandi M, Porikli F (2017) Going deeper into action recognition: A survey. IVC

Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. NeurIPS

Ho J, Chan W, Saharia C, Whang J, Gao R, Gritsenko A, Kingma DP, Poole B, Norouzi M, Fleet DJ, et al (2022a) Imagen video: High definition video generation with diffusion models. arxiv

Ho J, Salimans T, Gritsenko A, Chan W, Norouzi M, Fleet DJ (2022b) Video diffusion models. NeurIPS

Hoai M, De la Torre F (2014) Max-margin early event detectors. IJCV

Hoai M, Zisserman A (2015) Improving human action recognition using score distribution and ranking. In: ACCV

Hong J, Zhang H, Gharbi M, Fisher M, Fatahalian K (2022a) Spotting temporally precise, fine-grained events in video. In: ECCV

Hong W, Ding M, Zheng W, Liu X, Tang J (2022b) Cogvideo: Large-scale pretraining for text-to-video generation via transformers. arxiv

Hong X, Lan Y, Pang L, Guo J, Cheng X (2021) Transformation driven visual reasoning. In: CVPR

Höppe T, Mehrjou A, Bauer S, Nielsen D, Dittadi A (2024) Diffusion models for video prediction and infilling. IEEE TMLR

Hou J, Wu X, Wang R, Luo J, Jia Y (2020) Confidence-guided self refinement for action prediction in untrimmed videos. IEEE T-IP

Hou R, Chen C, Shah M (2017) Tube convolutional neural network (t-cnn) for action detection in videos. In: ICCV

Hu D, Nie F, Li X (2019) Deep multimodal clustering for unsupervised audiovisual learning. In: CVPR

Hu H, Dong S, Zhao Y, Lian D, Li Z, Gao S (2022a) Transrac: Encoding multi-scale temporal correlation with transformers for repetitive action counting. In: CVPR

Hu JF, Zheng WS, Ma L, Wang G, Lai J, Zhang J (2018) Early action prediction by soft regression. IEEE TPAMI

Hu X, Chen Z, Owens A (2022b) Mix and localize: Localizing sound sources in mixtures. In: CVPR

Hu X, Dai J, Li M, Peng C, Li Y, Du S (2022c) Online human action detection and anticipation in videos: A survey. Neurocomputing

Hu X, Huang Z, Huang A, Xu J, Zhou S (2023) A dynamic multi-scale voxel flow network for video prediction. In: CVPR

Hu Y, Luo C, Chen Z (2022d) Make it move: controllable image-to-video generation with text descriptions. In: CVPR

Huang D, Chen P, Zeng R, Du Q, Tan M, Gan C (2020a) Location-aware graph convolutional networks for video question answering. In: AAAI

Huang DA, Kitani KM (2014) Action-reaction: Forecasting the dynamics of human interaction. In: ECCV

Huang DA, Ramanathan V, Mahajan D, Torresani L, Paluri M, Fei-Fei L, Niebles JC (2018) What makes a video a video: Analyzing temporal information in video understanding models and datasets. In: CVPR

Huang PY, Xu H, Li J, Baevski A, Auli M, Galuba W, Metze F, Feichtenhofer C (2022) Masked autoencoders that listen. NeurIPS

Huang PY, Sharma V, Xu H, Ryali C, Li Y, Li SW, Ghosh G, Malik J, Feichtenhofer C, et al (2023) Mavil: Masked audio-video learners. In: NeurIPS

Huang Y, Dai Q, Lu Y (2019) Decoupling localization and classification in single shot temporal action detection. In: ICME

Huang Y, Zhang Y, Elachqar O, Cheng Y (2020b) Inset: Sentence infilling with inter-sentential transformer. In: ACL

Huh J, Chalk J, Kazakos E, Damen D, Zisserman A (2023) Epic-sounds: A large-scale dataset of actions that sound. In: ICASSP

Hussain Z, Sheng M, Zhang WE (2019) Different approaches for human activity recognition: A survey. arxiv

Hussein N, Gavves E, Smeulders AW (2019) Timeception for complex action recognition. In: CVPR

Hwang JJ, Ke TW, Shi J, Yu SX (2019) Adversarial structure matching for structured prediction tasks. In: CVPR

Iashin V, Rahtu E (2020a) A better use of audio-visual cues: Dense video captioning with bi-modal transformer. In: BMVC

Iashin V, Rahtu E (2020b) Multi-modal dense video captioning. In: CVPRw

Ibrahim MS, Muralidharan S, Deng Z, Vahdat A, Mori G (2016) A hierarchical deep temporal model for group activity recognition. In: CVPR

Ikizler-Cinbis N, Sclaroff S (2010) Object, scene and actions: Combining multiple features for human action recognition. In: ECCV

Ionescu C, Papava D, Olaru V, Sminchisescu C (2013) Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. IEEE TPAMI

Iosifidis A, Tefas A, Pitas I (2012) View-invariant action recognition based on artificial neural networks. IEEE TNNLS

Ippolito D, Grangier D, Callison-Burch C, Eck D (2019) Unsupervised hierarchical story infilling. In: WNU

Isard M, Blake A (1998) Condensation—conditional density propagation for visual tracking. IJCV
Islam MM, Ho N, Yang X, Nagarajan T, Torresani L, Bertasius G (2024) Video recap: Recursive captioning of hour-long videos. In: CVPR
Jaegle A, Gimeno F, Brock A, Vinyals O, Zisserman A, Carreira J (2021) Perceiver: General perception with iterative attention. In: ICML
Jain A, Tompson J, LeCun Y, Bregler C (2015) Modeep: A deep learning framework using motion features for human pose estimation. In: ACCV
Jain M, Van Gemert J, Jégou H, Bouthemy P, Snoek CG (2014) Action localization with tubelets from motion. In: CVPR
Jang Y, Song Y, Yu Y, Kim Y, Kim G (2017) Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: CVPR
Jeannerod M (1994) The representing brain: Neural correlates of motor intention and imagery. BBS
Jhuang H, Gall J, Zuffi S, Schmid C, Black MJ (2013) Towards understanding action recognition. In: ICCV
Ji J, Krishna R, Fei-Fei L, Niebles JC (2020) Action genome: Actions as compositions of spatio-temporal scene graphs. In: CVPR
Ji S, Xu W, Yang M, Yu K (2012) 3d convolutional neural networks for human action recognition. IEEE TPAMI
Jia K, Yeung DY (2008) Human action recognition using local spatio-temporal discriminant embedding. In: CVPR
Jiang B, Huang X, Yang C, Yuan J (2019a) Cross-modal video moment retrieval with spatial and language-temporal attention. In: ICMR
Jiang B, Wang M, Gan W, Wu W, Yan J (2019b) Stm: Spatiotemporal and motion encoding for action recognition. In: ICCV
Jiang B, Chen X, Liu W, Yu J, Yu G, Chen T (2023) Motiongpt: Human motion as a foreign language. NeurIPS
Jiang J, Nan Z, Chen H, Chen S, Zheng N (2021) Predicting short-term next-active-object through visual attention and hand position. Neurocomputing
Jiang P, Han Y (2020) Reasoning with heterogeneous graph alignment for video question answering. In: AAAI
Jiang YG, Ye G, Chang SF, Ellis D, Loui AC (2011) Consumer video understanding: A benchmark database and an evaluation of human and machine performance. In: ICMR
Jin B, Hu Y, Tang Q, Niu J, Shi Z, Han Y, Li X (2020) Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction. In: CVPR
Jin X, Li X, Xiao H, Shen X, Lin Z, Yang J, Chen Y, Dong J, Liu L, Jie Z, et al (2017) Video scene parsing with predictive feature learning. In: ICCV
Joo H, Simon T, Li X, Liu H, Tan L, Gui L, Banerjee S, Godisart TS, Nabbe B, Matthews I, Kanade T, Nobuhara S, Sheikh Y (2017) Panoptic studio: A massively multiview system for social interaction capture. IEEE TPAMI
Ju C, Zheng K, Liu J, Zhao P, Zhang Y, Chang J, Tian Q, Wang Y (2023) Distilling vision-language pre-training to collaborate with weakly-supervised temporal action localization. In: CVPR
Junejo IN, Dexter E, Laptev I, Perez P (2010) View-independent action recognition from temporal self-similarities. IEEE TPAMI
Kahatapitiya K, Arnab A, Nagrani A, Ryoo MS (2024) Victr: Video-conditioned text representations for activity recognition. In: CVPR
Kaiser L, Gomez AN, Shazeer N, Vaswani A, Parmar N, Jones L, Uszkoreit J (2017) One model to learn them all. arxiv
Kalogeiton V, Weinzaepfel P, Ferrari V, Schmid C (2017) Action tubelet detector for spatio-temporal action localization. In: ICCV
Karpathy A, Toderici G, Shetty S, Leung T, Sukthankar R, Fei-Fei L (2014) Large-scale video classification with convolutional neural networks. In: CVPR
Kataoka H, Miyashita Y, Hayashi M, Iwata K, Satoh Y (2016) Recognition of transitional action for short-term action prediction using discriminative temporal cnn feature. In: BMVC
Kay W, Carreira J, Simonyan K, Zhang B, Hillier C, Vijayanarasimhan S, Viola F, Green T, Back T, Natsev P, et al (2017) The kinetics human action video dataset. arxiv
Kazakos E, Nagrani A, Zisserman A, Damen D (2021) Slow-fast auditory streams for audio recognition. In: ICASSP
Ke Q, Fritz M, Schiele B (2019) Time-conditioned action anticipation in one shot. In: CVPR
Ke Y, Sukthankar R, Hebert M (2007) Spatio-temporal shape and flow correlation for action recognition. In: CVPR
Kilner JM (2011) More than one pathway to action understanding. Trends in cognitive sciences
Kim B, Lee J, Kang J, Kim ES, Kim HJ (2021a) Hotr: End-to-end human-object interaction detection with transformers. In: CVPR
Kim H, Jain M, Lee JT, Yun S, Porikli F (2021b) Efficient action recognition via dynamic knowledge propagation. In: ICCV
Kim M, Kwon H, Wang C, Kwak S, Cho M (2021c) Relational self-attention: What's missing in attention for video understanding. NeurIPS
Kim M, Gao S, Hsu YC, Shen Y, Jin H (2024a) Token fusion: Bridging the gap between token pruning and token merging. In: WACV
Kim M, Kim HB, Moon J, Choi J, Kim ST (2024b) Do you remember? dense video captioning with cross-modal memory retrieval. In: CVPR
Kitani KM, Ziebart BD, Bagnell JA, Hebert M (2012) Activity forecasting. In: ECCV
Ko D, Choi J, Ko J, Noh S, On KW, Kim ES, Kim HJ (2022) Video-text representation learning via differentiable weak temporal alignment. In: CVPR
Kohler E, Keysers C, Umilta MA, Fogassi L, Gallese V, Rizzolatti G (2002) Hearing sounds, understanding actions: action representation in mirror neurons. Science
Kondratyuk D, Yuan L, Li Y, Zhang L, Tan M, Brown M, Gong B (2021) Movinets: Mobile video networks for efficient video recognition. In: CVPR
Kong Q, Cao Y, Iqbal T, Wang Y, Wang W, Plumbley MD (2020) Panns: Large-scale pretrained audio neural networks for audio pattern recognition. IEEE/ACM TASLP
Kong Y, Fu Y (2022) Human action recognition and prediction: A survey. IJCV
Kong Y, Kit D, Fu Y (2014) A discriminative model with multiple temporal scales for action prediction. In: ECCV
Kong Y, Gao S, Sun B, Fu Y (2018) Action prediction from videos via memorizing hard-to-predict samples. In: AAAI
Koppula HS, Saxena A (2015) Anticipating human activities using object affordances for reactive robotic response. IEEE TPAMI
Koppula HS, Gupta R, Saxena A (2013) Learning human activities and object affordances from rgb-d videos. IJRR
Korbar B, Tran D, Torresani L (2019) Scsampler: Sampling salient clips from video for efficient action recognition. In: ICCV

Körner M, Denzler J (2013) Temporal self-similarity for appearance-based action recognition in multi-view setups. In: CAIP
Koutini K, Schlüter J, Eghbal-Zadeh H, Widmer G (2022) Efficient training of audio transformers with patchout. In: Interspeech
Krishna R, Hata K, Ren F, Fei-Fei L, Carlos Niebles J (2017) Dense-captioning events in videos. In: ICCV
Kuehne H, Jhuang H, Garrote E, Poggio T, Serre T (2011) Hmdb: a large video database for human motion recognition. In: ICCV
Kuehne H, Arslan A, Serre T (2014) The language of actions: Recovering the syntax and semantics of goal-directed human activities. In: CVPR
Kumar A, Rawat YS (2022) End-to-end semi-supervised learning for video action detection. In: CVPR
Kuo W, Piergiovanni A, Kim D, Luo X, Caine B, Li W, Ogale A, Zhou L, Dai A, Chen Z, et al (2023) Mammut: A simple architecture for joint learning for multimodal tasks. TMLR
Kviatkovsky I, Rivlin E, Shimshoni I (2014) Online action recognition using covariance of shape and motion. CVIU
Kwon T, Tekin B, Stühmer J, Bogo F, Pollefeys M (2021) H2o: Two hands manipulating objects for first person interaction recognition. In: ICCV
Laptev I, Lindeberg T (2003) Space-time interest points. In: ICCV
Laptev I, Pérez P (2007) Retrieving actions in movies. In: ICCV
Laptev I, Marszalek M, Schmid C, Rozenfeld B (2008) Learning realistic human actions from movies. In: CVPR
Le QV, Zou WY, Yeung SY, Ng AY (2011) Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: CVPR
Lee H, Battle A, Raina R, Ng A (2006) Efficient sparse coding algorithms. NeurIPS
Lee J, Lee Y, Kim J, Kosiorek A, Choi S, Teh YW (2019) Set transformer: A framework for attention-based permutation-invariant neural networks. In: ICML
Lei J, Yu L, Bansal M, Berg TL (2018) Tvqa: Localized, compositional video question answering. arxiv
Lei J, Berg TL, Bansal M (2021a) Detecting moments and highlights in videos via natural language queries. NeurIPS
Lei J, Li L, Zhou L, Gan Z, Berg TL, Bansal M, Liu J (2021b) Less is more: Clipbert for video-and-language learning via sparse sampling. In: CVPR
Li D, Qiu Z, Dai Q, Yao T, Mei T (2018a) Recurrent tubelet proposal and recognition networks for action detection. In: ECCV
Li D, Li J, Li H, Niebles JC, Hoi SC (2022a) Align and prompt: Video-and-language pre-training with entity prompts. In: CVPR
Li G, Cai G, Zeng X, Zhao R (2022b) Scale-aware spatio-temporal relation learning for video anomaly detection. In: ECCV
Li J, Li D, Savarese S, Hoi S (2023a) Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML
Li J, Wei P, Han W, Fan L (2023b) Intentqa: Context-aware video intent reasoning. In: CVPR
Li K, Fu Y (2014) Prediction of human activity by discovering temporal sequence patterns. IEEE TPAMI
Li K, Hu J, Fu Y (2012) Modeling complex temporal composition of actionlets for activity prediction. In: ECCV
Li K, Wang Y, Peng G, Song G, Liu Y, Li H, Qiao Y (2022c) Uniformer: Unified transformer for efficient spatial-temporal representation learning. In: ICLR
Li K, He Y, Wang Y, Li Y, Wang W, Luo P, Wang Y, Wang L, Qiao Y (2023c) Videochat: Chat-centric video understanding. arxiv
Li K, Wang Y, He Y, Li Y, Wang Y, Liu Y, Wang Z, Xu J, Chen G, Luo P, et al (2024a) Mvbench: A comprehensive multi-modal video understanding benchmark. In: CVPR
Li L, Chen YC, Cheng Y, Gan Z, Yu L, Liu J (2020a) Hero: Hierarchical encoder for video+ language omni-representation pre-training. In: EMNLP
Li T, Wang Z, Liu S, Lin WY (2021a) Deep unsupervised anomaly detection. In: WACV
Li T, Slavcheva M, Zollhoefer M, Green S, Lassner C, Kim C, Schmidt T, Lovegrove S, Goesele M, Newcombe R, et al (2022d) Neural 3d video synthesis from multi-view video. In: CVPR
Li W, Fritz M (2016) Recognition of ongoing complex activities by sequence prediction over a hierarchical label space. In: WACV
Li X, Xu H (2024) Repetitive Action Counting With Motion Feature Learning. In: WACV
Li X, Song J, Gao L, Liu X, Huang W, He X, Gan C (2019) Beyond rnns: Positional self-attention with co-attention for video question answering. In: AAAI
Li Y, Ye Z, Rehg JM (2015) Delving into egocentric actions. In: CVPR
Li Y, Li Y, Vasconcelos N (2018b) Resound: Towards action recognition without representation bias. In: ECCV
Li Y, Yao T, Pan Y, Chao H, Mei T (2018c) Jointly localizing and describing events for dense video captioning. In: CVPR
Li Y, Wang Z, Wang L, Wu G (2020b) Actions as moving points. In: ECCV
Li Y, Chen L, He R, Wang Z, Wu G, Wang L (2021b) Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In: ICCV
Li Y, Wang X, Xiao J, Chua TS (2022e) Equivariant and invariant grounding for video question answering. In: MM
Li Y, Wu CY, Fan H, Mangalam K, Xiong B, Malik J, Feichtenhofer C (2022f) Mvitv2: Improved multiscale vision transformers for classification and detection. In: CVPR
Li Y, Xiao J, Feng C, Wang X, Chua TS (2023d) Discovering spatio-temporal rationales for video question answering. In: ICCV
Li Z, Ma X, Shang Q, Zhu W, Ci H, Qiao Y, Wang Y (2024b) Efficient action counting with dynamic queries. arxiv
Liang C, Wang W, Zhou T, Yang Y (2022) Visual abductive reasoning. In: CVPR
Liang X, Lee L, Dai W, Xing EP (2017) Dual motion gan for future-flow embedded video prediction. In: ICCV
Lin C, Xu C, Luo D, Wang Y, Tai Y, Wang C, Li J, Huang F, Fu Y (2021a) Learning salient boundary feature for anchor-free temporal action localization. In: CVPR
Lin J, Gan C, Han S (2019) Tsm: Temporal shift module for efficient video understanding. In: ICCV
Lin KE, Xiao L, Liu F, Yang G, Ramamoorthi R (2021b) Deep 3d mask volume for view synthesis of dynamic scenes. In: ICCV
Lin KQ, Wang J, Soldan M, Wray M, Yan R, Xu EZ, Gao D, Tu RC, Zhao W, Kong W, et al (2022) Egocentric video-language pretraining. NeurIPS
Lin T, Zhao X, Su H, Wang C, Yang M (2018) Bsn: Boundary sensitive network for temporal action proposal generation. In: ECCV
Lin YB, Bertasius G (2024) Siamese vision transformers are scalable audio-visual learners. arxiv

Lin YB, Sung YL, Lei J, Bansal M, Bertasius G (2023) Vision transformers are parameter-efficient audio-visual learners. In: CVPR
Lipman Y, Chen RT, Ben-Hamu H, Nickel M, Le M (2022) Flow matching for generative modeling. arxiv
Liu D, Qu X, Dong J, Zhou P, Cheng Y, Wei W, Xu Z, Xie Y (2021a) Context-aware biaffine localizing network for temporal sentence grounding. In: CVPR
Liu D, Qu X, Di X, Cheng Y, Xu Z, Zhou P (2022a) Memory-guided semantic learning network for temporal sentence grounding. In: AAAI
Liu F, Liu J, Wang W, Lu H (2021b) Hair: Hierarchical visual-semantic relational reasoning for video question answering. In: ICCV
Liu H, Liu X, Kong Q, Wang W, Plumbley MD (2022b) Learning the spectrogram temporal resolution for audio classification. In: AAAI
Liu H, Li C, Wu Q, Lee YJ (2024a) Visual instruction tuning. NeurIPS
Liu J, Shah M (2008) Learning human actions via information maximization. In: CVPR
Liu J, Ali S, Shah M (2008) Recognizing human actions using multiple features. In: CVPR
Liu J, Luo J, Shah M (2009) Recognizing realistic actions from videos "in the wild"'. In: CVPR
Liu M, Wang X, Nie L, Tian Q, Chen B, Chua TS (2018a) Cross-modal moment localization in videos. In: MM
Liu M, Tang S, Li Y, Rehg JM (2020) Forecasting human-object interaction: joint prediction of motor attention and actions in first person video. In: ECCV
Liu M, Zhang M, Liu J, Dai H, Yang MH, Ji S, Feng Z, Gong B (2023) Video timeline modeling for news story understanding. NeurIPS
Liu Q, Wang Z (2020) Progressive boundary refinement network for temporal action detection. In: AAAI
Liu S, Tripathi S, Majumdar S, Wang X (2022c) Joint hand motion and interaction hotspots prediction from egocentric videos. In: CVPR
Liu S, Zhang CL, Zhao C, Ghanem B (2024b) End-to-end temporal action detection with 1b parameters across 1000 frames. In: CVPR
Liu T, Lam KM (2022) A hybrid egocentric activity anticipation framework via memory-augmented recurrent and one-shot representation forecasting. In: CVPR
Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC (2016) Ssd: Single shot multibox detector. In: ECCV
Liu W, Luo W, Lian D, Gao S (2018b) Future frame prediction for anomaly detection–a new baseline. In: CVPR
Liu W, Tekin B, Coskun H, Vineet V, Fua P, Pollefeys M (2022d) Learning to align sequential actions in the wild. In: CVPR
Liu X, Bai S, Bai X (2022e) An empirical study of end-to-end temporal action detection. In: CVPR
Liu Y, Wei P, Zhu SC (2017a) Jointly recognizing object fluents and tasks in egocentric videos. In: ICCV
Liu Y, Albanie S, Nagrani A, Zisserman A (2019) Use what you have: Video retrieval using representations from collaborative experts. In: BMVC
Liu Y, Wang L, Wang Y, Ma X, Qiao Y (2022f) Fineaction: A fine-grained video dataset for temporal action localization. IEEE T-IP
Liu Y, Zhang K, Li Y, Yan Z, Gao C, Chen R, Yuan Z, Huang Y, Sun H, Gao J, et al (2024c) Sora: A review on background, technology, limitations, and opportunities of large vision models. arxiv
Liu Z, Yeh RA, Tang X, Liu Y, Agarwala A (2017b) Video frame synthesis using deep voxel flow. In: ICCV
Liu Z, Wang L, Tang W, Yuan J, Zheng N, Hua G (2021c) Weakly supervised temporal action localization through learning explicit subspaces for action and context. In: AAAI
Liu Z, Mao H, Wu CY, Feichtenhofer C, Darrell T, Xie S (2022g) A convnet for the 2020s. In: CVPR
Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, Hu H (2022h) Video swin transformer. In: CVPR
Lu C, Ferrier NJ (2004) Repetitive Motion Analysis: Segmentation and Event Classification. IEEE TPAMI
Lu C, Shi J, Jia J (2013) Abnormal event detection at 150 fps in matlab. In: ICCV
Luc P, Couprie C, Lecun Y, Verbeek J (2018) Predicting future instance segmentation by forecasting convolutional features. In: ECCV
Luc P, Clark A, Dieleman S, Casas DdL, Doron Y, Cassirer A, Simonyan K (2020) Transformation-based adversarial video prediction on large-scale data. arxiv
Luo C, Yuille AL (2019) Grouped spatial-temporal aggregation for efficient action recognition. In: ICCV
Luo W, Liu W, Gao S (2017) A revisit of sparse coding based anomaly detection in stacked rnn framework. In: ICCV
Luo Z, Guillory D, Shi B, Ke W, Wan F, Darrell T, Xu H (2020) Weakly-supervised action localization with expectation-maximization multi-instance learning. In: ECCV
Luo Z, Xie W, Kapoor S, Liang Y, Cooper M, Niebles JC, Adeli E, Li FF (2021) Moma: Multi-object multi-actor activity parsing. NeurIPS
Ma C, Guo Q, Jiang Y, Luo P, Yuan Z, Qi X (2022) Rethinking resolution in the context of efficient video recognition. NeurIPS
Ma S, Zeng Z, McDuff D, Song Y (2021) Active contrastive learning of audio-visual video representations. In: ICLR
Maaz M, Rasheed H, Khan S, Khan FS (2023) Video-chatgpt: Towards detailed video understanding via large vision and language models. arxiv
Mangalam K, Akshulakov R, Malik J (2023) Egoschema: A diagnostic benchmark for very long-form video language understanding. NeurIPS
Markovitz A, Sharir G, Friedman I, Zelnik-Manor L, Avidan S (2020) Graph embedded pose clustering for anomaly detection. In: CVPR
Marszalek M, Laptev I, Schmid C (2009) Actions in context. In: CVPR
Martin M, Roitberg A, Haurilet M, Horne M, Reiß S, Voit M, Stiefelhagen R (2019) Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In: ICCV
Mascaró EV, Ahn H, Lee D (2023) Intention-conditioned long-term human egocentric action anticipation. In: WACV
Matthews I, Cootes TF, Bangham JA, Cox S, Harvey R (2002) Extraction of visual features for lipreading. IEEE TPAMI
Mavroudi E, Afouras T, Torresani L (2023) Learning to ground instructional articles in videos through narrations. In: ICCV
Menapace W, Lathuiliere S, Tulyakov S, Siarohin A, Ricci E (2021) Playable video generation. In: CVPR
Meng Y, Lin CC, Panda R, Sattigeri P, Karlinsky L, Oliva A, Saenko K, Feris R (2020) Ar-net: Adaptive frame resolution for efficient action recognition. In: ECCV

Metaxas D, Zhang S (2013) A review of motion analysis methods for human nonverbal communication computing. IVC
Mettes P, Van Gemert JC, Snoek CG (2016) Spot on: Action localization from pointly-supervised proposals. In: ECCV
Micorek J, Possegger H, Narnhofer D, Bischof H, Kozinski M (2024) Mulde: Multiscale log-density estimation via denoising score matching for video anomaly detection. In: CVPR
Miech A, Zhukov D, Alayrac JB, Tapaswi M, Laptev I, Sivic J (2019) Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In: CVPR
Miech A, Alayrac JB, Laptev I, Sivic J, Zisserman A (2020a) Rareact: A video dataset of unusual interactions. arxiv
Miech A, Alayrac JB, Smaira L, Laptev I, Sivic J, Zisserman A (2020b) End-to-end learning of visual representations from uncurated instructional videos. In: CVPR
Mikolajczyk K, Uemura H (2008) Action recognition with motion-appearance vocabulary forest. In: CVPR
Min J, Buch S, Nagrani A, Cho M, Schmid C (2024) Morevqa: Exploring modular reasoning models for video question answering. In: CVPR
Misra I, Zitnick CL, Hebert M (2016) Shuffle and learn: unsupervised learning using temporal order verification. In: ECCV
Mithun NC, Li J, Metze F, Roy-Chowdhury AK (2018) Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In: ICMR
Mittal H, Agarwal N, Lo SY, Lee K (2024) Can't make an omelette without breaking some eggs: Plausible action anticipation using large video-language models. In: CVPR
Mo S, Morgado P (2023) A unified audio-visual learning framework for localization, separation, and recognition. In: ICML
Moeslund TB, Granum E (2001) A survey of computer vision-based human motion capture. CVIU
Moeslund TB, Hilton A, Krüger V (2006) A survey of advances in vision-based human motion capture and analysis. CVIU
Moltisanti D, Fidler S, Damen D (2019) Action recognition from single timestamp supervision in untrimmed videos. In: CVPR
Moltisanti D, Keller F, Bilen H, Sevilla-Lara L (2023) Learning action changes by measuring verb-adverb textual relationships. In: CVPR
Monfort M, Andonian A, Zhou B, Ramakrishnan K, Bargal SA, Yan T, Brown L, Fan Q, Gutfreund D, Vondrick C, et al (2019) Moments in time dataset: one million videos for event understanding. IEEE TPAMI
Monfort M, Jin S, Liu A, Harwath D, Feris R, Glass J, Oliva A (2021) Spoken moments: Learning joint audio-visual representations from video descriptions. In: CVPR
Moon G, Yu SI, Wen H, Shiratori T, Lee KM (2020) Interhand2. 6m: A dataset and baseline for 3d interacting hand pose estimation from a single rgb image. In: ECCV
Morais R, Le V, Tran T, Saha B, Mansour M, Venkatesh S (2019) Learning regularity in skeleton trajectories for anomaly detection in videos. In: CVPR
Morgado P, Vasconcelos N, Misra I (2021) Audio-visual instance discrimination with cross-modal agreement. In: CVPR
Morrongiello BA, Fenwick KD, Nutley T (1998) Developmental changes in associations between auditory-visual events. Infant Behavior and Development
Mounir R, Vijayaraghavan S, Sarkar S (2024) Streamer: Streaming representation learning and event segmentation in a hierarchical manner. NeurIPS
Mueller F, Mehta D, Sotnychenko O, Sridhar S, Casas D, Theobalt C (2017) Real-time hand tracking under occlusion from an egocentric rgb-d sensor. In: CVPR
Mun J, Yang L, Ren Z, Xu N, Han B (2019) Streamlined dense video captioning. In: CVPR
Munro J, Damen D (2020) Multi-modal domain adaptation for fine-grained action recognition. In: CVPR
Mur-Labadia L, Martinez-Cantin R, Guerrero J, Farinella GM, Furnari A (2024) Aff-ttention! affordances and attention models for short-term object interaction anticipation. arxiv arXiv:240601194
Nag S, Zhu X, Deng J, Song YZ, Xiang T (2023) Difftad: Temporal action detection with proposal denoising diffusion. In: ICCV
Nagarajan T, Grauman K (2018) Attributes as operators: factorizing unseen attribute-object compositions. In: ECCV
Nagarajan T, Feichtenhofer C, Grauman K (2019) Grounded human-object interaction hotspots from video. In: CVPR
Nagrani A, Yang S, Arnab A, Jansen A, Schmid C, Sun C (2021) Attention bottlenecks for multimodal fusion. In: NeurIPS
Nan G, Qiao R, Xiao Y, Liu J, Leng S, Zhang H, Lu W (2021) Interventional video grounding with dual contrastive learning. In: CVPR
Nawhal M, Jyothi AA, Mori G (2022) Rethinking learning approaches for long-term action anticipation. In: ECCV
Nguyen TN, Meunier J (2019) Anomaly detection in video sequence with appearance-motion correspondence. In: ICCV
Nguyen TT, Nguyen P, Luu K (2024) Hig: Hierarchical interlacement graph approach to scene graph generation in video understanding. In: CVPR
Ni B, Paramathayalan VR, Moulin P (2014) Multiple granularity analysis for fine-grained action detection. In: CVPR
Nie X, Chen X, Jin H, Zhu Z, Yan Y, Qi D (2024) Triplet attention transformer for spatiotemporal predictive learning. In: WACV
Niebles JC, Wang H, Fei-Fei L (2008) Unsupervised learning of human action categories using spatial-temporal words. IJCV
Niebles JC, Chen CW, Fei-Fei L (2010) Modeling temporal structure of decomposable motion segments for activity classification. In: ECCV
Nikankin Y, Haim N, Irani M (2023) Sinfusion: training diffusion models on a single image or video. In: ICML
Nowozin S, Bakir G, Tsuda K (2007) Discriminative subsequence mining for action classification. In: ICCV
Ntinou I, Sanchez E, Tzimiropoulos G (2024) Multiscale vision transformers meet bipartite matching for efficient single-stage action localization. In: CVPR
Nugroho MA, Woo S, Lee S, Kim C (2023) Audio-visual glance network for efficient video recognition. In: ICCV
Ohkawa T, He K, Sener F, Hodan T, Tran L, Keskin C (2023) AssemblyHands: towards egocentric activity understanding via 3d hand pose estimation. In: CVPR
Oikonomopoulos A, Patras I, Pantic M (2005) Spatiotemporal saliency for human action recognition. In: ICME
Oncescu AM, Henriques JF, Liu Y, Zisserman A, Albanie S (2021) Queryd: A video dataset with high-quality text and audio narrations. In: ICASSP
Oneata D, Verbeek J, Schmid C (2013) Action and event recognition with fisher vectors on a compact feature set. In: ICCV
Ong KE, Ng XL, Li Y, Ai W, Zhao K, Yeo SY, Liu J (2023) Chaotic world: A large and challenging benchmark for human behavior understanding in chaotic events. In: ICCV

Oprea S, Martinez-Gonzalez P, Garcia-Garcia A, Castro-Vargas JA, Orts-Escolano S, Garcia-Rodriguez J, Argyros A (2022) A review on deep learning techniques for video prediction. IEEE TPAMI

Ortega JD, Kose N, Cañas P, Chao MA, Unnervik A, Nieto M, Otaegui O, Salgado L (2020) Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis. In: ECCV

Otani M, Nakashima Y, Rahtu E, Heikkilä J, Yokoya N (2016) Learning joint representations of videos and sentences with web image search. In: ECCVw

Owens A, Isola P, McDermott J, Torralba A, Adelson EH, Freeman WT (2016) Visually indicated sounds. In: CVPR

Pan J, Chen S, Shou MZ, Liu Y, Shao J, Li H (2021) Actor-context-actor relation network for spatio-temporal action localization. In: CVPR

Panagiotakis C, Karvounas G, Argyros A (2018) Unsupervised Detection of Periodic Segments in Videos. In: ICIP

Pareek P, Thakkar A (2021) A survey on video-based human action recognition: recent updates, datasets, challenges, and applications. UMT-AIR

Park H, Noh J, Ham B (2020) Learning memory-guided normality for anomaly detection. In: CVPR

Park J, Lee J, Sohn K (2021a) Bridge to answer: Structure-aware graph interaction network for video question answering. In: CVPR

Park JS, Shen S, Farhadi A, Darrell T, Choi Y, Rohrbach A (2022) Exposing the limits of video-text models through contrast sets. In: NAACL

Park S, Kim K, Lee J, Choo J, Lee J, Kim S, Choi E (2021b) Vid-ode: Continuous-time video generation with neural ordinary differential equation. In: AAAI

Park W, Kim D, Lu Y, Cho M (2019) Relational knowledge distillation. In: CVPR

Parmar P, Morris BT (2019) What and how well you performed? a multitask learning approach to action quality assessment. In: CVPR

Patron-Perez A, Marszalek M, Zisserman A, Reid I (2010) High five: Recognising human interactions in tv shows. In: BMVC

Patsch C, Zhang J, Wu Y, Zakour M, Salihu D, Steinbach E (2024) Long-term action anticipation based on contextual alignment. In: ICASSP

Paul S, Roy S, Roy-Chowdhury AK (2018) W-talc: Weakly-supervised temporal activity localization and classification. In: ECCV

Pei M, Jia Y, Zhu SC (2011) Parsing video events with goal inference and intent prediction. In: ICCV

Peng S, Zhang Y, Xu Y, Wang Q, Shuai Q, Bao H, Zhou X (2021) Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In: CVPR

Peng X, Schmid C (2016) Multi-region two-stream r-cnn for action detection. In: ECCV

Pian W, Mo S, Guo Y, Tian Y (2023) Audio-visual class-incremental learning. In: ICCV

Pickup LC, Pan Z, Wei D, Shih Y, Zhang C, Zisserman A, Scholkopf B, Freeman WT (2014) Seeing the arrow of time. In: CVPR

Piergiovanni A, Ryoo M (2020) Avid dataset: Anonymized videos from diverse countries. NeurIPS

Piergiovanni A, Angelova A, Toshev A, Ryoo MS (2020) Adversarial generative grammars for human activity prediction. In: ECCV

Piergiovanni A, Noble I, Kim D, Ryoo MS, Gomes V, Angelova A (2024) Mirasol3b: A multimodal autoregressive model for time-aligned and contextual modalities. In: CVPR

Pirsiavash H, Ramanan D (2012) Detecting activities of daily living in first-person camera views. In: CVPR

Pishchulin L, Andriluka M, Gehler P, Schiele B (2013) Strong appearance and expressive spatial models for human pose estimation. In: CVPR

Plizzari C, Goletto G, Furnari A, Bansal S, Ragusa F, Farinella GM, Damen D, Tommasi T (2024) An outlook into the future of egocentric vision. IJCV

Pogalin E, Smeulders AW, Thean AH (2008) Visual Quasi-Periodicity. In: CVPR

Poppe R (2007) Vision-based human motion analysis: An overview. CVIU

Poppe R (2010) A survey on vision-based human action recognition. IVC

Price W, Vondrick C, Damen D (2022) Unweavenet: Unweaving activity stories. In: CVPR

Pu Y, Wu X, Yang L, Wang S (2023) Learning prompt-enhanced context features for weakly-supervised video anomaly detection. arxiv

Purwanto D, Chen YT, Fang WH (2021) Dance with self-attention: A new look of conditional random fields on anomaly detection in videos. In: ICCV

Qian L, Li J, Wu Y, Ye Y, Fei H, Chua TS, Zhuang Y, Tang S (2024) Momentor: Advancing video large language model with fine-grained temporal reasoning. In: ICML

Qing Z, Su H, Gan W, Wang D, Wu W, Wang X, Qiao Y, Yan J, Gao C, Sang N (2021) Temporal context aggregation network for temporal action proposal refinement. In: CVPR

Qiu Z, Yao T, Mei T (2017) Learning spatio-temporal representation with pseudo-3d residual networks. In: ICCV

Qiu Z, Yao T, Ngo CW, Tian X, Mei T (2019) Learning spatio-temporal representation with local and global diffusion. In: CVPR

Qu X, Tang P, Zou Z, Cheng Y, Dong J, Zhou P, Xu Z (2020) Fine-grained iterative attention network for temporal language localization in videos. In: MM

Radevski G, Grujicic D, Blaschko M, Moens MF, Tuytelaars T (2023) Multimodal distillation for egocentric action recognition. In: ICCV

Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, et al (2021) Learning transferable visual models from natural language supervision. In: ICLR

Ragusa F, Furnari A, Livatino S, Farinella GM (2021) The meccano dataset: Understanding human-object interactions from egocentric videos in an industrial-like domain. In: WACV

Rahman T, Xu B, Sigal L (2019) Watch, listen and tell: Multi-modal weakly supervised dense event captioning. In: CVPR

Rahmani H, Mahmood A, Huynh DQ, Mian A (2014) Real time action recognition using histograms of depth gradients and random decision forests. In: WACV

Rai N, Chen H, Ji J, Desai R, Kozuka K, Ishizaka S, Adeli E, Niebles JC (2021) Home action genome: Cooperative compositional action understanding. In: CVPR

Ramachandra B, Jones MJ, Vatsavai RR (2020) A survey of single-scene video anomaly detection. IEEE TPAMI

Randall M (2009) Movie narrative charts. URL https://xkcd.com/657/

Rangrej SB, Liang KJ, Hassner T, Clark JJ (2023) Glitr: Glimpse transformers with spatiotemporal consistency for online action prediction. In: WACV

Rasouli A (2020) Deep learning for vision-based prediction: A survey. arxiv
Recasens A, Lin J, Carreira J, Jaegle D, Wang L, Alayrac Jb, Luc P, Miech A, Smaira L, Hemsley R, et al (2023) Zorro: the masked multimodal transformer. arxiv
Reddy KK, Shah M (2013) Recognizing 50 human action categories of web videos. MVA
Redmon J, Divvala S, Girshick R, Farhadi A (2016) You only look once: Unified, real-time object detection. In: CVPR
Regneri M, Rohrbach M, Wetzel D, Thater S, Schiele B, Pinkal M (2013) Grounding action descriptions in videos. TACL
Ren S, Yao L, Li S, Sun X, Hou L (2024) Timechat: A time-sensitive multimodal large language model for long video understanding. In: CVPR
Rizve MN, Mittal G, Yu Y, Hall M, Sajeev S, Shah M, Chen M (2023) Pivotal: Prior-driven supervision for weakly-supervised temporal action localization. In: CVPR
Rizzolatti G, Fogassi L, Gallese V (2001) Neurophysiological mechanisms underlying the understanding and imitation of action. Nature reviews neuroscience
Rodin I, Furnari A, Mavroeidis D, Farinella GM (2021) Predicting the future from first person (egocentric) vision: A survey. CVIU
Rodriguez MD, Ahmed J, Shah M (2008) Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: CVPR
Rohr K (1994) Towards model-based recognition of human movements in image sequences. CVGIP
Rohrbach M, Amin S, Andriluka M, Schiele B (2012) A database for fine grained activity detection of cooking activities. In: CVPR
Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: CVPR
Roy D, Fernando B (2021) Action anticipation using pairwise human-object interactions and transformers. IEEE T-IP
Roy D, Fernando B (2022) Action anticipation using latent goal learning. In: WACV
Roy D, Rajendiran R, Fernando B (2024) Interaction region visual transformer for egocentric action anticipation. In: WACV
Runia TF, Snoek CG, Smeulders AW (2018) Real-World Repetition Estimation by Div, Grad and Curl. In: CVPR
Ryali C, Hu YT, Bolya D, Wei C, Fan H, Huang PY, Aggarwal V, Chowdhury A, Poursaeed O, Hoffman J, et al (2023) Hiera: A hierarchical vision transformer without the bells-and-whistles. In: ICML
Ryoo M, Piergiovanni A, Arnab A, Dehghani M, Angelova A (2021) Tokenlearner: Adaptive space-time tokenization for videos. NeurIPS
Ryoo MS (2011) Human activity prediction: Early recognition of ongoing activities from streaming videos. In: ICCV
Ryoo MS, Aggarwal JK (2009) Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In: ICCV
Sadanand S, Corso JJ (2012) Action bank: A high-level representation of activity in video. In: CVPR
Saini N, Pham K, Shrivastava A (2022) Disentangling visual embeddings for attributes and objects. In: CVPR
Saini N, Wang H, Swaminathan A, Jayasundara V, He B, Gupta K, Shrivastava A (2023) Chop & learn: Recognizing and generating object-state compositions. In: ICCV
Saito M, Matsumoto E, Saito S (2017) Temporal generative adversarial nets with singular value clipping. In: ICCV
Sakoe H, Chiba S (1978) Dynamic programming algorithm optimization for spoken word recognition. IEEE TASSP
Sandvine I (2024) Global internet phenomena report. North America and Latin America
Schiappa MC, Rawat YS, Shah M (2023) Self-supervised learning for videos: A survey. CSUR
Schuldt C, Laptev I, Caputo B (2004) Recognizing human actions: a local svm approach. In: ICPR
Selva J, Johansen AS, Escalera S, Nasrollahi K, Moeslund TB, Clapés A (2023) Video transformers: A survey. IEEE TPAMI
Sener F, Chatterjee D, Shelepov D, He K, Singhania D, Wang R, Yao A (2022) Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: CVPR
Seo PH, Nagrani A, Schmid C (2021) Look before you speak: Visually contextualized utterances. In: CVPR
Seo PH, Nagrani A, Arnab A, Schmid C (2022) End-to-end generative pretraining for multimodal video captioning. In: CVPR
Sermanet P, Xu K, Levine S (2017) Unsupervised perceptual rewards for imitation learning. In: ICLRw
Sermanet P, Lynch C, Chebotar Y, Hsu J, Jang E, Schaal S, Levine S, Brain G (2018) Time-contrastive networks: Self-supervised learning from video. In: ICRA
Sevilla-Lara L, Liao Y, Güney F, Jampani V, Geiger A, Black MJ (2019) On the integration of optical flow and action recognition. In: GCPR
Shahroudy A, Liu J, Ng TT, Wang G (2016) Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In: CVPR
Shao D, Zhao Y, Dai B, Lin D (2020) Finegym: A hierarchical video dataset for fine-grained action understanding. In: CVPR
Shao J, Wang X, Quan R, Zheng J, Yang J, Yang Y (2023) Action sensitivity learning for temporal action localization. In: ICCV
Sharma S, Kiros R, Salakhutdinov R (2015) Action recognition using visual attention. In: ICLR
Shechtman E, Irani M (2005) Space-time behavior based correlation. In: CVPR
Sheikh Y, Sheikh M, Shah M (2005) Exploring the space of a human action. In: ICCV
Shen X, Li X, Elhoseiny M (2023) Mostgan-v: Video generation with temporal motion styles. In: CVPR
Shen Y, Elhamifar E (2024) Progress-aware online action segmentation for egocentric procedural task videos. In: CVPR
Shen Y, Ni B, Li Z, Zhuang N (2018) Egocentric activity prediction via event modulated attention. In: ECCV
Shen Z, Li J, Su Z, Li M, Chen Y, Jiang YG, Xue X (2017) Weakly supervised dense video captioning. In: CVPR
Shi B, Ji L, Liang Y, Duan N, Chen P, Niu Z, Zhou M (2019) Dense procedure captioning in narrated instructional videos. In: ACL
Shi D, Zhong Y, Cao Q, Ma L, Li J, Tao D (2023) Tridet: Temporal action detection with relative boundary modeling. In: CVPR
Shou MZ, Lei SW, Wang W, Ghadiyaram D, Feiszli M (2021) Generic event boundary detection: A benchmark for event segmentation. In: ICCV
Shou Z, Wang D, Chang SF (2016) Temporal action localization in untrimmed videos via multi-stage cnns. In: CVPR
Shou Z, Chan J, Zareian A, Miyazawa K, Chang SF (2017) Cdc: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos. In:

CVPR
Shou Z, Gao H, Zhang L, Miyazawa K, Chang SF (2018) Autoloc: Weakly-supervised temporal action localization in untrimmed videos. In: ECCV
Shrivastava G, Shrivastava A (2024) Video prediction by modeling videos as continuous multi-dimensional processes. In: CVPR
Sigal L, Balan AO, Black MJ (2010) Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. IJCV
Sigurdsson GA, Varol G, Wang X, Farhadi A, Laptev I, Gupta A (2016) Hollywood in homes: Crowdsourcing data collection for activity understanding. In: ECCV
Sigurdsson GA, Gupta A, Schmid C, Farhadi A, Alahari K (2018) Charades-ego: A large-scale dataset of paired third and first person videos. arxiv
Simonyan K, Zisserman A (2014) Two-stream convolutional networks for action recognition in videos. NeurIPS
Singer U, Polyak A, Hayes T, Yin X, An J, Zhang S, Hu Q, Yang H, Ashual O, Gafni O, et al (2023) Make-a-video: Text-to-video generation without text-video data. In: ICLR
Singh B, Marks TK, Jones M, Tuzel O, Shao M (2016) A multi-stream bi-directional recurrent neural network for fine-grained action detection. In: CVPR
Singh G, Saha S, Sapienza M, Torr PH, Cuzzolin F (2017) Online real-time multiple spatiotemporal action localisation and prediction. In: ICCV
Singh N, Wu CW, Orife I, Kalayeh M (2024) Looking similar sounding different: Leveraging counterfactual cross-modal pairs for audiovisual representation learning. In: CVPR
Sinha S, Stergiou A, Damen D (2024) Every shot counts: Using exemplars for repetition counting in videos. arxiv
Skorokhodov I, Tulyakov S, Elhoseiny M (2022) Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In: CVPR
Smaira L, Carreira J, Noland E, Clancy E, Wu A, Zisserman A (2020) A short note on the kinetics-700-2020 human action dataset. arxiv
Smith J, De Mello S, Kautz J, Linderman S, Byeon W (2024) Convolutional state space models for long-range spatiotemporal modeling. NeurIPS
Song E, Chai W, Wang G, Zhang Y, Zhou H, Wu F, Chi H, Guo X, Ye T, Zhang Y, et al (2024) Moviechat: From dense token to sparse memory for long video understanding. In: CVPR
Song L, Zhang S, Yu G, Sun H (2019) Tacnet: Transition-aware context network for spatio-temporal action detection. In: CVPR
Song L, Yu G, Yuan J, Liu Z (2021) Human pose estimation and its application to action recognition: A survey. JVCIR
Soomro K, Zamir AR, Shah M (2012) Ucf101: A dataset of 101 human actions classes from videos in the wild. arxiv
Soomro K, Idrees H, Shah M (2015) Action localization in videos through context walk. In: ICCV
Souček T, Alayrac JB, Miech A, Laptev I, Sivic J (2022) Look for the change: Learning object states and state-modifying actions from untrimmed web videos. In: CVPR
Souček T, Damen D, Wray M, Laptev I, Sivic J, et al (2024) Genhowto: Learning to generate actions and state transformations from instructional videos. In: CVPR
Spunt RP, Satpute AB, Lieberman MD (2011) Identifying the what, why, and how of an observed action: an fmri study of mentalizing and mechanizing during action observation. JCN
Srivastava N, Mansimov E, Salakhudinov R (2015) Unsupervised learning of video representations using lstms. In: ICML
Srivastava S, Sharma G (2024a) Omnivec: Learning robust representations with cross modal sharing. In: WACV
Srivastava S, Sharma G (2024b) Omnivec2-a novel transformer based network for large scale multimodal and multitask learning. In: CVPR
Stein S, McKenna SJ (2013) Combining embedded accelerometers with computer vision for recognizing food preparation activities. In: UbiComp
Stergiou A, Damen D (2023a) Play it back: Iterative attention for audio recognition. In: ICASSP
Stergiou A, Damen D (2023b) The wisdom of crowds: Temporal progressive attention for early action prediction. In: CVPR
Stergiou A, Deligiannis N (2023) Leaping into memories: Space-time deep feature synthesis. In: ICCV
Stergiou A, Poppe R (2019) Analyzing human–human interactions: A survey. CVIU
Stergiou A, Poppe R (2021a) Learn to cycle: Time-consistent feature discovery for action recognition. PRL
Stergiou A, Poppe R (2021b) Multi-temporal convolutions for human action recognition in videos. In: IJCNN
Stergiou A, De Weerdt B, Deligiannis N (2024) Holistic representation learning for multitask trajectory anomaly detection. In: WACV
Sudhakaran S, Escalera S, Lanz O (2020) Gate-shift networks for video action recognition. In: CVPR
Sultani W, Chen C, Shah M (2018) Real-world anomaly detection in surveillance videos. In: CVPR
Sun C, Shrivastava A, Vondrick C, Murphy K, Sukthankar R, Schmid C (2018) Actor-centric relation network. In: ECCV
Sun C, Myers A, Vondrick C, Murphy K, Schmid C (2019a) Videobert: A joint model for video and language representation learning. In: CVPR
Sun C, Shrivastava A, Vondrick C, Sukthankar R, Murphy K, Schmid C (2019b) Relational action forecasting. In: CVPR
Sun L, Jia K, Yeung DY, Shi BE (2015) Human action recognition using factorized spatio-temporal convolutional networks. In: ICCV
Sun P, Cao J, Jiang Y, Yuan Z, Bai S, Kitani K, Luo P (2022a) Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In: CVPR
Sun X, Chen M, Hauptmann A (2009) Action recognition via local descriptors and holistic features. In: CVPRw
Sun X, Panda R, Chen CFR, Oliva A, Feris R, Saenko K (2021) Dynamic network quantization for efficient video inference. In: ICCV
Sun Z, Ke Q, Rahmani H, Bennamoun M, Wang G, Liu J (2022b) Human action recognition from various data modalities: A review. IEEE TPAMI
Sung J, Ponce C, Selman B, Saxena A (2012) Unstructured human activity detection from rgbd images. In: ICRA
Surís D, Liu R, Vondrick C (2021) Learning the predictability of the future. In: CVPR
Tan C, Gao Z, Wu L, Xu Y, Xia J, Li S, Li SZ (2023a) Temporal attention unit: Towards efficient spatiotemporal predictive learning. In: CVPR
Tan J, Tang J, Wang L, Wu G (2021) Relaxed transformer decoders for direct action proposal generation. In: ICCV
Tan S, Nagarajan T, Grauman K (2023b) Egodistill: Egocentric head motion distillation for efficient video understanding. NeurIPS

Tang J, Xia J, Mu X, Pang B, Lu C (2020a) Asynchronous interaction aggregation for action detection. In: ECCV
Tang Y, Ding D, Rao Y, Zheng Y, Zhang D, Zhao L, Lu J, Zhou J (2019) Coin: A large-scale dataset for comprehensive instructional video analysis. In: CVPR
Tang Y, Ni Z, Zhou J, Zhang D, Lu J, Wu Y, Zhou J (2020b) Uncertainty-aware score distribution learning for action quality assessment. In: CVPR
Tang Y, Dong P, Tang Z, Chu X, Liang J (2024) Vmrnn: Integrating vision mamba and lstm for efficient and accurate spatiotemporal forecasting. In: CVPR
Taylor GW, Fergus R, LeCun Y, Bregler C (2010) Convolutional learning of spatio-temporal features. In: ECCV
Thakur S, Beyan C, Morerio P, Murino V, Del Bue A (2024) Anticipating next active objects for egocentric videos. IEEE Access
Thangali A, Sclaroff S (2005) Periodic motion detection and estimation via space-time sampling. In: WACV
Thoker FM, Doughty H, Snoek CGM (2023) Tubelet-contrastive self-supervision for video-efficient generalization. In: ICCV
Thompson EL, Bird G, Catmur C (2019) Conceptualizing and testing action understanding. NBR
Thurau C, Hlavác V (2008) Pose primitive based human action recognition in videos or still images. In: CVPR
Tian Y, Li D, Xu C (2020) Unified multisensory perception: Weakly-supervised audio-visual video parsing. In: ECCV
Tian Y, Pang G, Chen Y, Singh R, Verjans JW, Carneiro G (2021) Weakly-supervised video anomaly detection with robust temporal feature magnitude learning. In: ICCV
Torabi A, Tandon N, Sigal L (2016) Learning language-visual embedding for movie understanding with natural-language. arXiv
la Torre Frade FD, Hodgins JK, Bargteil AW, Artal XM, Macey JC, Castells ACI, Beltran J (2008) Guide to the carnegie mellon university multimodal activity (cmu-mmac) database. Tech. rep., CMU
Touvron H, Lavril T, Izacard G, Martinet X, Lachaux MA, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, et al (2023) Llama: Open and efficient foundation language models. arxiv
Tran D, Bourdev L, Fergus R, Torresani L, Paluri M (2015) Learning spatiotemporal features with 3d convolutional networks. In: ICCV
Tran D, Wang H, Torresani L, Ray J, LeCun Y, Paluri M (2018) A closer look at spatiotemporal convolutions for action recognition. In: CVPR
Tran D, Wang H, Torresani L, Feiszli M (2019) Video classification with channel-separated convolutional networks. In: ICCV
Tran KN, Kakadiaris IA, Shah SK (2012) Part-based motion descriptor image for human action recognition. PR
Tschernezki V, Darkhalil A, Zhu Z, Fouhey D, Laina I, Larlus D, Damen D, Vedaldi A (2024) Epic fields: Marrying 3d geometry and video understanding. NeurIPS
Tsuchida S, Fukayama S, Hamasaki M, Goto M (2019) Aist dance video database: Multi-genre, multi-dancer, and multi-camera database for dance information processing. In: ISMIR
Turaga P, Chellappa R, Subrahmanian VS, Udrea O (2008) Machine recognition of human activities: A survey. IEEE TCSVT
Uithol S, van Rooij I, Bekkering H, Haselager P (2011) Understanding motor resonance. Social neuroscience
Ullah A, Ahmad J, Muhammad K, Sajjad M, Baik SW (2017) Action recognition in video sequences using deep bi-directional lstm with cnn features. IEEE access
Ulutan O, Rallapalli S, Srivatsa M, Torres C, Manjunath B (2020) Actor conditioned attention maps for video action detection. In: WACV
Vaina LM, Jaulent MC (1991) Object structure and action requirements: A compatibility model for functional recognition. IJIS
Van Gemeren C, Poppe R, Veltkamp RC (2016) Spatio-temporal detection of fine-grained dyadic human interactions. In: HBU
Varol G, Laptev I, Schmid C (2017) Long-term temporal convolutions for action recognition. IEEE TPAMI
Villegas R, Yang J, Zou Y, Sohn S, Lin X, Lee H (2017) Learning to generate long-term future via hierarchical prediction. In: ICML
Villegas R, Erhan D, Lee H, et al (2018) Hierarchical long-term video prediction without supervision. In: ICML
Villegas R, Babaeizadeh M, Kindermans PJ, Moraldo H, Zhang H, Saffar MT, Castro S, Kunze J, Erhan D (2022) Phenaki: Variable length video generation from open domain textual descriptions. In: ICLR
Vishwakarma S, Agrawal A (2013) A survey on activity recognition and behavior understanding in video surveillance. TVC
Voleti V, Jolicoeur-Martineau A, Pal C (2022) Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. NeurIPS
Vondrick C, Pirsiavash H, Torralba A (2016a) Anticipating visual representations from unlabeled video. In: CVPR
Vondrick C, Pirsiavash H, Torralba A (2016b) Generating videos with scene dynamics. NeurIPS
Wang B, Zhao Y, Yang L, Long T, Li X (2023a) Temporal action localization in the deep learning era: A survey. IEEE TPAMI
Wang H, Schmid C (2013) Action recognition with improved trajectories. In: ICCV
Wang J, Cherian A (2019) Gods: Generalized one-class discriminative subspaces for anomaly detection. In: ICCV
Wang J, Jiang W, Ma L, Liu W, Xu Y (2018a) Bidirectional attentive fusion with context gating for dense video captioning. In: CVPR
Wang J, Jiao J, Liu YH (2020a) Self-supervised video representation learning by pace prediction. In: ECCV
Wang J, Ma L, Jiang W (2020b) Temporally grounding language queries in videos by contextual boundary-aware prediction. In: AAAI
Wang J, Gao Y, Li K, Hu J, Jiang X, Guo X, Ji R, Sun X (2021a) Enhancing unsupervised video representation learning by decoupling the scene and the motion. In: AAAI
Wang J, Ge Y, Cai G, Yan R, Lin X, Shan Y, Qie X, Shou MZ (2022a) Object-aware video-language pre-training for retrieval. In: CVPR
Wang J, Ge Y, Yan R, Ge Y, Lin KQ, Tsutsui S, Lin X, Cai G, Wu J, Shan Y, et al (2023b) All in one: Exploring unified video-language pre-training. In: CVPR
Wang J, Chen D, Luo C, He B, Yuan L, Wu Z, Jiang YG (2024a) Omnivid: A generative framework for universal video understanding. In: CVPR
Wang L, Li Y, Lazebnik S (2016a) Learning deep structure-preserving image-text embeddings. In: CVPR
Wang L, Xiong Y, Wang Z, Qiao Y, Lin D, Tang X, Van Gool L (2016b) Temporal segment networks: Towards good practices for deep action recognition. In: ECCV
Wang L, Xiong Y, Lin D, Van Gool L (2017a) Untrimmednets for weakly supervised action recognition and detection.

In: CVPR
Wang L, Li W, Li W, Van Gool L (2018b) Appearance-and-relation networks for video classification. In: CVPR
Wang L, Huang B, Zhao Z, Tong Z, He Y, Wang Y, Wang Y, Qiao Y (2023c) Videomae v2: Scaling video masked autoencoders with dual masking. In: CVPR
Wang M, Ni B, Yang X (2017b) Recurrent modeling of interaction context for collective activity recognition. In: CVPR
Wang P, Li W, Ogunbona P, Wan J, Escalera S (2018c) Rgb-d-based human motion recognition with deep learning: A survey. CVIU
Wang Q, Zhao L, Yuan L, Liu T, Peng X (2023d) Learning from semantic alignment between unpaired multiviews for egocentric video recognition. In: ICCV
Wang R, Chen D, Wu Z, Chen Y, Dai X, Liu M, Yuan L, Jiang YG (2023e) Masked video distillation: Rethinking masked feature modeling for self-supervised video representation learning. In: CVPR
Wang T, Zheng H, Yu M, Tian Q, Hu H (2020c) Event-centric hierarchical representation for dense video captioning. IEEE TCSVT
Wang T, Zhang R, Lu Z, Zheng F, Cheng R, Luo P (2021b) End-to-end dense video captioning with parallel decoding. In: ICCV
Wang W, Tran D, Feiszli M (2020d) What makes training multi-modal classification networks hard? In: CVPR
Wang W, Chang F, Zhang J, Yan R, Liu C, Wang B, Shou MZ (2023f) Magi-net: Meta negative network for early activity prediction. IEEE T-IP
Wang X, Gupta A (2018) Videos as space-time region graphs. In: ECCV
Wang X, Girshick R, Gupta A, He K (2018d) Non-local neural networks. In: CVPR
Wang X, Hu JF, Lai JH, Zhang J, Zheng WS (2019a) Progressive teacher-student learning for early action prediction. In: CVPR
Wang X, Wu J, Chen J, Li L, Wang YF, Wang WY (2019b) Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In: CVPR
Wang Y, Huang K, Tan T (2007) Human activity recognition based on r transform. In: CVPR
Wang Y, Long M, Wang J, Yu PS (2017c) Spatiotemporal pyramid network for video action recognition. In: CVPR
Wang Y, Gao Z, Long M, Wang J, Philip SY (2018e) Predrnn++: Towards a resolution of the deep-in-time dilemma in spatiotemporal predictive learning. In: ICML
Wang Y, Wu J, Long M, Tenenbaum JB (2020e) Probabilistic video prediction from noisy data with a posterior confidence. In: CVPR
Wang Y, Chen Z, Jiang H, Song S, Han Y, Huang G (2021c) Adaptive focus for efficient video recognition. In: ICCV
Wang Y, Yue Y, Lin Y, Jiang H, Lai Z, Kulikov V, Orlov N, Shi H, Huang G (2022b) Adafocus v2: End-to-end training of spatial dynamic networks for video recognition. In: CVPR
Wang Y, Yue Y, Xu X, Hassani A, Kulikov V, Orlov N, Song S, Shi H, Huang G (2022c) Adafocusv3: On unified spatial-temporal dynamic video recognition. In: ECCV
Wang Y, Li K, Li X, Yu J, He Y, Chen G, Pei B, Zheng R, Xu J, Wang Z, et al (2024b) Internvideo2: Scaling video foundation models for multimodal video understanding. arxiv
Wang Z, Zhong Y, Miao Y, Ma L, Specia L (2022d) Contrastive video-language learning with fine-grained frame sampling. arXiv preprint arXiv:221005039
Wei C, Fan H, Xie S, Wu CY, Yuille A, Feichtenhofer C (2022) Masked feature prediction for self-supervised visual pre-training. In: CVPR
Weinland D, Ronfard R, Boyer E (2011) A survey of vision-based methods for action representation, segmentation and recognition. CVIU
Weinzaepfel P, Harchaoui Z, Schmid C (2015) Learning to track for spatio-temporal action localization. In: ICCV
Weinzaepfel P, Martin X, Schmid C (2016) Towards weakly-supervised action localization. arxiv
Wong SF, Cipolla R (2007) Extracting spatiotemporal interest points using global information. In: ICCV
Wray M, Larlus D, Csurka G, Damen D (2019) Fine-grained action retrieval through multiple parts-of-speech embeddings. In: CVPR
Wray M, Doughty H, Damen D (2021) On semantic similarity in video retrieval. In: CVPR
Wu C, Zhang J, Savarese S, Saxena A (2015) Watch-n-patch: Unsupervised understanding of actions and relations. In: CVPR
Wu C, Huang L, Zhang Q, Li B, Ji L, Yang F, Sapiro G, Duan N (2021a) Godiva: Generating open-domain videos from natural descriptions. arxiv
Wu C, Liang J, Ji L, Yang F, Fang Y, Jiang D, Duan N (2022a) Nüwa: Visual synthesis pre-training for neural visual world creation. In: ECCV
Wu CY, Feichtenhofer C, Fan H, He K, Krahenbuhl P, Girshick R (2019a) Long-term feature banks for detailed video understanding. In: CVPR
Wu CY, Li Y, Mangalam K, Fan H, Xiong B, Malik J, Feichtenhofer C (2022b) Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In: CVPR
Wu H, Yao Z, Wang J, Long M (2021b) Motionrnn: A flexible model for video prediction with spacetime-varying motions. In: CVPR
Wu H, Chen K, Liu H, Zhuge M, Li B, Qiao R, Shu X, Gan B, Xu L, Ren B, et al (2023a) Newsnet: A novel dataset for hierarchical temporal segmentation. In: CVPR
Wu JZ, Ge Y, Wang X, Lei SW, Gu Y, Shi Y, Hsu W, Shan Y, Qie X, Shou MZ (2023b) Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In: CVPR
Wu P, Liu J, Shi Y, Sun Y, Shao F, Wu Z, Yang Z (2020a) Not only look, but also listen: Learning multimodal violence detection under weak supervision. In: ECCV
Wu Q, Cui R, Li Y, Zhu H (2024) Haltingvt: Adaptive token halting transformer for efficient video recognition. In: ICASSP
Wu T, Cao M, Gao Z, Wu G, Wang L (2023c) Stmixer: A one-stage sparse action detector. In: CVPR
Wu X, Wang R, Hou J, Lin H, Luo J (2021c) Spatial–temporal relation reasoning for action prediction in videos. IJCV
Wu X, Zhao J, Wang R (2021d) Anticipating future relations via graph growing for action prediction. In: AAAI
Wu Y, Yang Y (2021) Exploring heterogeneous clues for weakly-supervised audio-visual video parsing. In: CVPR
Wu Y, Zhu L, Wang X, Yang Y, Wu F (2020b) Learning to anticipate egocentric actions by imagination. IEEE T-IP
Wu Z, Xiong C, Jiang YG, Davis LS (2019b) Liteeval: A coarse-to-fine framework for resource efficient video recognition. NeurIPS
Wu Z, Xiong C, Ma CY, Socher R, Davis LS (2019c) Adaframe: Adaptive frame selection for fast video recognition. In: CVPR

Wu Z, Li H, Xiong C, Jiang YG, Davis LS (2020c) A dynamic frame selection framework for fast video recognition. IEEE TPAMI
Xia B, Wang Z, Wu W, Wang H, Han J (2022a) Temporal saliency query network for efficient video recognition. In: ECCV
Xia B, Wu W, Wang H, Su R, He D, Yang H, Fan X, Ouyang W (2022b) Nsnet: Non-saliency suppression sampler for efficient video recognition. In: ECCV
Xiao F, Lee YJ, Grauman K, Malik J, Feichtenhofer C (2020) Audiovisual slowfast networks for video recognition. arXiv
Xiao J, Shang X, Yao A, Chua TS (2021) Next-qa: Next phase of question-answering to explaining temporal actions. In: CVPR
Xiao J, Zhou P, Chua TS, Yan S (2022) Video graph transformer for video question answering. In: ECCV
Xiao J, Zhou P, Yao A, Li Y, Hong R, Yan S, Chua TS (2023) Contrastive video question answering via video graph transformer. IEEE TPAMI
Xiao J, Yao A, Li Y, Chua TS (2024) Can i trust your answer? visually grounded video question answering. In: CVPR
Xing Z, Dai Q, Hu H, Chen J, Wu Z, Jiang YG (2023) Svformer: Semi-supervised video transformer for action recognition. In: CVPR
Xiong Y, Zhao Y, Wang L, Lin D, Tang X (2017) A pursuit of temporal accuracy in general activity detection. arxiv
Xu D, Zhao Z, Xiao J, Wu F, Zhang H, He X, Zhuang Y (2017a) Video question answering via gradually refined attention over appearance and motion. In: MM
Xu D, Xiao J, Zhao Z, Shao J, Xie D, Zhuang Y (2019a) Self-supervised spatiotemporal learning via video clip order prediction. In: CVPR
Xu H, Das A, Saenko K (2017b) R-c3d: Region convolutional 3d network for temporal activity detection. In: ICCV
Xu H, He K, Plummer BA, Sigal L, Sclaroff S, Saenko K (2019b) Multilevel language and vision integration for text-to-clip retrieval. In: AAAI
Xu H, Ghosh G, Huang PY, Okhonko D, Aghajanyan A, Metze F, Zettlemoyer L, Feichtenhofer C (2021) Videoclip: Contrastive pre-training for zero-shot video-text understanding. In: EMNLP
Xu H, Ye Q, Yan M, Shi Y, Ye J, Xu Y, Li C, Bi B, Qian Q, Wang W, et al (2023a) mplug-2: A modularized multi-modal foundation model across text, image and video. In: ICML
Xu J, Mei T, Yao T, Rui Y (2016) Msr-vtt: A large video description dataset for bridging video and language. In: CVPR
Xu R, Xiong C, Chen W, Corso J (2015a) Jointly modeling deep video and compositional text to bridge vision and language in a unified framework. In: AAAI
Xu W, Yu J, Miao Z, Wan L, Ji Q (2019c) Prediction-cgan: Human action prediction with conditional generative adversarial networks. In: MM
Xu X, Li YL, Lu C (2023b) Dynamic context removal: A general training strategy for robust models on video action predictive tasks. IJCV
Xu Z, Qing L, Miao J (2015b) Activity auto-completion: Predicting human activities from partial videos. In: CVPR
Xue H, Hang T, Zeng Y, Sun Y, Liu B, Yang H, Fu J, Guo B (2022) Advancing high-resolution video-language representation with large-scale video transcriptions. In: CVPR
Xue Z, Marculescu R (2023) Dynamic multimodal fusion. In: CVPR
Xue Z, Song Y, Grauman K, Torresani L (2023) Egocentric video task translation. In: CVPR
Xue Z, Ashutosh K, Grauman K (2024) Learning object state changes in videos: An open-world perspective. In: CVPR
Yan S, Xiong X, Arnab A, Lu Z, Zhang M, Sun C, Schmid C (2022) Multiview transformers for video recognition. In: CVPR
Yan S, Xiong X, Nagrani A, Arnab A, Wang Z, Ge W, Ross D, Schmid C (2023) Unloc: A unified framework for video localization tasks. In: ICCV
Yan W, Zhang Y, Abbeel P, Srinivas A (2021) Videogpt: Video generation using vq-vae and transformers. arXiv
Yang A, Miech A, Sivic J, Laptev I, Schmid C (2021) Just ask: Learning to answer questions from millions of narrated videos. In: CVPR
Yang A, Miech A, Sivic J, Laptev I, Schmid C (2022a) Learning to answer visual questions from web videos. IEEE TPAMI
Yang A, Miech A, Sivic J, Laptev I, Schmid C (2022b) Tubedetr: Spatio-temporal video grounding with transformers. In: CVPR
Yang A, Miech A, Sivic J, Laptev I, Schmid C (2022c) Zero-shot video question answering via frozen bidirectional language models. NeurIPS
Yang A, Nagrani A, Seo PH, Miech A, Pont-Tuset J, Laptev I, Sivic J, Schmid C (2023a) Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In: CVPR
Yang A, Nagrani A, Laptev I, Sivic J, Schmid C (2024a) Vidchapters-7m: Video chapters at scale. NeurIPS
Yang C, Xu Y, Shi J, Dai B, Zhou B (2020a) Temporal pyramid network for action recognition. In: CVPR
Yang D, Liu Y (2024) Active object detection with knowledge aggregation and distillation from large models. In: CVPR
Yang P, Hu VT, Mettes P, Snoek CG (2020b) Localizing the common action among a few videos. In: ECCV
Yang S, Zhang L, Liu Y, Jiang Z, He Y (2023b) Video diffusion models with local-global context guidance. In: IJCAI
Yang X, Yang X, Liu MY, Xiao F, Davis LS, Kautz J (2019) Step: Spatio-temporal progressive learning for video action detection. In: CVPR
Yang Z, Liu J, Wu P (2024b) Text prompt with normality guidance for weakly supervised video anomaly detection. In: CVPR
Yao B, Fei-Fei L (2010) Modeling mutual context of object and human pose in human-object interaction activities. In: CVPR
Yao G, Lei T, Zhong J (2019) A review of convolutional-neural-network-based action recognition. PRL
Yao Z, Cheng X, Zou Y (2023) PoseRAC: Pose Saliency Transformer for Repetitive Action Counting. arxiv
Ye X, Bilodeau GA (2022) Vptr: Efficient transformers for video prediction. In: ICPR
Ye X, Bilodeau GA (2023) A unified model for continuous conditional video prediction. In: CVPRw
Ye X, Bilodeau GA (2024) Stdiff: Spatio-temporal diffusion for continuous stochastic video prediction. In: AAAI
Ye Y, Zhao Z, Li Y, Chen L, Xiao J, Zhuang Y (2017) Video question answering via attribute-augmented attention network learning. In: SIGIR
Yeung S, Russakovsky O, Mori G, Fei-Fei L (2016) End-to-end learning of action detection from frame glimpses in videos. In: CVPR
Yeung S, Russakovsky O, Jin N, Andriluka M, Mori G, Fei-Fei L (2018) Every moment counts: Dense detailed labeling of actions in complex videos. IJCV

Yilmaz A, Shah M (2006) Matching actions in presence of camera motion. CVIU
Yilmaz A, Javed O, Shah M (2006) Object tracking: A survey. CSUR
Yoon JS, Kim K, Gallo O, Park HS, Kautz J (2020) Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In: CVPR
Yu J, Wang Z, Vasudevan V, Yeung L, Seyedhosseini M, Wu Y (2022a) Coca: Contrastive captioners are image-text foundation models. arxiv
Yu J, Li X, Zhao X, Zhang H, Wang YX (2023a) Video state-changing object segmentation. In: ICCV
Yu S, Tack J, Mo S, Kim H, Kim J, Ha JW, Shin J (2021) Generating videos with dynamics-aware implicit generative adversarial networks. In: ICLR
Yu S, Tack J, Mo S, Kim H, Kim J, Ha JW, Shin J (2022b) Generating videos with dynamics-aware implicit generative adversarial networks. In: ICLR
Yu S, Cho J, Yadav P, Bansal M (2023b) Self-chained image-language model for video localization and question answering. NeurIPS
Yu S, Sohn K, Kim S, Shin J (2023c) Video probabilistic diffusion models in projected latent space. In: CVPR
Yu S, Nie W, Huang DA, Li B, Shin J, Anandkumar A (2024) Efficient video diffusion models via content-frame motion-latent decomposition. In: ICLR
Yu Y, Ko H, Choi J, Kim G (2017) End-to-end concept word detection for video captioning, retrieval, and question answering. In: CVPR
Yuan Y, Mei T, Zhu W (2019) To find where you talk: Temporal sentence localization in video with attention based location regression. In: AAAI
Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, Vinyals O, Monga R, Toderici G (2015) Beyond short snippets: Deep networks for video classification. In: CVPR
Zaheer MZ, Mahmood A, Astrid M, Lee SI (2020a) Claws: Clustering assisted weakly supervised learning with normalcy suppression for anomalous event detection. In: ECCV
Zaheer MZ, Mahmood A, Shin H, Lee SI (2020b) A self-reasoning framework for anomaly detection using video-level labels. IEEE SPL
Zanella L, Menapace W, Mancini M, Wang Y, Ricci E (2024) Harnessing large language models for training-free video anomaly detection. In: CVPR
Zatsarynna O, Abu Farha Y, Gall J (2021) Multi-modal temporal convolutional network for anticipating actions in egocentric videos. In: CVPRw, pp 2249–2258
Zatsarynna O, Bahrami E, Farha YA, Francesca G, Gall J (2024) Gated temporal diffusion for stochastic long-term dense anticipation. In: ECCV
Zellers R, Lu X, Hessel J, Yu Y, Park JS, Cao J, Farhadi A, Choi Y (2021) Merlot: Multimodal neural script knowledge models. NeurIPS
Zellers R, Lu J, Lu X, Yu Y, Zhao Y, Salehi M, Kusupati A, Hessel J, Farhadi A, Choi Y (2022) Merlot reserve: Neural script knowledge through vision and language and sound. In: CVPR
Zelnik-Manor L, Irani M (2001) Event-based analysis of video. In: CVPR
Zeng KH, Chen TH, Chuang CY, Liao YH, Niebles JC, Sun M (2017) Leveraging video descriptions to learn video question answering. In: AAAI
Zeng R, Huang W, Tan M, Rong Y, Zhao P, Huang J, Gan C (2019) Graph convolutional networks for temporal action localization. In: ICCV
Zeng Y, Wei G, Zheng J, Zou J, Wei Y, Zhang Y, Li H (2024) Make pixels dance: High-dynamic video generation. In: CVPR
Zha X, Zhu W, Xun L, Yang S, Liu J (2021) Shifted chunk transformer for spatio-temporal representational learning. NeurIPS
Zhai Y, Wang L, Tang W, Zhang Q, Yuan J, Hua G (2020) Two-stream consensus network for weakly-supervised temporal action localization. In: ECCV
Zhang B, Wang L, Wang Z, Qiao Y, Wang H (2016) Real-time action recognition with enhanced motion vector cnns. In: CVPR
Zhang C, Cao M, Yang D, Chen J, Zou Y (2021a) Cola: Weakly-supervised temporal action localization with snippet contrastive learning. In: CVPR
Zhang C, Yang T, Weng J, Cao M, Wang J, Zou Y (2022a) Unsupervised pre-training for temporal action localization tasks. In: CVPR
Zhang CL, Wu J, Li Y (2022b) Actionformer: Localizing moments of actions with transformers. In: ECCV
Zhang H, Xu X, Han G, He S (2020) Context-aware and scale-insensitive temporal repetition counting. In: CVPR
Zhang H, Sun A, Jing W, Zhen L, Zhou JT, Goh RSM (2021b) Natural language video localization: A revisit in span-based question answering framework. IEEE TPAMI
Zhang H, Li X, Bing L (2023a) Video-llama: An instruction-tuned audio-visual language model for video understanding. In: EMNLP
Zhang H, Liu D, Zheng Q, Su B (2023b) Modeling video as stochastic processes for fine-grained video representation learning. In: CVPR
Zhang HB, Zhang YX, Zhong B, Lei Q, Yang L, Du JX, Chen DS (2019a) A comprehensive survey of vision-based human action recognition methods. Sensors
Zhang J, Qing L, Miao J (2019b) Temporal convolutional network with complementary inner bag loss for weakly supervised anomaly detection. In: ICIP
Zhang L, Rao A, Agrawala M (2023c) Adding conditional control to text-to-image diffusion models. In: ICCV
Zhang W, Zhu M, Derpanis KG (2013) From actemes to action: A strongly-supervised representation for detailed action understanding. In: ICCV
Zhang X, Yoon J, Bansal M, Yao H (2024a) Multimodal representation learning by alternating unimodal adaptation. In: CVPR
Zhang Y, Tokmakov P, Hebert M, Schmid C (2019c) A structured model for action detection. In: CVPR
Zhang Y, Shao L, Snoek CG (2021c) Repetitive Activity Counting by Sight and Sound. In: CVPR
Zhang Y, Bai Y, Wang H, Xu Y, Fu Y (2022c) Look more but care less in video recognition. NeurIPS
Zhang Z, Hu J, Cheng W, Paudel D, Yang J (2024b) Extdm: Distribution extrapolation diffusion model for video prediction. In: CVPR
Zhao B, Fei-Fei L, Xing EP (2011) Online detection of unusual events in videos via dynamic sparse coding. In: CVPR
Zhao C, Thabet AK, Ghanem B (2021) Video self-stitching graph network for temporal action localization. In: ICCV
Zhao C, Liu S, Mangalam K, Ghanem B (2023) Re2tal: Rewiring pretrained video backbones for reversible temporal action localization. In: CVPR
Zhao H, Wildes RP (2019) Spatiotemporal feature residual propagation for action prediction. In: ICCV
Zhao H, Wildes RP (2020) On diverse asynchronous activity anticipation. In: ECCV

Zhao H, Gan C, Rouditchenko A, Vondrick C, McDermott J, Torralba A (2018) The sound of pixels. In: ECCV
Zhao H, Torralba A, Torresani L, Yan Z (2019) Hacs: Human action clips and segments dataset for recognition and temporal localization. In: ICCV
Zhao J, Snoek CG (2019) Dance with flow: Two-in-one stream action detection. In: CVPR
Zhao J, Zhang Y, Li X, Chen H, Shuai B, Xu M, Liu C, Kundu K, Xiong Y, Modolo D, et al (2022) Tuber: Tubelet transformer for video action detection. In: CVPR
Zhao L, Gundavarapu NB, Yuan L, Zhou H, Yan S, Sun JJ, Friedman L, Qian R, Weyand T, Zhao Y, et al (2024a) Videoprism: A foundational visual encoder for video understanding. ICML
Zhao Z, Huang X, Zhou H, Yao K, Ding E, Wang J, Wang X, Liu W, Feng B (2024b) Skim then focus: Integrating contextual and fine-grained views for repetitive action counting. arxiv
Zheng C, Wu W, Chen C, Yang T, Zhu S, Shen J, Kehtarnavaz N, Shah M (2020) Deep learning-based human pose estimation: A survey. CSUR
Zheng N, Song X, Su T, Liu W, Yan Y, Nie L (2023) Egocentric early action prediction via adversarial knowledge distillation. ACM TOMM
Zhong JX, Li N, Kong W, Liu S, Li TH, Li G (2019) Graph convolutional label noise cleaner: Train a plug-and-play action classifier for anomaly detection. In: CVPR
Zhong Y, Liang L, Zharkov I, Neumann U (2023a) Mmvp: Motion-matrix-based video prediction. In: ICCV
Zhong Z, Martin M, Voit M, Gall J, Beyerer J (2023b) A survey on deep learning techniques for action anticipation. arxiv
Zhong Z, Schneider D, Voit M, Stiefelhagen R, Beyerer J (2023c) Anticipative feature fusion transformer for multimodal action anticipation. In: WACV
Zhou B, Andonian A, Oliva A, Torralba A (2018a) Temporal relational reasoning in videos. In: ECCV
Zhou H, Martín-Martín R, Kapadia M, Savarese S, Niebles JC (2023) Procedure-aware pretraining for instructional video understanding. In: CVPR
Zhou J, Wang J, Zhang J, Sun W, Zhang J, Birchfield S, Guo D, Kong L, Wang M, Zhong Y (2022) Audio–visual segmentation. In: ECCV
Zhou L, Xu C, Corso JJ (2018b) Towards automatic learning of procedures from web instructional videos. In: AAAI
Zhou L, Zhou Y, Corso JJ, Socher R, Xiong C (2018c) End-to-end dense video captioning with masked transformer. In: CVPR
Zhou X, Arnab A, Buch S, Yan S, Myers A, Xiong X, Nagrani A, Schmid C (2024) Streaming dense video captioning. In: CVPR
Zhou Y, Berg TL (2015) Temporal perception and prediction in ego-centric video. In: ICCV
Zhou Y, Sun X, Zha ZJ, Zeng W (2018d) Mict: Mixed 3d/2d convolutional tube for human action recognition. In: CVPR
Zhu L, Yang Y (2020) Actbert: Learning global-local video-text representations. In: CVPR
Zhu W, Hu J, Sun G, Cao X, Qiao Y (2016) A key volume mining deep framework for action recognition. In: CVPR
Zhu Y, Newsam S (2019) Motion-aware feature for improved video anomaly detection. In: BMVC
Zhu Y, Shen X, Xia R (2023) Personality-aware human-centric multimodal reasoning: A new task, dataset and baselines. arxiv
Zhu Y, Zhang G, Tan J, Wu G, Wang L (2024) Dual detrs for multi-label temporal action detection. In: CVPR
Zhuang S, Li K, Chen X, Wang Y, Liu Z, Qiao Y, Wang Y (2024) Vlogger: Make your dream a vlog. In: CVPR
Zhuo T, Cheng Z, Zhang P, Wong Y, Kankanhalli M (2019) Explainable video action reasoning via prior knowledge and state transitions. In: MM
Zong M, Wang R, Chen X, Chen Z, Gong Y (2021) Motion saliency based multi-stream multiplier resnets for action recognition. IVC