Video Action Understanding: A Tutorial

MATTHEW HUTCHINSON and VIJAY GADEPALLY, MIT Lincoln Laboratory Supercomputing Center (LLSC)

Many believe that the successes of deep learning on image understanding problems can be replicated in the realm of video understanding. However, the span of video action problems and the set of proposed deep learning solutions is arguably wider and more diverse than those of their 2D image siblings. Finding, identifying, and predicting actions are a few of the most salient tasks in video action understanding. This tutorial clarifies a taxonomy of video action problems, highlights datasets and metrics used to baseline each problem, describes common data preparation methods, and presents the building blocks of state-of-the-art deep learning model architectures.

CCS Concepts: • **Computing methodologies** → **Supervised learning**; **Machine learning**; *Ensemble methods*; **Activity recognition and understanding**; *Image representations*.

Additional Key Words and Phrases: video understanding, action understanding, action recognition, action prediction, action proposal, action localization, action detection

1 INTRODUCTION

Video understanding is a natural extension of deep learning research efforts in computer vision. The image understanding field has benefited greatly from the application of artificial neural network (ANN) machine learning (ML) methods. Many image understanding problems—object recognition, scene classification, semantic segmentation, etc.—have workable deep learning "solutions." FixEfficientNet-L2 currently boasts 88.5%/98.7% Top-1/Top-5 accuracy on the ImageNet object classification task [211, 253]. Hikvision Model D scores 90.99% Top-5 accuracy on the Places2 scene classification task [211, 322]. HRNet-OCR yields a mean IoU of 85.1% on the Cityscapes semantic segmentation test [11, 40]. Naturally, many hope that deep learning methods can achieve similar levels of success on video understanding problems.

Drawing from Diba et al. (2019), *semantic video understanding* is a combination of understanding the scene/environment, objects, actions, events, attributes, and concepts [48]. This article focuses on the action understanding component and is presented as a tutorial by introducing a common set of terms and tools, explaining basic and fundamental concepts, and providing concrete examples. We intend this to be accessible to a general computer science audience and assume readers have a basic understanding of supervised learning—the paradigm of learning from input-output examples.

1.1 Action Understanding

While the literature often uses the terms *action* and *activity* synonymously [28, 36, 121], we prefer to use action in this article for a few reasons. First, action is the dominant term across the field, and we would need significant reason to divert from that. Second, the use of activity is generally biased towards human actors rather than non-human actors and phenomenon. We prefer action for its broader applicability. Third, activity recognition is a term already used in several non-video domains [34, 135, 265]. Meanwhile, action recognition is a primarily computer vision and video-based term.

But what is an action? Kang and Wildes (2016) [118] consider an action to be "a motion created by the human body, which may or may not be cyclic." Zhu et al. (2016) [325] define action as an "intentional, purposive, conscious and subjectively meaningful activity." Several human action surveys create a spectrum of action complexity from gestures to interactions or group activities

Authors' address: Matthew Hutchinson, hutchinson@alum.mit.edu; Vijay Gadepally, vijayg@ll.mit.edu, MIT Lincoln Laboratory Supercomputing Center (LLSC), 244 Wood Street, Lexington, Massachusetts, 02421-6426.

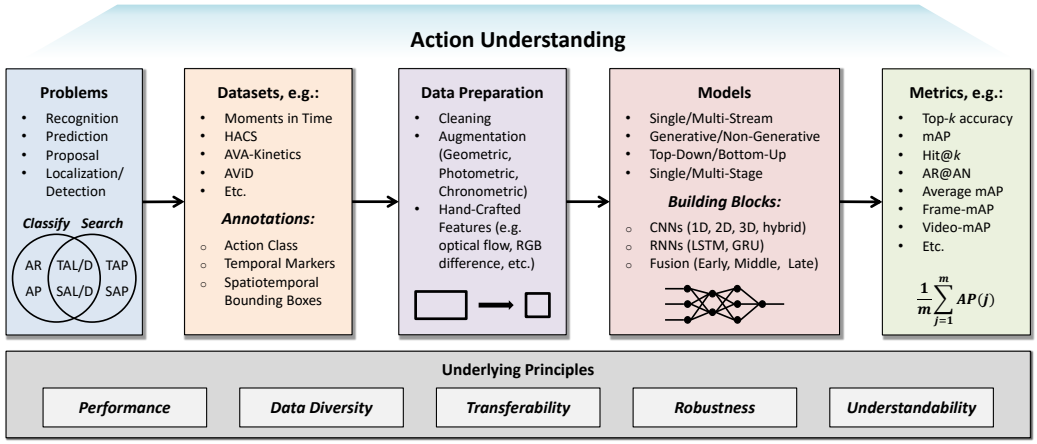

Fig. 1. Overview of action understanding steps (problem formulation, dataset selection, data preparation, model development, and metric-based evaluation) and underlying principles (computational performance, data diversity, transferability, robustness, and understandability. This serves as the framework for this tutorial.

[36, 83, 325]. Unlike these surveys, we will use a broader definition of action, one that includes actions of both human and non-human actors because (1) video datasets are being introduced that use this broader definition [177, 178], (2) most deep learning metrics and methods are equally applicable to both settings, and (3) the colloquial use of action has no distinction between human and non-human actors. Merriam-Webster's Dictionary and the Oxford English Dictionary define action as "an act done" and "something done or performed", respectively [174, 197]. Therefore, this article defines *action* as something done or performed intentionally or unintentionally by a human or non-human actor from which a human observer could derive meaning. This includes everything from low-level gestures and motions to high-level group interactions.

As shown in Figure 1, *action understanding* encompasses action problems, video action datasets, data preparation techniques, deep learning models, and evaluation metrics. Underlying these steps are computer vision and supervised learning principles of computational performance, data diversity, transferability, model robustness, and understandability.

1.2 Related Work and Our Contribution

Table 1 shows a selection of surveys written in the last decade on action understanding. Of the more recent examples, Kong and Fu [126], Xia and Zhan [290], and Rasouli [202] are the most thorough in their independent directions. Despite all of these works, few focus on more than one or two action problems or present more than a narrow coverage of video action datasets. The vast majority only consider a narrow (human) definition of actions. Additionally, the few that cover metrics generally do so shallowly. Relative to the literature noted above, we present this article as a tutorial and contribute the following:

- Clear definitions of recognition, prediction, proposal, and localization/detection problems.
- An extensive and up-to-date catalog of video action datasets.
- Descriptions of the oft neglected, yet important methods of data preparation.
- Explanations of common deep learning model building blocks.
- Groupings of state-of-the-art model architectures.
- Formal definitions of evaluation metrics across the span of action problems.

Table 1. Coverage of surveys on action understanding. Tabular information includes year of publication, number of citations on Google Scholar as of August 2020, action coverage: human (H) and non-human (N), topic coverage: datasets (Ds), metrics (Mc), models/methods (Md), and problem coverage: action recognition (AR), action proposal (AP), temporal action proposal (TAP), temporal action localization/detection (TAL/D), spatiotemporal action localization/detection (SAL/D).

			Actions		Topics			Problems				
Survey	Year	Cited	H	N	Ds	Mc	Md	AR	AP	TAP	TAL/D	SAL/D
Poppe [196]	2010	2,252	✓		✓		✓	✓				
Weinland et al. [281]	2011	1,050	✓		✓		✓	✓				
Ahad et al. [5]	2011	28	✓		✓		✓			✓		
Chaquet et al. [28]	2013	364	✓		✓		✓	✓				
Guo and Lai [83]	2014	170	✓		✓		✓	✓	✓			
Cheng et al. [36]	2015	116	✓		✓		✓	✓	✓			
Zhu et al. [325]	2016	69	✓				✓	✓				
Kang and Wildes [118]	2016	31	✓		✓	✓	✓	✓	✓		✓	
Zhang et al. [311]	2016	168	✓		✓		✓	✓			✓	
Herath et al. [89]	2017	342	✓		✓		✓	✓				
Koohzadi and Charkari [128]	2017	27	✓				✓	✓				
Asadi-Aghbolaghi et al. [9]	2017	103	✓		✓		✓	✓				
Kong and Fu [126]	2018	91	✓		✓		✓	✓	✓			
Zhang et al. [310]	2019	60	✓		✓		✓	✓	✓		✓	
Bhoi [14]	2019	1	✓	✓	✓		✓					✓
Singh and Vishwakarma [233]	2019	18	✓		✓			✓			✓	
Xia and Zhan [290]	2020	0	✓	✓	✓	✓				✓	✓	
Rasouli [202]	2020	0	✓	✓	✓	✓	✓		✓			
Ours	2020		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

Our paper is organized in the following way. Section 2 defines and organizes action understanding problems. Section 3 catalogs video action datasets by annotation type which directly relates to the problems for which they are applicable. Section 4 provides an introduction to video data and data preparation techniques. Section 5 presents basic model building blocks and organizes state-of-the-art methods. Section 6 defines standard metrics used across these problems, formally shows how they are calculated, and points to examples of their usage in high-profile action understanding competitions. Section 7 summarizes and concludes the tutorial.

2 PROBLEMS

Several problems fall under the umbrella of action understanding. In this section, we introduce a taxonomy of these problems, provide definitions, and indicate disagreements in the literature.

2.1 Taxonomy

As shown in Figure 2, we organize the main action understanding problems into overlapping classify and search bins. Classification problems involve labeling videos by their action class. Search problems involve temporally or spatiotemporally finding action instances.

2.1.1 Definitions. The following are the action understanding problems covered in this tutorial:

Action Recognition (AR) is the process of classifying a complete input (either an entire video or a specified segment) by the action occurring in the input. If the action instance spans the entire length of the input, then the problem is known as *trimmed action recognition*. If the action instance does not span the entire input, then the problem is known as *untrimmed action recognition*. Untrimmed action recognition is generally more challenging because a model would need to complete the action classification task while disregarding non-action background segments of the input.

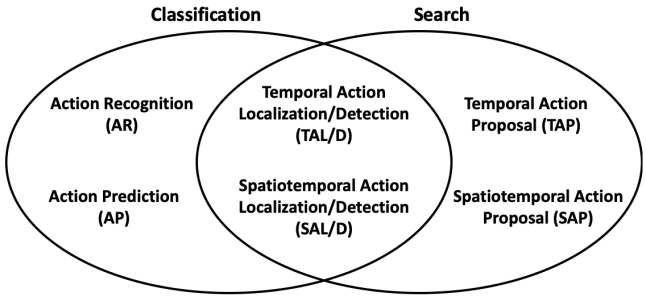

Fig. 2. Action understanding problem taxonomy.

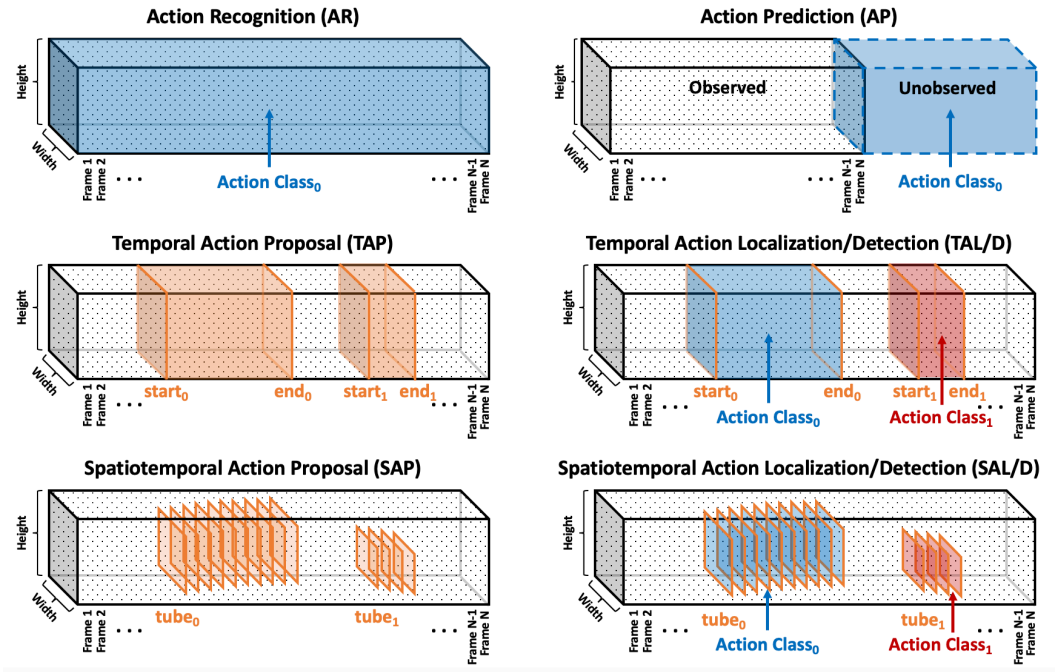

Fig. 3. Action understanding problems: action recognition (AR), action prediction (AP), temporal action proposal (TAP), temporal action localization/detection (TAL/D), spatiotemporal action proposal (SAP), and spatiotemporal action localization/detection (SAL/D). Video is depicted as a 3D volume where N frames are densely stacked along a temporal dimension.

Action Prediction (AP) is the process of classifying an incomplete input by the action yet to be observed. One sub-problem is *action anticipation (AA)* in which no portion of the action has yet to be observed and classification is entirely based on observed contextual clues. Another is *early action prediction (EAP)* in which a portion, but not the entirety, of the action instance has been observed. Both AR and AP are classification problems, but AP often requires a dataset with temporal annotations so that there is a clear delimiter between a "before-action" segment and "during-action" segment for AA or between "start-action" and "end-action" for EAP.

Temporal Action Proposal (TAP) is the process of partitioning an input video into segments (consecutive series of frames) of action and inaction by indicating start and end markers of each action instance. *Temporal Action Localization/Detection (TAL/D)* is the process of creating temporal action proposals and classifying each action.

Spatiotemporal Action Proposal (SAP) is the process of partitioning an input video by both space (bounding boxes) and time (per-frame OR start and end markers of a segment) between regions of action and inaction. If a linking strategy is applied to bounding boxes across several frames, the regions of actions that are constrained in the spatial and temporal dimensions are often referred to as *tubes* or *tubelets*. *Spatiotemporal Action Localization/Detection (SAL/D)* is the process of creating spatiotemporal action proposals and classifying each frame's bounding boxes (or action tubes when a linking strategy is applied).

2.1.2 Literature Observations. This taxonomy and these definitions are intended to clarify several term discrepancies in the literature. First, recognition and classification are sometimes used interchangeably (e.g. [75, 123, 186]). We believe that should be avoided because both recognition (an identification task) and prediction (an anticipation task) require arranging inputs into categories (i.e. classification). To use recognition and classification synonymously would incorrectly equate recognition and prediction. Second, localization and detection are often used interchangeably (e.g. [27, 227, 319]). However, in this case, because the task involved finding and identifying, we feel the terms are appropriate. While detection appears slightly more prevalent in the temporal action literature and localization appears slightly more prevalent in the spatiotemporal action literature, this article will remain neutral and use localization/detection (L/D) together as a single term. Third, action proposal and action proposal generation are used interchangeably (e.g. [67, 157, 164]). We chose to use the former because it is shorter and proposal can be defined as the act of generating a proposal. Referring to proposal generation is redundant. An important takeaway is that there are many examples in the literature where different terms are referring to the same video action problem (e.g. [14] and [53]). Similarly, there are many examples where the same terms are referring to different video action problems (e.g. [308] and [292]). To compound the issue, many video action datasets can be applied to more than one of these problems. We encourage those entering the field to carefully examine a paper's purpose before assuming it is related to a particular line of interest.

Another notable observation from the literature is that while TAP and TAL/D are sometimes studied independently, SAP is not studied outside of a SAL/D framework. Therefore, the remainder of this article will not refer to SAP independently of SAL/D.

2.2 Related Problems

Here, we highlight a few video problems related to but not included in our main taxonomy.

Action instance segmentation (AIS) is the labeling of individual instances or examples of an action within the same video input even when these action instances may overlap in both space and time. Therefore, AIS is a constraint that can be placed on top of TAL/D or SAL/D. For example, a model performing SAL/D on a video of a concert may identify the frames and bounding boxes sections where the audience is shown and label the proposed temporal segment with the action "clapping." Applying the AIS constraint on top of this would require the model to divide the bounding boxes into each individual clapping member of the audience and track these individual actions across time. Useful action instance segmentation literature includes Weinland et al. (2011) [281], Saha et al. (2017) [217], Ji et al. (2018) [107], and Saha et al. (2020) [218].

Dense captioning is the generation of sentence descriptions for videos. This problem spans several of the video understanding semantic components and is worth noting because it is often paired with action understanding problems in public challenges [71, 72, 234, 235]. Similarly, video captioning

datasets (such as MSVD [32], MVAD [252], MPII-MD [209] and ActivityNet Captions [130]) will sometimes be included in video action understanding dataset lists. For more on video captioning, Li et al. (2019) [146] present a survey on methods, datasets, difficulties, and trends.

Action spotting (AS), proposed by Alwassel et al. (2018) [8] is the process of finding any temporal occurrence of an action in a video while observing as little as possible. This differs from TAL/D in two ways. First, AS requires only finding a single frame within the action instance segment rather than start and end markers. Second, AS is concerned with the efficiency of the search process.

Object tracking is the process of detecting objects and linking detections between frames to track them across time. Object tracking is a relevant related problem because some metrics used for object detection in videos were adopted in video action detection [54, 136]. We recommend Yao et al. (2019) [299] for a recent and broad survey on video object segmentation and tracking.

3 DATASETS

Data is critical to successful machine learning model. In this section, we catalog video action datasets, describe the diversity of foundational and emerging benchmarks, and highlight competitions using these datasets that have been the pinnacle drivers of model development and progress in the field.

3.1 Video Action Dataset Catalog

The last two decades has seen huge growth in available video action datasets. To the best of our knowledge, we have organized the most comprehensive collection of these datasets in the literature. We catalog 137 video action datasets sorted by release year. Due to the scale of this catalog, 30 of the most historically influential, current state-of-the-art, and emerging benchmarks datasets are highlighted in Table 2 while the full catalog can be found in Appendix A.

3.1.1 Criteria. We include a dataset in our catalog if it meets the following criteria:

(1) The dataset was released between 2004 and 2020.
(2) The dataset contains single-channel (B/W) or three-channel (RGB) videos.
(3) The dataset includes annotations of each video or defined segments of each video.
(4) The dataset contains at least 2 action classes.
(5) The dataset contains at least one of the following types of annotations: (C) action class labels, (T) temporal start/end segment markers or frame-level labels, or (S) spatiotemporal frame-level bounding boxes or masks.

3.1.2 Content. For each dataset, we report the name, release year, citations on Google Scholar as of August 2020, number of action classes, number of action instances, types of actors: human (H) and/or non-human (N), annotations: class (C), temporal (T), or spatiotemporal (S), and theme/purpose. We chose to include the total number of action instances rather than total number of videos because supervised learning (the predominant action understanding deep learning paradigm) is dependent on the number of positively labeled examples in the training set. Including annotation type is critical because those determine the types of action understanding problems for which the datasets are useful. The theme/purpose is intended to provide some insight into the applicability of a particular dataset. While the catalog may not include a dataset for your specific research purpose, we hope that it helps in finding suitable data for pretraining and transfer learning.

3.1.3 Trends. By plotting these datasets by year and size in Figure 4, several trends and observations emerge. First, these datasets have grown considerably over the past two decades in both number of action classes and number of action instances. This trend is present across all of the use cases and has occurred over several orders of magnitude. Larger datasets are essential for training deep learning models with often millions of parameters. Second, datasets only useful for classification

Table 2. 30 historically influential, current state-of-the-art, and emerging benchmarks of video action datasets. Tabular information includes dataset name, year of publication, citations on Google Scholar as of August 2020, number of action classes, number of action instances, actors: human (H) and/or non-human (N), annotations: action class (C), temporal markers (T), spatiotemporal bounding boxes/masks (S), and theme/purpose. The full catalog can be found in Appendix A.

			Action		Actors		Annotations			
Video Dataset	Year	Cited	Classes	Instances	H	N	C	T	S	Theme/Purpose
KTH [219]	2004	3,853	6	2,391	✓		✓			B/W, static background
Weizmann [15]	2005	1,890	10	90	✓		✓			human motions
Coffee & Cigarettes [138]	2007	491	2	246	✓		✓	✓	✓	movies and TV
Hollywood2 [172]	2009	1,312	12	3,669	✓		✓			movies
VIRAT [188]	2011	536	23	~10,000	✓		✓	✓	✓	surveillance, aerial-view
HMDB51 [133]	2011	1,928	51	~7,000	✓		✓			human motions
UCF101 [238]	2012	2,470	101	13,320	✓		✓			web videos, expand UCF50
ADL [195]	2012	619	18	~1,200	✓		✓		✓	egocentric, daily activities
THUMOS'13 [101, 112, 238]	2013	146	*101	13,320	✓		✓		✓	web videos, extend UCF101
J-HMDB-21 [106]	2013	458	51	928	✓		✓		✓	re-annotate HMDB51 subset
Sports-1M [119]	2014	4,361	487	1,000,000	✓		✓			multi-label, sports
MEXaction2 [42]	2015	n/a	2	1,975	✓		✓	✓		culturally relevant actions
ActivityNet200 (v2.3) [21]	2016	797	200	23,064	✓		✓	✓		untrimmed web videos
Kinetics-400 [120]	2017	810	400	306,245	✓		✓			diverse web videos
AVA [81]	2017	270	80	>392,416	✓		✓		✓	atomic visual actions
Moments in Time (MiT) [177]	2017	137	339	836,144	✓	✓	✓			intra-class variation, web videos
MultiTHUMOS [301]	2017	231	65	~16,000	✓		✓	✓		multi-label, extends THUMOS
Kinetics-600 [24]	2018	52	600	495,547	✓		✓			extends Kinetics-400
EGTEA Gaze+ [151]	2018	52	106	10,325	✓		✓	✓	✓	egocentric, kitchen
Something-Something-v2 [170]	2018	5	174	220,847	✓		✓			extends Something-Something
Charades-Ego [229]	2018	19	157	68,536	✓		✓	✓		egocentric, daily activities
Jester [173]	2019	12	27	148,092	✓		✓			crowd-sourced, gestures
Kinetics-700 [25]	2019	33	700	~650,000	✓		✓			extends Kinetics-600
Multi-MiT [178]	2019	1	313	~1,020,000	✓	✓	✓			multi-label, extends MiT
HACS Clips [315]	2019	31	200	~1,500,000	✓		✓			trimmed web videos
HACS Segments [315]	2019	31	200	~139,000	✓		✓	✓		extends and improves SLAC
NTU RGB-D 120 [162]	2019	55	120	114,480	✓		✓			extends NTU RGB-D 60
EPIC-KITCHENS-100 [44]	2020	6	97	~90,000	✓		✓	✓	✓	extends EPIC-KITCHENS-55
AVA-Kinetics [142]	2020	5	80	>238,000	✓		✓		✓	adds annotations, AVA+Kinetics
AViD [194]	2020	0	887	~450,000	✓	✓	✓			diverse peoples, anonymized faces

*Only 24 classes have spatiotemporal annotations. This subset is also known as UCF101-24.

(mainly AR) are considerably larger and more prevalent than temporally or spatiotemporally annotated datasets. This is expected because temporal markers or spatiotemporal bounding boxes are more challenging to create. An annotator may require only a few seconds to identify whether a particular video contains a given action but would need much more time to mark the start and end of an action. Additionally, solving AR is often considered a prerequisite for effective TAL/D or SAL/D. Therefore, recognition research has generally preceded localization/detection research.

3.2 Foundational and Emerging Benchmarks

Below, we describe datasets and dataset families in three groups: (1) datasets with only action class annotations primarily for AR, (2) datasets with temporal annotations most useful for TAP, TAL/D, and sometimes AP and (3) datasets with spatiotemporal annotations most useful for SAL/D. Because many of the earlier influential video action datasets such as KTH, Weizmann, etc. are described at length in previous survey papers [5, 28, 126], we focus the following descriptions on the current largest and highest quality datasets.

3.2.1 Action Recognition Datasets. Table 5 plots AR-focused datasets by number of classes and number of instances. Here we describe some of the largest and highest quality among them.

Sports-1M [119] was produced in 2014 as a large-scale video classification benchmark for comparing CNNs. Examples of the 487 sports action classes include "cycling", "snowboarding", and

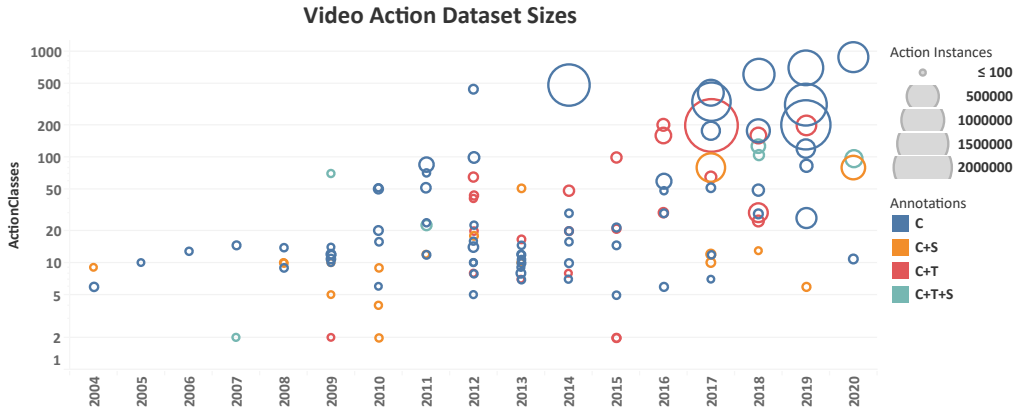

Fig. 4. Trends in video action dataset sizes from 2004 to mid-2020. Both the number of action classes and action instances in these datasets have increased by several orders of magnitude. Note that the action classes dimension is log-scaled. Datasets have increased by several orders of magnitude in both number of action classes and number of action instances.

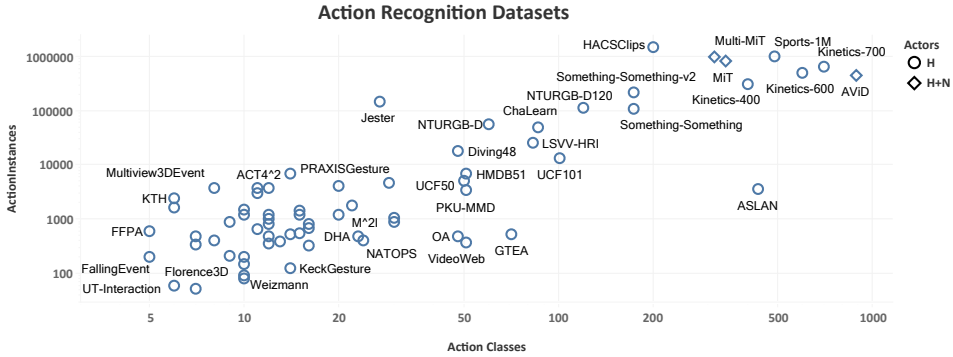

Fig. 5. Datasets with only action class annotations mainly useful for AR. Note that the plot is log-scaled in both action instances and action classes dimensions.

"american football". Note that some inter-class variation is low (e.g. classes include 23 types of billiards, 6 types of bowling, and 7 types of American football). Videos were collected from YouTube and weakly annotated using text metadata. The dataset consists of one million videos with a 70/20/10 training/validation/test split. On average, videos are ~5.5 minutes long, and approximately 5% are annotated with > 1 class. As one of the first large-scale datasets, Sports-1M was critical for demonstrating the effectiveness of CNN architectures for feature learning.

Something-Something [80] (a.k.a. 20BN-SOMETHING-SOMETHING) was produced in 2017 as a human-object interaction benchmark. Examples of the 174 classes include "holding something", "turning something upside down", and "folding something". Video creation was crowd-sourced through Amazon Mechanical Turk (AMT). The dataset consists of 108,499 videos with an 80/10/10 training/validation/test split. Each single-instance video lasts for 2-6 seconds. The dataset was expanded to *Something-Something-v2* [170] in 2018 by increasing the size to 220,847 videos, adding

object annotations, reducing label noise, and improved video resolution. These datasets are important benchmarks for human-object interaction due to their scale and quality.

The *Kinetics* dataset family was produced as "a large-scale, high quality dataset of URL links" to human action video clips focusing on human-object interactions and human-human interactions. *Kinetics-400* [120] was released in 2017, and examples of the 400 human actions include "hugging", "mowing lawn", and "washing dishes". Video clips were collected from YouTube and annotated by AMT crowd-workers. The dataset consists of 306,245 videos. Within each class, 50 are reserved for validation and 100 are reserved for testing. Each single-instance video lasts for ∼10 seconds. The dataset was expanded to *Kinetics-600* [24] in 2018 by increasing the number of classes to 600 and the number of videos to 495,547. The dataset was expanded again to *Kinetics-700* [25] in 2019 by increasing to 700 classes and 650,317 videos. These are among the most cited human action datasets in the field and continue to serve as a standard benchmark and pretraining source.

NTU RGB-D[222] was produced in 2016 as "a large-scale dataset for RGB-D human action recognition." The multi-modal nature provides depth maps, 3D skeletons, and infrared in addition to RGB video. Examples of the 60 human actions include "put on headphone", "toss a coin", and "eat meal". Videos were captured with a Microsoft Kinect v2 in a variety of settings. The dataset consists of 56,880 single-instance video clips from 40 different subjects in 80 different views. Training and validation splits are not specified. The dataset was improved to *NTU RGB-D 120* [162] in 2019 by increasing the number of classes to 120, videos to 114,480, subjects to 106, and views to 155. This serves as a state-of-the-art benchmark for human AR with non-RGB modalities.

Moments in Time (MiT) [177] was produced in 2018 with a focus on broadening action understanding to include people, objects, animals, and natural phenomenon. Examples of the 339 diverse action classes include "running", "opening", and "picking". Videos clips were collected from a variety of internet sources and annotated by AMT crowd-workers. The dataset consists of 903,964 videos with a roughly 89/4/7 training/validation/test split. Each single-instance video lasts for 3 seconds. The dataset was improved to *Multi-Moments in Time (M-MiT)* [178] in 2019 by increasing the number of videos to 1.02 million, pruning vague classes, and increasing the number of labels per video (2.01 million total labels). MiT and M-MiT are interesting benchmarks because of the focus on inter-class and intra-class variation.

Jester [173] (a.k.a. 20BN-JESTER) was produced in 2019 as "a large collection of densely labeled video clips that show humans performing pre-defined hand gestures in front of laptop camera or webcam." Examples of the 27 human hand gestures include "drumming fingers", "shaking hand", and "swiping down". Data creation was crowd-sourced through AMT. The dataset consists of 148,092 videos with an 80/10/10 training/validation/test split. Each single-instance video lasts for ∼3 seconds. The Jester dataset is the first large-scale, semantically low-level human AR dataset.

Anonymized Videos from Diverse countries (AViD) [194] was produced in 2020 with the intent of (1) avoiding the western bias of many datasets by providing human actions (and some non-human actions) from a diverse set of people and cultures, (2) anonymizing all human faces to protect the privacy of the individuals, and (3) ensuring that all videos in the dataset are static with a creative commons license. Most of the 887 classes are drawn from Kinetics [25], Charades [230], and MiT [177] while removing duplicates and any actions that involve the face (e.g. "smiling"). 159 actions not found in any of those datasets are also added. Web videos in 22 different languages were annotated by AMT crowd-workers. The dataset consists of approximately 450,000 videos with a 90/10 training/validation split. Each single-instance video lasts between 3 and 15 seconds. We believe AViD will quickly become a foundational benchmark because of the emphasis on diversity of actors and privacy standards.

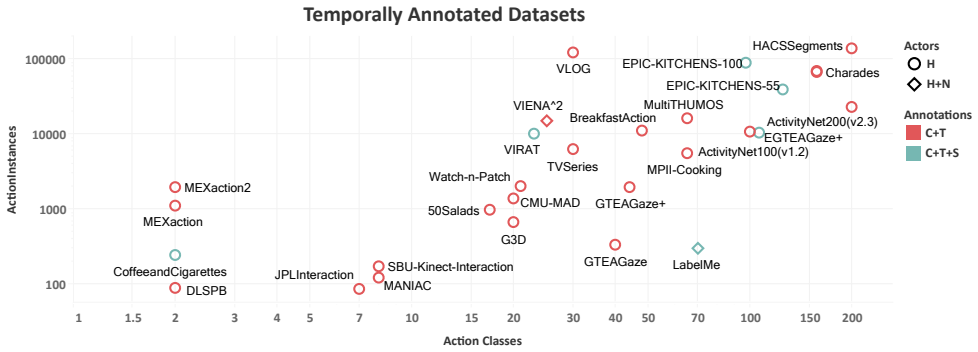

Fig. 6. Datasets with temporal annotations useful for TAP, TAL/D, and possibly AP. Note that the plot is log-scaled in both action instances and action classes dimensions. The SLAC dataset [317] is excluded because while it has a very large number of temporally annotated action instances, the dataset was of poor quality. HACS Segments was developed out of SLAC and has significantly fewer temporally annotated action instances.

3.2.2 Temporally Annotated Datasets. Table 6 plots temporally annotated datasets by number of classes and action instances. Here we describe some of the largest and highest quality among them.

The *ActivityNet* dataset [21, 88] family was produced "to compare algorithms for human activity understanding: global video classification, trimmed activity recognition and activity detection." Example human action classes include "Drinking coffee", "Getting a tattoo", and "Ironing clothes". *ActivityNet 100 (v1.2)* was released in 2015. The 100-class dataset consists of 9,682 videos divided into a 4,819 videos (7,151 instances) training set, a 2,383 videos (3,582 instances) validation set, and a 2,480 videos test set. *ActivityNet 200 (v1.3)* was released in 2016. The 200-class dataset consists of 19,994 videos divided into a 10,024 videos (15,410 instances) training set, a 4,926 videos (7,654 instances) validation set, and a 5,044 videos test set. On average, action instances are 51.4 seconds long. Web videos were temporal annotated by AMT crowd-workers. ActivityNet has remained as a foundational benchmark for TAP and TAL/D because of the dataset scope and size. It is also commonly applied as an untrimmed multi-label AR benchmark.

Charades [230] was produced in 2016 as a crowd-sourced dataset of daily human activities. Examples of the 157 classes include "pouring into cup", "running", and "folding towel". The dataset consists of 9,848 videos (66,500 temporal action annotations) with a roughly 80/20 training/validation split. Videos were filmed in 267 homes with an average length of 30.1 seconds and an average of 6.8 actions per video. Action instances average 12.8 seconds long. *Charades-Ego* was released in 2018 using similar methodologies and the same 157 classes. However, in this dataset, an egocentric (first-person) view and a third-person view is available for each video. The dataset consists of 7,860 videos (68.8 hours) capturing 68,536 temporally annotated action instances. Charades has served as a TAL/D benchmark along with ActivityNet, but it also has found a use as a multi-label AR benchmark because of the high average number of actions per video. Charades-Ego presents a multi-view quality unique among large-scale daily human action datasets.

MultiTHUMOS [301] was produced in 2017 as an extension of the dataset used in the 2014 THUMOS Challenge [113]. Examples of the 65 human action classes include "throw", "hug", and "talkToCamera". The dataset consists of 413 videos (30 hours) with 38,690 multi-label, frame-level annotations (an average of 1.5 per frame). The total number of action instances—where an instance is a set of sequential frames with the same action annotation—is not reported. The number of action

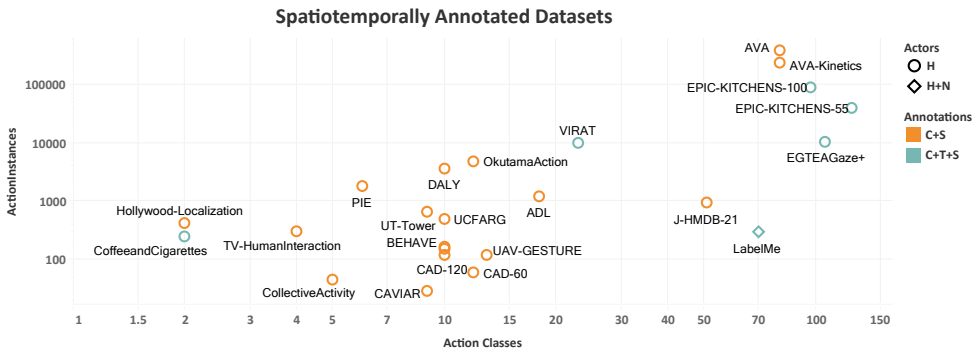

Fig. 7. Datasets with spatiotemporal annotations useful for SAP and SAL/D. Note the the plot is log-scaled in both action instances and action classes dimensions.

instances per class is extremely variable ranging from "VolleyballSet" with 15 to "Run" with 3,500. Each action instance lasts on average for 3.3 seconds with some lasting only 66 milliseconds (2 frames). Like Charades, the MultiTHUMOS dataset offers a benchmark for multi-label TAP and TAL/D. It stands out due to its dense multi-labeling scheme.

VLOG [64] was produced in 2018 as an implicitly gathered large-scale daily human actions dataset. Unlike previous daily human action datasets [80, 129, 230] in which the videos were created, VLOG was compiled from internet daily lifestyle video blogs (vlogs) and annotated by crowd-workers. The method improves diversity of participants and scenes. The dataset consists of 144,000 videos (14 days, 8 hours) using a 50/25/25 training/validation/test split. The 30 classes are the objects with which the person is interacting (e.g. "Bag", "Laptop", and "Toothbrush"). Clips are labeled with these hand/object classes and temporally annotated with the state (positive/negative) of hand-object contact. Because of the collection and annotation methods, VLOG brings actions in daily life datasets closer on par with other temporally annotated large-scale datasets.

HACS Segments [315] was produced in 2019 as "a new large-scale dataset for recognition and temporal localization of human actions collected from Web videos". Both HACS Segments and *HACS Clips* (the AR portion) are improvements on the *SLAC* dataset produced in the 2017 [317]. HACS uses the same 200 human action classes as ActivityNet 200 (1.3). Videos were collected from YouTube and temporally annotated by crowd-workers. HACS Segments consists of 50,000 videos with a 76/12/12 training/validation/test split. The dataset contains 139,000 action instances (referred to as segments). Compared to ActivityNet, the number of action instances per video is greater (2.8 versus 1.5), and the average action instance duration is shorter (40.6 versus 51.4). HACS Segments is an emerging benchmark and provides a more challenging task for human TAP and TAL/D.

3.2.3 Spatiotemporally Annotated Datasets. Table 7 plots spatiotemporally annotated datasets. Here we describe some of the largest and highest quality among them. We also describe two smaller but still highly relevant datasets: UCF101-24 and J-HMDB-21.

VIRAT [188] was created in 2011 as "a new large-scale surveillance video dataset designed to assess the performance of event recognition algorithms in realistic scenes." It includes both ground and aerial surveillance videos. Examples of the 23 classes include "picking up", "getting in a vehicle", and "exiting a facility". The dataset consists of 17 videos (29 hours) with between 10 and 1,500 action instances per class. Due to the camera to action distance across the varying views, the human to video height ratio is between 2% and 20%. Crowd-workers created bounding boxes around moving

objects and temporal event annotations. While this is a smaller dataset, VIRAT is the highest quality surveillance-based spatiotemporal dataset and is used in the latest SAL/D competitions [234, 235].

UCF101-24, the spatiotemporally labelled data subset of *THUMOS'13* [112], was produced in 2013 as part of the THUMOS'13 challenge. Examples of the 24 human action classes include "BasketballDunk", "IceDancing", "Surfing", and "WalkingWithDog". Note, that the majority of the classes are sports. It consists of 3,207 videos from the original UCF101 dataset [238]. Each video contains one or more spatiotemporally annotated action instances. While multiple instances within a video will have separate spatial and temporal boundaries, they will have the same action class label. Videos average ~7 seconds long. The dataset is organized into three train/test splits. While a small dataset, UCF101-24 remains a foundational benchmark for SAL/D.

J-HMDB-21 [106] was produced in 2013 for pose-based action recognition. Examples of the 21 human action classes include "brush hair", "climb stairs", and "shoot bow". The dataset consists of 928 videos from the original HMDB51 dataset [133] and is divided into three 70/30 train/test splits similar to UCF101. Each video contains one action instance that lasts for the entire duration of the video. 2D joint masks and human-background segmentation annotations were created by AMT crowd-workers. Because all of the action classes are human actions, bounding boxes could easily be derived from the joint masks or segmentation masks. Along with UCF101-24, J-HMDB serves as a early foundational benchmark for SAL/D.

EPIC-KITCHENS-55 [43] was produced in 2018 as a large-scale benchmark for egocentric kitchen activities. Examples of 149 human action classes include "put", "open", "pour", and "peel". Videos were captured by head-mounted GoPro cameras on 32 individuals in 4 cities who were instructed to film anytime they entered their kitchen. AMT crowd-workers located relevant actions and objects as well as created final action segment start/end annotations and object bounding boxes. The dataset consists of 432 videos (55 hours) divided into a 272 video train/validation set, 106 video test set 1 (for previously seen kitchens), and a 54 video test set 2 (for previously unseen kitchens). These sets correspond to 28,561, 8,064, and 2,939 action instances, respectively. The dataset was improved to *EPIC-KITCHENS-100* [44] in 2020 by increasing the number of videos to 700 (100 hours), action instances to 89,879, participants to 37, and environments to 34. Annotation quality was also improved. This dataset serves as a state-of-the-art egocentric kitchen activities benchmark.

Atomic Visual Actions (AVA) [81] was produced in 2017 as the first large-scale spatiotemporally annotated diverse human action dataset. Examples of the 80 classes include "swim", "write", and "drive". The dataset consists of 437 15-minute videos with an approximately 55/15/30 training/validation/test split. When only using the 60 most prominent classes (i.e. excluding those with fewer than 25 action instances), the dataset contains 214,622 training, 57,472 validation, and 120,322 test action instances. Videos were gathered from YouTube and segments were annotated by crowd-workers. Ground truth "tracklets" were calculated between manually annotated sections. Because of the dataset scale, AVA serves a large-scale multi-label benchmark for TAL/D.

The *AVA-Kinetics* dataset [142] was produced in 2020 with the purpose of using an existing large-scale human action recognition dataset to create a large-scale spatiotemporally annotated atomic video action dataset. The dataset consists of combining a subset of videos from Kinetics-700 [25] and all videos from AVA [81] for a total of 238,906 videos with a roughly 59/14/27 training/validation/test split. For each 10-second video from Kinetics-700, a combination of algorithm and human crowd-workers created a bounding box for the frame with the highest person detection. Crowd-workers then labeled the set of action instances performed by the person using the 80 possible action classes from the AVA dataset. This dataset is an emerging benchmark because it improves upon AVA by dramatically expanding the number of annotated frames and increases the visual diversity.

3.3 Competitions

Several competitions have introduced state-of-the-art datasets, galvanized model development, and standardized metrics. THUMOS Challenges [79, 112, 113] conducted through the International Conference on Computer Vision (ICCV) in 2013, the European Conference on Computer Vision (ECCV) in 2014, and the Conference on Computer Vision and Pattern Recognition (CVPR) in 2015. These primarily focused on AR and TAL/D tasks. ActivityNet Large Scale Activity Recognition Challenges [21, 71, 72, 234, 235] were held at CVPR from 2016 through 2020 and have slowly expanded into scope encompassing trimmed AR, untrimmed AR, TAP, TAL/D, and SAL/D competitions. Other challenges have been modeled off THUMOS and ActivityNet such as the Workshop on Multi-modal Video Analysis and Moments in Time Challenge[1] held at ICCV in 2019. We provide an overview of these competitions in Appendix A.

4 DATA PREPARATION

While some datasets are available in pre-processed forms, others are presented *raw*—using the original frame rate, frame dimensions, and duration. *Data preparation* is the process of transforming data prior to learning. This step is essential to extract relevant features, fit model input specifications, and prevent overfitting during training. Key preparation processes include:

- *Data cleaning* is the process of removing detecting and removing incomplete or irrelevant portions of the dataset. For datasets that simply link to YouTube or other web videos (e.g. [24, 25, 120, 315]), this step of determining which videos are still active on the site could be very important and affect the dataset quality.
- *Data augmentation* is the process of transforming data to fit model input specifications and increase data diversity. Data diversity helps prevent *overfitting*—when a model too closely matches training data and fails to generalize to unseen examples. Overfitting can occur when the model learns undesired, low-level biases rather than desired, high-level semantics.
- *Hand-crafted feature extraction* is the process of transforming raw RGB video data into a specified feature space to provide insights that a model may not be able to independently learn. With video data, motion representations are the most common extracted features.

4.1 Video Data

Video is composed of a series of still-image frames where each frame is made of rows and columns of *pixels*, the smallest elements of raster images. In standard 3-channel red-green-blue (RGB) video, each pixel is a 3-tuple with an intensity value from 0 to 255 for each of the three color channels. RGB-D video contributes a fourth channel that represents depth often determined by a depth sensor such as the Microsoft Kinect.[2]

As used throughout this article, a common abstraction to represent video is a 3-dimensional (3D) volume in which frames are densely stacked along a temporal dimension. However, with multi-channel pixels, this volume actually has four dimensions. The desired order of these dimensions can vary between software packages with *(frames, channels, height, width)* known as channels first (NCHW) and *(frames, height, width, channels)* known as channels last (NHWC). This order can lead to performance improvements or degradation depending on the training environment (e.g. Theano and MXNet[3] versus CNTK and TensorFlow[4]).

[1] https://sites.google.com/view/multimodalvideo/home
[2] https://developer.microsoft.com/en-us/windows/kinect/
[3] http://deeplearning.net/software/theano/ and https://mxnet.apache.org/versions/1.6/
[4] https://docs.microsoft.com/en-us/cognitive-toolkit/ and https://www.tensorflow.org/

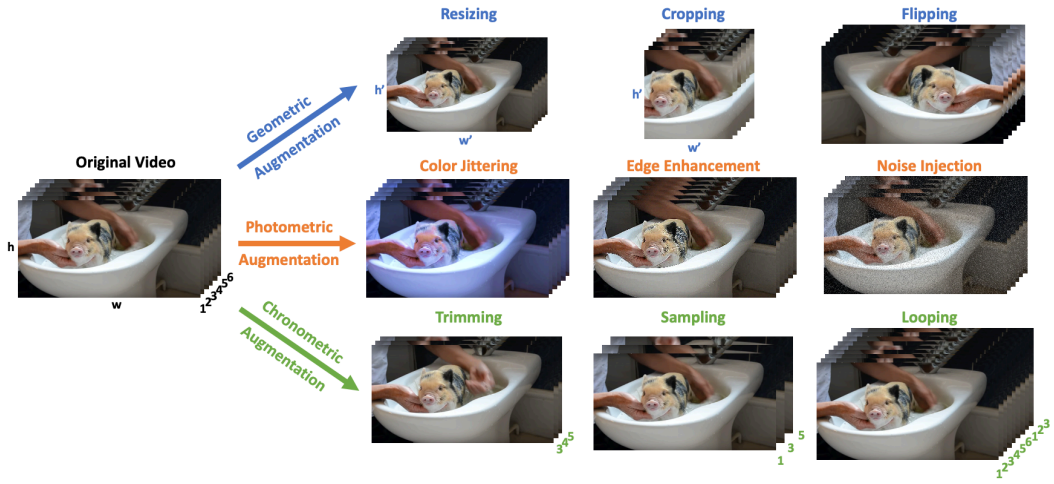

Fig. 8. Common video augmentations. (Frames from the Moments in Time dataset [177], class *washing*)

4.2 Data Augmentation

4.2.1 Geometric Augmentation Methods. In the context of video, *geometric augmentation methods* are transformations that alter the geometry of frames [249]. To be effective, these must be applied equally across all frames. If separate geometric transformations are applied on different frames, a video could quickly loose its semantic meaning. Common geometric augmentations include:

- *Resizing*—the process of scaling a video's frames from a given height and width (h, w) to a new height and width (h', w') via spatial up-sampling or down-sampling [158]. Ratio jittering [272] is resizing that permutes the aspect ratio done for data diversification.
- *Cropping*—the process of transforming a video's frames from a given height and width (h, w) to a new, smaller height and width (h', w') via removing exterior rows or columns. Techniques include random cropping [29, 131, 225] and corner cropping [273].
- *Horizontal (left-right) flipping*—the process of mirroring a video's frames across the vertical axis (i.e. reversing the order of columns in each frame). Random horizontal flipping is a popular and computationally efficient method of introducing data diversity [26, 131, 231, 273].

Other geometric augmentation methods that are less popular for video include *vertical flipping*, *shearing*, *piecewise affine transforming*, and *rotating*. Shorten and Khoshgoftaar (2019) [225] present a survey on image augmentation which describes some of these alternative techniques that could easily be applied to video. While some might be more likely to change the semantic meaning of actions. For example, jumping is an action generally predicated on an actor moving upward. Vertical flipping or a 180 degree rotation would change the apparent direction of motion possibly confusing the model into believing the action is falling.

4.2.2 Photometric Augmentation Methods. In the context of video, *photometric augmentation methods* are transformations that alter the color-space of the pixels making up each frame [249]. Unlike geometric augmentation, these transformations can generally be applied on a per-frame basis and are overall less common in the action understanding literature. These include:

- *Color jittering*—the process of transforming a video's hue, saturation, contrast, or brightness. This can be done randomly [84, 231, 285] or via a specific adjustments [131, 223].

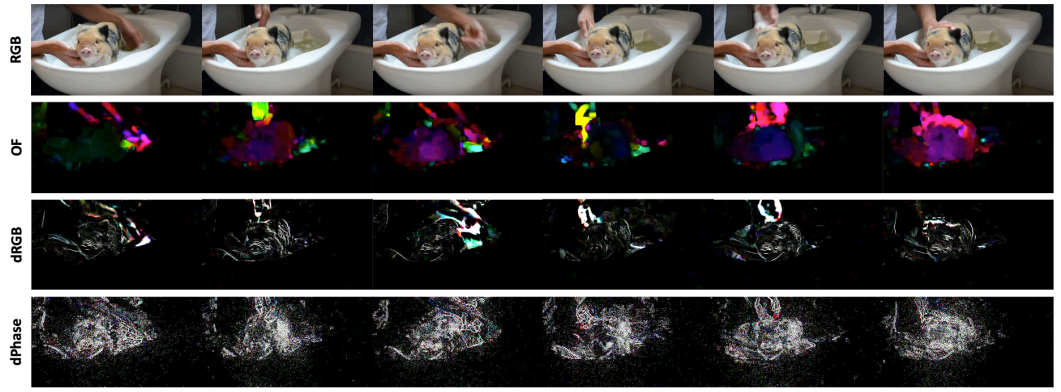

Fig. 9. **(1)** Original RGB, **(2)** dense optical flow (OF) computed using the Farneback method [55] and OpenCV packages [19] (color indicates direction), **(3)** RGB difference/derivative (dRGB), and **(4)** phase difference/derivative (dPhase) computed using the approach described in [92]. (Video frames from the Moments in Time dataset [177], class *washing*)

- *Edge enhancement*—the process of increasing the appearance of contours in a video's frames. In some settings, this speed up the learning process since it has been shown that the first few layers in convolutions neural networks learn to detect edges and gradients [131].

Other photometric augmentation methods that may be useful in future settings are *superpixelization*, *random gray* [84], *random erasing* [225], and *vignetting* [84]. However, currently these are not only absent from the action understanding but uncommon in the image understanding

4.2.3 Chronometric Augmentation Methods. Because the literature does not appear to have a term for transformations that affect the duration of the video input, we refer to these as *chronometric augmentation* following the naming pattern of geometric and photometric. These transformations are generally used to fit a model's input specifications rather than increase data diversity.

- *Trimming*—the process of altering the start and end of a video—essentially temporal cropping. This may be useful to remove the portion of the video that does not include the labeled action.
- *Sampling*—the process of extracting frames from a video—essentially temporal resizing. This can be done from specific frame indices [26, 60] or randomly selected frame indices [231, 273].
- *Looping*—the process of repeating a video's frames to increase the duration—essentially temporal padding [26]. This might be necessary when a video segment has fewer frames than the model's input specifies.

4.3 Hand-Crafted Feature Extraction

While shallow learning has become less common since the deep learning revolution, several hand-crafted motion features have found their way into state-of-the-art deep learning models [26, 60, 231, 273]. These motion representations generally fall under two classical field theories: Lagrangian flow [190] and Eulerian flow [286].

4.3.1 Lagrangian Motion Representations. Lagrangian flow fields track individual parcel or particle motion. In the video context, this refers to tracking pixels by looking at nearby appearance information in adjacent frames to see if that pixel has moved. The most common Lagrangian motion representation is *optical flow (OF)* [73]. Many methods exist for computing this feature: the Lucas–Kanade method [166], the Horn–Schunck method [93], the TV-$L1$ approach [306], the

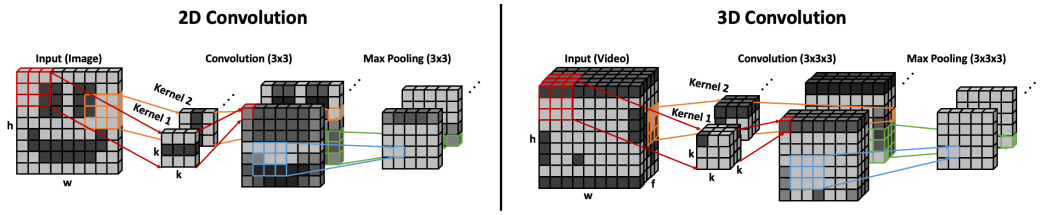

Fig. 10. An example of 2D and 3D convolutional layers and and max pooling layer on single-channel image and video inputs. Note that these filter kernels were chosen randomly and do not necessarily lead to good embedded features.

Farneback method [55], and others [13]. It is also possible to warp OF to attempt to reduce background or camera motion [264]. This technique (*WarpFlow*) requires computing the *homography*, a transformation between two planes, between frames. Optical flow has been noted for its usefulness in action understanding because it is invariant to appearance [221].

4.3.2 Eulerian Motion Representations. Eulerian flow fields represent motion through a particular spatial location. In the video context, this refers to determining visual information differences at a particular spatial location across frames. Two Eulerian motion representations are *RGB difference/derivative (dRGB)* [92, 272, 273] and *phase difference/derivative (dPhase)* [92]. RGB difference is the difference between RGB pixel intensities at equivalent spatial locations in adjacent frames. To compute this, one frame is subtracted from another. Phase difference requires converting each frame into frequency domain before taking the difference and converting back to the time domain.

5 MODELS

The past decade of action understanding research has seen a paradigm shift from primarily shallow, hand-crafted approaches to deep learning where multi-layer artificial neural networks are able to learn complex non-linear relations in structured data. In this section, we describe network building blocks that are common across the diversity of action understanding models and organize state-of-the-art models into architecture families.

5.1 Model Building Blocks

5.1.1 Convolutional Neural Networks. No deep learning architecture component has had a greater impact on action understanding (and computer vision at large) than convolutional neural networks (CNNs), also commonly referred to as ConvNets. A CNN is primarily composed of convolutional, pooling, normalization, and fully-connected layers. For further details, a multitude of tutorials exist on utilizing standard CNN layers (e.g. [139, 140]). CNNs are useful in video understanding because the sharing of weights dramatically decreases the number of trainable parameters and therefore reduces computational cost compared to fully-connected networks. Generally, deeper models (i.e. those with more layers) outperform shallower models by increasing the *receptive field*—the portion of the input that contributes to the feature—of individual neurons in the network [168]. However, deep models can suffer from problems like exploding or vanishing gradients [90].

1-Dimensional CNNs (C1D), 2-Dimensional CNNs (C2D), and 3-Dimensional CNNs (C3D) are the backbone for many state-of-the-art models and use 1D, 2D, and 3D kernels, respectively. C1D is primarily applicable for convolutions along the time dimension of embedded features while C2D and C3D are primarily applicable for extracting feature vectors from individual frames or stacked frames. Single-channel examples of 2D and 3D convolutions are shown in Figure 10. Note that when using multi-channel inputs, the convolutional kernels must be expanded to include a

depth dimension with the same number of channels as the input tensor, and the output is summed across channels. The CNN literature is vast, but we briefly note a few influential developments consistently employed throughout the action understanding literature:

- *Residual networks (ResNets)* [87]—utilize skip connections to avoid vanishing gradients
- *Inception blocks* [245, 246]—utilize multi-size filters for computational efficiency
- *Dense connections (DenseNet)* [97]—utilize skip connections between each layer and every subsequent layer for strengthening feature propagation
- *Inflated networks* [26]—expand lower dimensional networks into a higher dimension in a way that benefits from lower dimensional pretrained weights (e.g. I3D)
- *Normalization* [103]—methods of suppressing the undesired effects of random initialization and random internal distribution shifts. These include *batch normalization (BN)* [103], *layer normalization (LN)* [10], *instance normalization (IN)* [259], and *group normalization (GN)* [288]

Recently, many *hybrid CNNs* have introduced new convolutional blocks, layers, and modules. Some focus on reducing the large computational costs of C3D: P3D [198], R(2+1)D [70, 257], ARTNet [270], MFNet [35], GST [167], and CSN [256]. Others focus on recognizing long-range temporal dependencies: LTC-CNN [261], NL [270], Timeception [99], and STDA [143]. Some unique modules include TSM [155] which shifts individual channels along the temporal dimension for improved 2D CNN performance and TrajectoryNet [318] which uses introduces a TDD-like [271] trajectory convolution to replace temporal convolutions.

5.1.2 Recurrent Neural Networks. The second most common artificial neural network architecture employed in action understanding is the *recurrent neural network (RNN)*. RNNs use a directed graph approach to process sequential inputs such as temporal data. This makes them valuable for action understanding because frames (or frame-based extracted vectors) can be fed as inputs. The most common type of RNN is the *long short-term memory (LSTM)* [91]. An LSTM cell uses an input/forget/output gate structure to perform long-range learning. The second most common type of RNN is the *gated recurrent unit (GRU)* [38]. A GRU cell uses a reset/update gate structure to perform less computationally intensive learning than LSTM cells. Several thorough tutorials cover RNN, LSTM, and GRU use and underlying principles (e.g. [33, 145, 224, 239]).

5.1.3 Fusion. The processes of combining input features, embedded features, or output features are known as *early fusion*, *middle fusion* (or *slow fusion*), and *late fusion* (or *ensemble*), respectively [119, 153, 216]. The simplest and most naïve form is averaging. However, recently *attention mechanisms*, processes that allow a model to focus on the most relevant information and disregard the least relevant information, have gained popularity.

5.2 State-of-the-Art Model Architectures

We focus here on grouping these methods into architecture families under each action problem and pointing to useful examples. Because of the rapidly evolving nature of the field, we recommend checking online scoreboards[5] for up-to-date performances on benchmark datasets.

5.2.1 Action Recognition Models. As shown in Figure 11, we broadly group AR architectures into a few families of varying levels of complexity. The first is *single-stream architectures* which sample or extract one 2D [86, 110, 119] or 3D [85, 108, 248, 254] input feature from a video and feed that into a CNN. The output of the CNN is the model's prediction. While surprisingly effective at some tasks [119, 177], single-stream methods often lack the temporal resolution to adequately perform AR without the application of state-of-the-art hybrid modules discussed in Section 5.1.1.

[5]https://paperswithcode.com/area/computer-vision

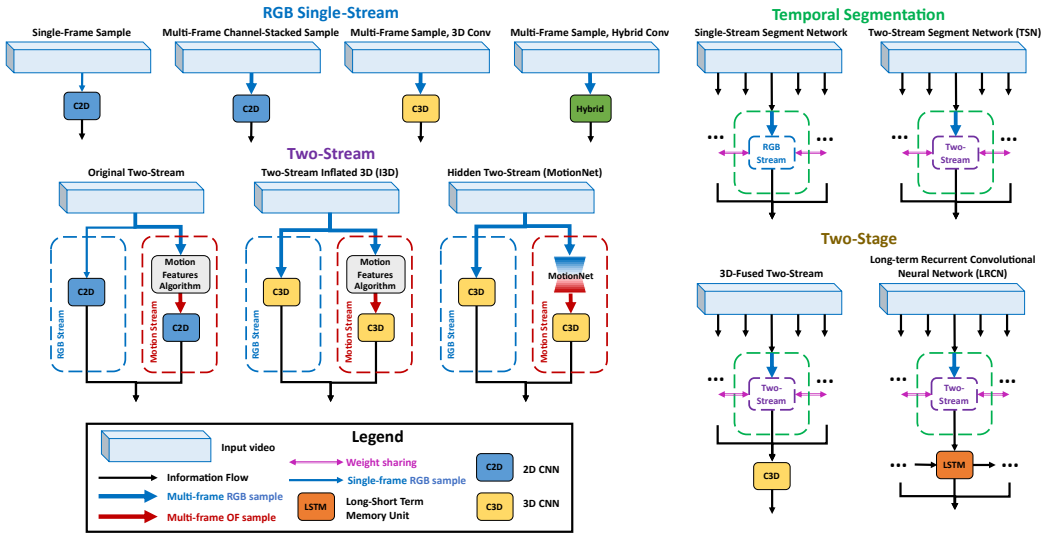

Fig. 11. Action Recognition Model Examples. RGB and Motion Single-Stream architectures train a 2D, 3D, or Hybrid CNN on one sampled feature. Two-stream architectures fuse RGB and Motion streams. Temporal Segmentation architectures divide a video into segments, process each segment on a single-stream or multi-stream architecture, and fuse outputs. Two-stage architectures use temporal segmentation to extract feature vectors and feed those into a convolutional or recurrent network.

The second family is *two-stream architectures* with one stream for RGB learning and one stream for motion feature learning [26, 231]. However, computing optical flow or other hand-crafted features is computationally expensive. Therefore, several recent models use a "hidden" motion stream where motion representations are learned rather than manually determined. These include MotionNet [326] which operates similarly to standard two-stream methods, and MARS [41] and D3D [243] which perform middle fusion between the streams. Feichetenhofer et al. (2017) [59] explores gating techniques between the streams. While these models are generally computationally constrained to two streams, more streams for additional modalities are possible [269, 274].

Built out of single-streams, two-streams, or multi-streams, the third family is *temporal segmentation architectures* which address long-term dependencies of actions. Temporal Segment Network (TSN) methods [272, 273] divide an input video into N segments, sample from those segments, and create video-level prediction by averaging segment level outputs. Model weights are shared between each segment stream. T-C3D [163], TRN [321], ECO [327], and SlowFast [58] build on temporal segmentation by performing multi-resolution segmentation and/or fusion.

The fourth family, at the highest level of complexity in our AR methods taxonomy, is *two-stage learning* where the first stage uses temporal segmentation methods to extract segment embedded feature vectors and the second stage trains on those features. These include 3D-fusion [60] and CNN+LSTM approaches [49, 116, 277, 289, 304]. Ma et al. (2019) conducted a side-by-side comparison of C3D and CNN+LSTM performance [169].

5.2.2 Action Prediction Models. Rasouli (2020) [202] noted that recurrent techniques dominate the approaches. We group these highly diverse action prediction models into *generative* or *non-generative* families. Generative architectures produce "future" features and then classify those predictions. This often takes the form of an encoder-decoder scheme. Examples include RED [69]

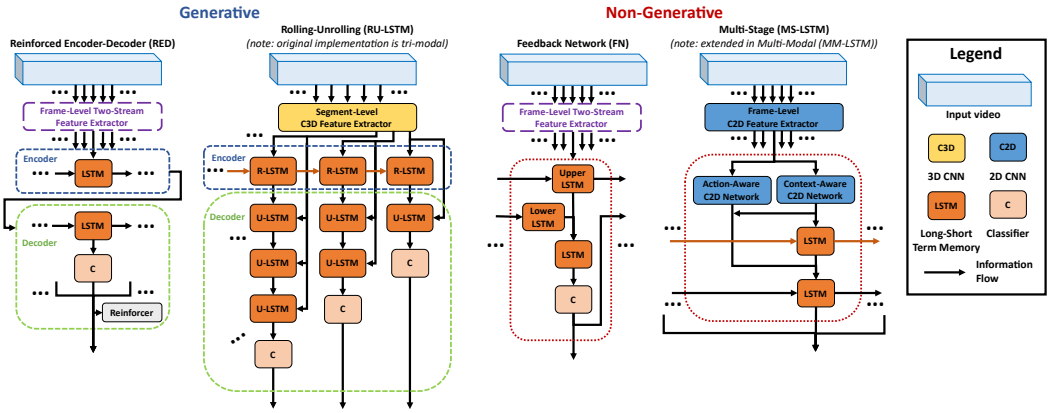

Fig. 12. Action Prediction Model Examples. Generative models create representations of future timesteps for prediction (typically via an encoder-decoder scheme). Non-generative models is a broad-sweeping category for those which create predictions directly from observed sections of the input.

which uses a reinforcement learning module to improve an encoder-decoder, IRL [307] which uses a C2D inverse-reinforcement learning strategy to predict future frames, Conv3D [82] which uses a C3D to generate unseen features for prediction, RGN [316] which uses a recursive generation and prediction scheme with a Kalman filter during training, and RU-LSTM [65, 66] which uses a multi-modal rolling-unrolling encoder-decoder with modality attention.

Non-generative architectures is a broad grouping of all other approaches. These create predictions directly from observed features. Examples include F-RNN-EL [104] which uses an exponential loss to bias a multi-modal CNN+LSTM fusion strategy towards the most recent predictions, MS-LSTM [215] which uses two LSTM stages for action-aware and context-aware learning, MM-LSTM [7] which extends MS-LSTM to arbitrarily many modalities, FN [46] which uses a three-stage LSTM approach, and TP-LSTM [275] which uses a temporal pyramid learning structure.

Many of these examples in this section were developed for action anticipation (when no portion of the action is yet observed), but they are also applicable for early action recognition (when a portion of the action has been observed). Additionally, action recognition models described in Section 5.2.1 may be applicable for some early action recognition tasks if they are able to derive enough semantic meaning from the provided portion and the video context.

5.2.3 Temporal Action Proposal Models. As shown in Figure 13, TAP approaches can be grouped into three families. The first family is *top-down architectures* which consists of models that use sliding windows to derive segment-level proposals. Examples include DAP [52] and SST [20] which use CNN feature extractors and recurrent networks, S-CNN [227] which uses multi-scale sliding windows, and TURN TAP [68] which uses a multi-scale pooling strategy.

The second family is *bottom-up-architectures* which use two-stream frame-level or short-segment-level extracted features to derive "actionness" confidence predictions. Various grouping strategies are then applied to these dense predictions to create full proposals. Examples include TAG [319] which uses a flooding algorithm to convert these into multi-scale groupings, BSN [157] and BMM [155] which use additional "startness" and "endness" feature for different proposal generation and proposal evaluation techniques, and RecapNet [276] which uses a residual causal network rather than a generic 1D CNN to compute confidence predictions. R-C3D [293] and TAL-Net [27] use region-based methods to adapt 2D object proposals in images to 1D action proposals in videos.

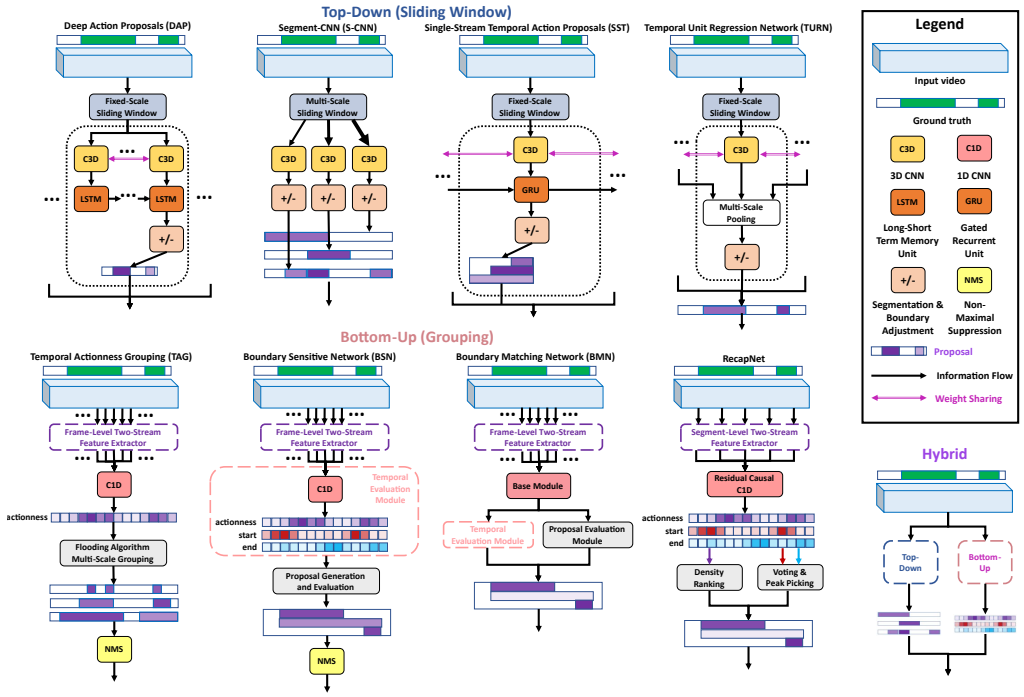

Fig. 13. Temporal Action Proposal Model Examples. Top-down models use a sliding window approach to create segment-level proposals. Bottom-up models use frame or short-segment level actionness score predictions with grouping strategies to produce proposals. Hybrid models use both top-down and bottom-up strategies in parallel.

Many of the bottom-up-architectures require non-maximal suppression (NMS) of outputs to reduce the weight of redundant proposals.

The third family is *hybrid architectures* which combine top-down and bottom-up approaches. These generally create segment proposals and actionness scores in parallel and then use actionness to refine the proposals. Examples include CDC [226], CTAP [67], MGG [164], and DPP [144].

5.2.4 Temporal Action Localization/Detection Models. There are two main families of TAL/D methods as shown in Figure 14. This taxonomy was introduced by Xia et al. (2020) [290]. The first family is *two-stage architectures* in which the first stage creates proposals and the second stage classifies them. Therefore, to create a two-stage architecture, you can pair any of the TAP model described in Section 5.2.3 with an AR model described in Section 5.2.1. It is worth noting that almost all papers that explore TAP methods also extend their work to TAL/D.

The second family is *one-stage architectures* in which proposal and classification happen together. Examples include SSAD [156] which creates a snippet-level action score sequence from which a 1D CNN extracts multi-scale detections, SS-TAD [228] in which parallel recurrent memory cells create proposals and classifications, Decouple-SSAD [98] which builds on SSAD with a three-stream decoupled-anchor network, GTAN [165] which uses multi-scale Gaussian kernels, Two-stream SSD [199] which fuses RGB detections with OF detections, and RBC [115] which completes boundary refinement prior to classification.

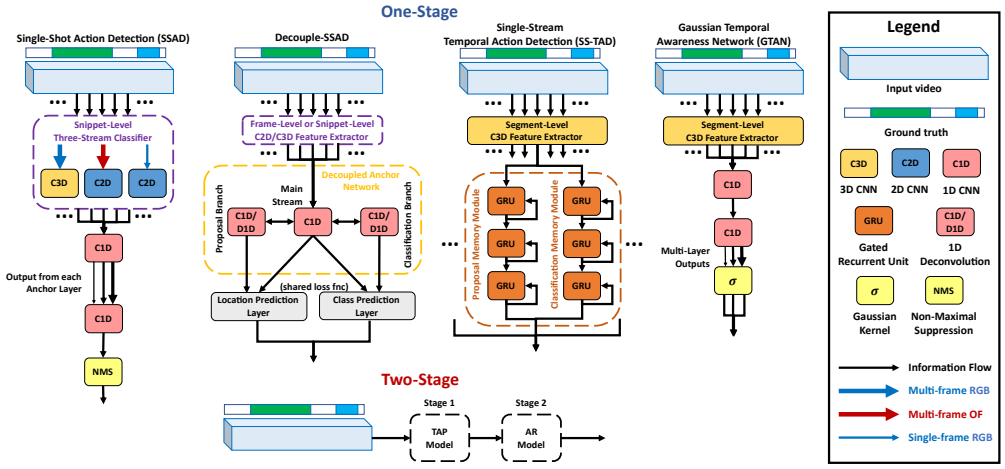

Fig. 14. Temporal Action Localization/Detection Model Examples. One-stage architectures conduct proposal and classification together while two-stage architectures create proposals and then use an action recognition model to classify each proposal.

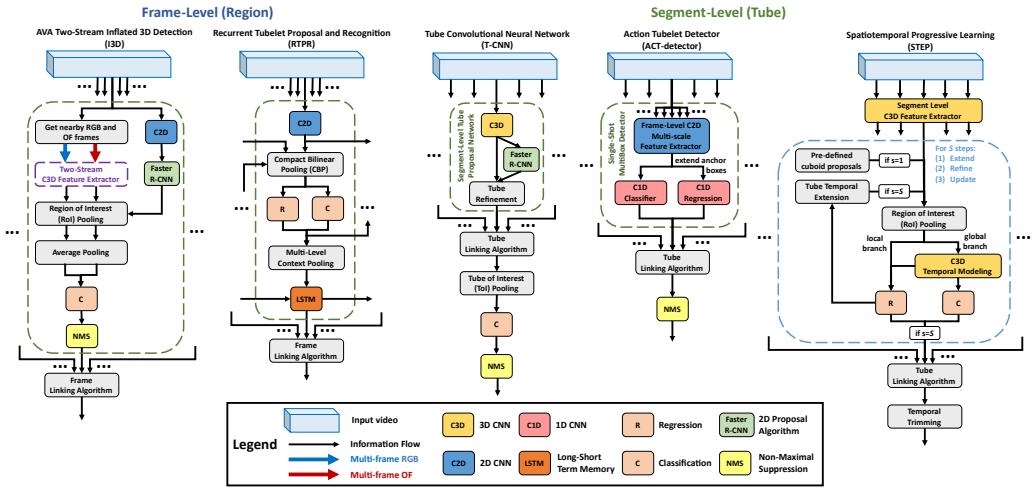

Fig. 15. Spatiotemporal Action Localization/Detection Model Examples. Frame-level (region) proposal models link frame-level detections together while segment-level (tube) proposal models create small "tubelets" for short segments and link the tubelets into longer tubes.

5.2.5 Spatiotemporal Action Localization/Detection Models. As shown in Figure 15, there are two main families of state-of-the-art SAL/D methods. The first is *frame-level (region) proposal architectures* which use various region proposal algorithms (e.g. R-CNN [77], Fast R-CNN [76], Faster R-CNN [207], early+late fusion Faster R-CNN [300]) to derive bounding boxes from frame then apply a frame linking algorithm. Examples include MR-TS [192], CPLA [297], ROAD [232], AVA I3D [81], RTPR [151], and PntMatch [300].

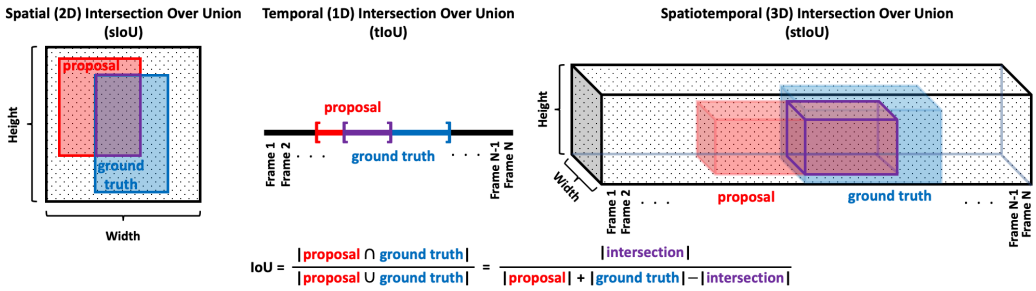

Fig. 16. Illustration of types of intersection over union: spatial, temporal, and spatiotemporal. IoU is also known as the *Jaccard index* or the *Jaccard similarity coefficient*.

The second family is *segment-level (tube) proposal architectures* which uses various methods to create segment-level temporally-small tubes or "tubelets" and then uses a tube linking algorithm. Examples of these models include T-CNN [94], ACT-detector [117], and STEP [296].

A few state-of-the-art models do not fit nicely in either of these families but are worth noting. Zhang et al. (2019) [312] use a tracking network and graph convolutional network to derive person-object detections. VATX [74] augments I3D approaches with a multi-head, multi-layer transformer. STAGE [251] introduces a temporal graph attention method.

6 METRICS

Choosing the right metric is critical to evaluating a model properly. In this section, we define commonly used metrics and point to examples of their usage. We will not cover binary classification metrics as the action datasets we have cataloged overwhelmingly have more than two classes. Note that any time we refer to an accuracy value, the error value can easily be computed as $e = 1 - a$. To clarify terms, we use following notation across the metrics:

- $X = \{x^{(1)}, ..., x^{(n)}\}$: the set of n input videos
- $Y = \{y^{(1)}, ..., y^{(n)}\}$: the set of n ground truth annotations for the input videos
- $M : X \rightarrow \widehat{Y}$: a function (a.k.a. model) mapping input videos to prediction annotations
- $\widehat{Y} = \{\hat{y}^{(1)}, ..., \hat{y}^{(n)}\}$: the set of n model outputs
- $C = \{1, ..., m\}$: the set of m action classes
- $TP_j : \mathbb{N} \rightarrow \{0, 1\}$: a function mapping rank in a list L_j to 1 if the item at that rank is a true positive, 0 otherwise

Several of these metrics also use forms of *intersection over union (IoU)*, a measure of similarity of two regions. Figure 16 depicts spatial IoU, temporal IoU, and spatiotemporal IoU.

6.1 Multi-class Classification Metrics

In action understanding, multi-class classification consists of problems where the model returns per-class confidence scores for each input video. This is done primarily with a *softmax loss* in which the confidence scores across classes for a given input sum to 1. Formally, $\forall i \in \{1, ..., n\}$:

- $y^{(i)} \in C$: the ground truth annotation for input $x^{(i)}$ is a single action class label
- $\hat{y}^{(i)} = \{p_1^{(i)}, ..., p_m^{(i)}\}$ where $p_j^{(i)} \in [0, 1]$ is the probability that video $x^{(i)}$ depicts action j
- $\sum_{j=1}^{m} p_j^{(i)} = 1$ if the model uses softmax output (as is common)

We define the two common metrics below. Other metrics that we will not cover include *F1-score* (micro-averaged and macro-averaged), *Cohen's Kappa*, *PR-AUC*, *ROC-AUC*, *partial AUC (pAUC)*, or

two-way pAUC. Sokolava and Lapalme [236] and Tharwat [250] present thorough evaluations of these and other multi-class classification metrics.

6.1.1 Top-k Categorical Accuracy (a_k). This metric measures the proportion of times when the ground truth label can be found in the top k predicted classes for that input. Top-1 accuracy, sometimes simply referred to as *accuracy*, is the most ubiquitous while Top-3 and Top-5 are other standard choices [21, 71, 72, 234, 235]. In some cases, several Top-k accuracies or errors are averaged. To calculate Top-k accuracy, let $\hat{y}_k^{(i)} \subseteq \hat{y}^{(i)}$ be the subset containing the k highest confidence scores for video $x^{(i)}$. The Top-k accuracy over the entire input set, where $\mathbb{1}$ is a 0-1 indicator function, is:

$$a_k = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\hat{y}_k^{(i)}}(y^{(i)}) \tag{1}$$

6.1.2 Mean Average Precision (mAP). This metric is the arithmetic mean of the interpolated average precision (*AP*) of each class, and it has been used in multiple THUMOS and ActivityNet challenges [21, 72, 79, 112, 113]. To calculate interpolated *AP* for a particular class, the model outputs must be ranked in decreasing confidence of that class. Formally, $\forall j \in \{1, .., m\}$, L_j is a ranked list of outputs such that $\forall a, b \in \{1, ..., n\}, p_j^{(a)} \geq p_j^{(b)}$. The prediction at rank r in list L_j is a true positive if that video's ground truth label $y^{(i)}$ is class j (i.e. $TP_j(r) = 1$ if $y^{(i)} = j$). Using these lists L_1,...,L_m, precision up to rank k in a given list, interpolated *AP* over all ranks with unique recall values for a given class, and *mAP* are calculated as:

$$P_j(k) = \frac{1}{k} \sum_{r=1}^{k} TP_j(r) \tag{2}$$

$$AP(j) = \frac{\sum_{k=1}^{n} P_j(k) * TP_j(k)}{\sum_{k=1}^{n} TP_j(k)} \tag{3}$$

$$mAP = \frac{1}{m} \sum_{j=1}^{m} AP(j) \tag{4}$$

6.2 Multi-label Classification Metrics

In the context of action understanding, multi-label classification consists of AR or AP problems in which the dataset has more than two classes and each video can be annotated with multiple action class labels. As in multi-class classification, the model returns per-class confidence scores for each input. However, in multi-label problems, it is common for the outputs to be calculated through a *sigmoid loss*. Unlike with softmax, confidence scores do not sum to 1. Formally, $\forall i \in \{1, ..., n\}$:

- $y^{(i)} \subseteq C$: the ground truth annotation for input $x^{(i)}$ is a set of action classes
- $\hat{y}^{(i)} = \{p^{(i)}, ..., p_m^{(i)}\}$ where $p_j^{(i)} \in [0, 1]$ is the probability that video $x^{(i)}$ depicts action j

We define two common metrics below. For more information on other metrics such as *exact match ratio* and *Hamming loss*, we recommend Tsoumakas and Katakis (2007) [258] and Wu and Zhou (2017) [287] which present surveys of multi-label classification metrics.

6.2.1 Mean Average Precision (mAP). This is the same metric as described in Section 6.1.2, and it is calculated very similarly for multi-label problems. The difference occurs when determining the true positives in each class list. Here, a prediction at rank r in list L_j is a true positive if class j is one of the video's ground truth labels (i.e. $TP_j(r) = 1$ if $j \in Y^{(i)}$). From there, precision up to rank k, interpolated *AP* for a particular class, and *mAP* are calculated as shown in Equations 2, 3, and 4. This metric is used for MultiTHUMOS [301], ActivityNet 1.3 [21] when applied as an untrimmed

AR problem, and Multi-Moments in Time [178]. One possible variant of multi-label *mAP* involves only computing *AP* for each class up to a specified rank k. Another variant involves only counting predictions as true positives if the confidence score is above a specific threshold (e.g. $t = 0.5$).

6.2.2 Hit@k. This metric indicates the proportion of times when any of the ground truth labels for an input can be found in the top k predicted classes for that input. Once again, 1 and 5 are standard choices for k [119]. Formally, let $\hat{y}_k^{(i)} \subseteq \hat{y}^{(i)}$ be the subset containing the k highest confidence scores for video $x^{(i)}$. A "hit" occurs if the intersection of the ground truth set of labels and the set of top-k predictions is non-empty:

$$Hit@k = \frac{1}{n} \sum_{i=1}^{n} [y^{(i)} \cap \hat{y}_k^{(i)} \neq \emptyset] \tag{5}$$

6.3 Temporal Proposal Metrics

Metrics for TAP are less varied than those for classification. Below, we define the two main ones found in the literature. Here, the model returns proposed temporal regions (start and end markers for each) and a confidence score for each proposal. Formally, $\forall i \in \{1, ...n\}$:

- $y^{(i)} = \{s_1^{(i)}, ...\}$: the ground truth annotation set of temporal segments where $s_j^{(i)}$ consists of start and end markers for input video $x^{(i)}$
- $\hat{y}^{(i)} = \{(\hat{s}_1^{(i)}, c_1^{(i)}), ...\}$ where $c_j^{(i)} \in [0, 1]$ is the probability (confidence) that proposal segment $\hat{s}_j^{(i)}$ matches a ground truth segment for input $x^{(i)}$
- tIoU($s_a^{(i)}$,$\hat{s}_b^{(i)}$): the temporal intersection over union between the ground truth a proposal

Intuitively, a model that produces more proposals will have a better chance of covering all of the ground truth segments. Therefore, TAP metrics include average number of proposals (*AN*), a hyperparameter that can be manually tuned. *AN* is defined as the total number of proposals divided by the total number of input videos. Formally,

$$AN = \frac{1}{n} \sum_{i=0}^{n} |\hat{y}^{(i)}| \tag{6}$$

6.3.1 Average Recall at Average Number of Proposals (AR@AN). Recall is a measure of sensitivity of the prediction model. In this context, a ground truth temporal segment $s_a^{(i)}$ is counted as a true positive if there exists a proposal segment $\hat{s}_b^{(i)}$ that has a tIoU with it greater than or equal to a given threshold t (i.e. $TP_a^{(i)}(t) = 1$ if tIoU$(s_a^{(i)}, \hat{s}_b^{(i)}) \geq t$). Recall is the proportion of all ground truth temporal segments for which there is a true positive prediction. Average recall is the mean of all recall values over thresholds from 0.5 to t_{max} (inclusive) with a step size of η. In the ActivityNet challenges, $t_{max} = 0.95$ and $\eta = 0.05$ [71, 72, 234]. Formally, recall at a particular threshold and *AN* and average recall at *AN* are calculated as:

$$R(t)@AN = \frac{1}{\sum_{i=1}^{n} |y^{(i)}|} \sum_{i=1}^{n} \sum_{j=\{1,...\}} TP_j^{(i)}(t) \tag{7}$$

$$AR@AN = \frac{1}{(t_{max} - 0.5)/\eta + 1} \sum_{l=0}^{(t_{max}-0.5)/\eta} R(0.5 + l\eta)@AN \tag{8}$$

6.3.2 Area Under the AR-AN Curve (AUC). Another metric for TAP is the area under the curve when *AR@AN* is plotted for various values of *AN*. Commonly, this is for values of 1 to 100 with a

step size of 1 [71, 72, 234]. Note that at an AN of 0 where no proposals are given, AR is trivially 0. Using $AR@AN$ from Equation 8, AR-AN AUC is calculated as:

$$AUC = \sum_{AN=1}^{100} \frac{AR@AN - AR@(AN-1)}{2} \tag{9}$$

6.4 Temporal Localization/Detection Metrics

Like temporal proposal, there are two main metrics for TAL/D and both are used across many challenges [21, 71, 72, 79, 113, 234, 235]. Here, the model returns proposed temporal regions (start and end markers for each), a class prediction for each proposal, and a confidence score for each proposal. Formally, $\forall i \in \{1, ..., n\}$:

- $y^{(i)} = \{(s_1^{(i)}, l_1^i), ...\}$: the ground truth annotation set of (temporal segment, class label) pairs for input $x^{(i)}$ where $s_j^{(i)}$ consists of start and end markers and $l_j^{(i)} \in C$
- $\hat{y}^{(i)} = \{(\hat{s}_1^{(i)}, \hat{l}_1^{(i)}, c_1^{(1)}), ...\}$ where $c_j^{(i)}$ is the probability (confidence) that proposal segment $\hat{s}_j^{(i)}$ matches a ground truth segment labeled with class $\hat{l}_j^{(i)}$ for input $x^{(i)}$
- tIoU($s_a^{(i)}$,$\hat{s}_b^{(i)}$): the temporal IoU between a ground truth segment and a proposal

6.4.1 Mean Average Precision at tIoU Threshold (mAP tIoU@t). This metric is the arithmetic mean of the interpolated average precision (AP) over all classes at a given tIoU threshold. First, all proposals for a given class are ranked in decreasing confidence. The difference from standard mAP described in Section 6.1.2 occurs when determining true positives. In this case, a proposal segment $\hat{s}_a^{(i)}$ at rank r in list L_j is counted as a true positive if there exists a ground truth segment $s_b^{(i)}$ that has a tIoU with it greater than or equal to a given threshold t, the predicted class label $\hat{l}_a^{(i)}$ matches the ground truth class label $l_b^{(i)}$, and that ground truth segment has not already been detected by another proposal higher in the ranked list (i.e. $TP_j(r) = 1$ if tIoU$(\hat{s}_a^{(i)}, s_b^{(i)}) \geq t$ and $\hat{l}_a^{(i)} = l_b^{(i)}$). This way, no redundant detections are allowed. Precision up to rank k, interpolated AP for a particular class, and mAP are calculated using Equations 2, 3, 4. Note that in this case, n in Equation 3 must be replaced with the number of prediction tuples for the class j.

6.4.2 Average Mean Average Precision (average mAP). The most common TAL/D metric is the arithmetic mean of mAP over multiple different tIoU thresholds from 0.5 to t_{max} with a given step size η. Commonly, $t_{max} = 0.95$ (inclusive) and $\eta = 0.05$ [21, 71, 72, 234, 235]. Therefore, $average\ mAP$ is computed as:

$$average\ mAP = mAP\ \text{tIoU@0.5:}t_{max}\text{:}\eta = \frac{1}{(t_{max} - 0.5)/\eta + 1} \sum_{j=0}^{(t_{max}-0.5)/\eta} mAP\ \text{tIoU@}(0.5 + j\eta) \tag{10}$$

6.5 Spatiotemporal Localization/Detection Metrics

SAL/D involves locating actions in both time and space as well as classifying the located actions. Here, the model generally returns frame-level proposed spatial regions (bounding boxes), a class prediction for each box, and a confidence score. Formally, $\forall i \in \{1, ..., n\}$:

- $y^{(i)} = \{(f_1^{(i)}, b_1^{(i)}, l_1^{(i)}), ...\}$: the ground truth annotation set of tuples for input $x^{(i)}$ where $f_j^{(i)}$ is the frame number counting up from 1, $b_j^{(i)}$ is a bounding box marking the upper left corner, the box's height, and the box's width, and $l_j^{(i)} \in C$

- $\hat{y}^{(i)} = \{(\hat{f}_1^{(i)}, \hat{b}_1^{(i)}, \hat{l}_1^{(i)}, c_1^{(i)}), ...\}$ where $c_j^{(i)}$ is the confidence that bounding box $\hat{b}_j^{(i)}$ at frame $\hat{f}_j^{(i)}$ matches a ground truth bounding box on the same frame labeled with class $\hat{l}_j^{(i)}$
- $tube_j^{(i)}$: a spatiotemporal tube in video $x^{(i)}$ is a linked set of bounding boxes $b_k^{(i)}, b_l^{(i)}, b_m^{(i)}, ...$ with the same class label ($l_k^{(i)} = l_l^{(i)} = l_m^{(i)} = ...$) in adjacent frames ($k = l - 1 = m - 2 = ...$)
- sIoU($b_a^{(i)}, \hat{b}_b^{(i)}$): the spatial IoU between a ground truth bounding box and a proposed bounding box (note: this requires $f_a^{(i)} = \hat{f}_b^{(i)}$)
- stIoU($tube_a^{(i)}, \widehat{tube}_b^{(i)}$): the spatiotemporal IoU between a ground truth and proposed tubes

6.5.1 Frame-Level Mean Average Precision (frame-mAP). This metric treats is useful because it evaluates the model independent of the linking strategy—the process of developing action instance tubes. It is utilized in several ActivityNet challenges [71, 234, 235]. Like several metrics above, this is the mean of the interpolated *AP* over all classes. For a given class, every prediction tuple is ranked in decreasing confidence. Here, a proposal box $\hat{b}_a^{(i)}$ at rank r in list L_j is counted as a true positive if there exists a ground truth box $b_b^{(i)}$ on the same frame with the same class label that has a sIoU with it greater than or equal to a given threshold t that has not already been detected by another proposed box higher in the ranked list (i.e. $TP_j(r) = 1$ if sIoU($\hat{b}_a^{(i)}, b_b^{(i)}$) $\geq t$ and $\hat{f}_a^{(i)} = f_b^{(i)}$ and $\hat{l}_a^{(i)} = l_b^{(i)}$). No redundant detections are allowed. Precision up to rank k, interpolated *AP* for a particular class, and *mAP* are calculated using Equations 2, 3 and 4. Note that n in Equation 3 must be replaced with the number of prediction tuples for the class j (i.e. the length of ranked list L_j).

6.5.2 Video-Level Mean Average Precision (video-mAP). This metric is useful for evaluating the linking strategy applied to connect bounding boxes of the same class label in adjacent frames. When using *frame-mAP*, longer actions would take up more frames and weight more when calculating *AP* and *mAP*. However, using this metric, each action instance is weighted equally regardless of the temporal duration of the occurrence. This *video-mAP* metric has been employed for use with both AVA and J-HMDB-21 datasets [81, 106]. Once bounding boxes of the same class label in adjacent frames are linked into tubes, every prediction tube of that class is ranked in decreasing confidence. Here, a proposal tube $\widehat{tube}_a^{(i)}$ at rank r in list L_j is counted as a true positive if there exists a ground truth tube $tube_b^{(i)}$ with the same class label that has a stIoU with it greater than or equal to a given threshold t that has not already been detected by another proposed tube higher in the ranked list (i.e. $TP_j(r) = 1$ if stIoU($\widehat{tube}_a^{(i)}, tube_b^{(i)}$) $\geq t$ and $\hat{l}_a^{(i)} = l_b^{(i)}$). No redundant detections are allowed. Precision up to rank k, interpolated *AP* for a particular class, and *mAP* are calculated using Equations 2, 3 and 4. Note that in this case, n in Equation 3 must be replaced with the number of prediction tubes for the class j.

7 CONCLUSION

In this tutorial, we presented the suite of problems encapsulated within action understanding, listed datasets useful as benchmarks and pretraining sources, described data preparation steps and strategies, organized deep learning model building blocks and state-of-the-art model families, and defined common metrics for assessing models. We hope that this tutorial has clarified terminology, expanded your understanding of these problems, and inspired you to pursue research in this rapidly evolving field at the intersection of computer vision and deep learning. This article has also demonstrated the similarities and differences between these action understanding problem spaces via common datasets, model building blocks, and metrics. To that end, we also hope that this can facilitate idea cross-pollination between the somewhat stove-piped action problem sub-fields.

ACKNOWLEDGMENTS

We like to thank the following individuals who have provided feedback on this article: Jeremy Kepner, Andrew Kirby, Alex Knapp, Alison Louthain, and Albert Reuther.

Research was sponsored by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement Number FA8750-19-2-1000. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the United States Air Force or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

REFERENCES

[1] 2010. CMU Graphics Lab Motion Capture Database. http://mocap.cs.cmu.edu/

[2] 2011. The ChaLearn Gesture Dataset (CGD 2011). http://gesture.chalearn.org/data

[3] 2019. Workshop on Multi-modal Video Analysis and Moments in Time Challenge. https://sites.google.com/view/multimodalvideo/home

[4] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. 2016. YouTube-8M: A Large-Scale Video Classification Benchmark. arXiv:1609.08675 [cs.CV]

[5] M. A. R. Ahad, J. Tan, H. Kim, and S. Ishikawa. 2011. Action dataset — A survey. In *SICE Annual Conference 2011*. 1650–1655.

[6] Eren Erdal Aksoy, Minija Tamosiunaite, and Florentin Wörgötter. 2015. Model-free incremental learning of the semantics of manipulation actions. *Robotics and Autonomous Systems* 71 (2015), 118 – 133. https://doi.org/10.1016/j.robot.2014.11.003 Emerging Spatial Competences: From Machine Perception to Sensorimotor Intelligence.

[7] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. 2019. VIENA2: A Driving Anticipation Dataset. In *Computer Vision – ACCV 2018*, C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler (Eds.). Springer International Publishing, Cham, 449–466.

[8] Humam Alwassel, Fabian Caba Heilbron, and Bernard Ghanem. 2018. Action Search: Spotting Actions in Videos and Its Application to Temporal Action Localization. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[9] M. Asadi-Aghbolaghi, A. Clapés, M. Bellantonio, H. J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, and S. Escalera. 2017. A Survey on Deep Learning Based Approaches for Action and Gesture Recognition in Image Sequences. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*. 476–483.

[10] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. 2016. Layer Normalization. arXiv:1607.06450 [stat.ML]

[11] Jean Barbier and Nicolas Macris. 2019. 0-1 phase transitions in sparse spiked matrix estimation. arXiv:1911.05030 [cs.IT]

[12] Mohammadamin Barekatain, Miquel Marti, Hsueh-Fu Shih, Samuel Murray, Kotaro Nakayama, Yutaka Matsuo, and Helmut Prendinger. 2017. Okutama-Action: An Aerial View Video Dataset for Concurrent Human Action Detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[13] S. S. Beauchemin and J. L. Barron. 1995. The Computation of Optical Flow. *ACM Comput. Surv.* 27, 3 (Sept. 1995), 433–466. https://doi.org/10.1145/212094.212141

[14] Amlaan Bhoi. 2019. Spatio-temporal Action Recognition: A Survey. arXiv:1901.09403 [cs.CV]

[15] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. 2005. Actions as space-time shapes. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, Vol. 2. 1395–1402 Vol. 2.

[16] Victoria Bloom, Vasileios Argyriou, and Dimitrios Makris. 2013. Dynamic Feature Selection for Online Action Recognition. In *Human Behavior Understanding*, Albert Ali Salah, Hayley Hung, Oya Aran, and Hatice Gunes (Eds.). Springer International Publishing, Cham, 64–76.

[17] V. Bloom, D. Makris, and V. Argyriou. 2012. G3D: A gaming action dataset and real time action recognition evaluation framework. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 7–12.

[18] Scott Blunsden and RB Fisher. 2010. The BEHAVE video dataset: ground truthed video for multi-person behavior classification. *Annals of the BMVA* 4, 1-12 (2010), 4.

[19] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[20] Shyamal Buch, Victor Escorcia, Chuanqi Shen, Bernard Ghanem, and Juan Carlos Niebles. 2017. SST: Single-Stream Temporal Action Proposals. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[22] Liangliang Cao, Zicheng Liu, and Thomas S Huang. 2010. Cross-dataset action detection. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 1998–2005.

[23] Vincenzo Carletti, Pasquale Foggia, Gennaro Percannella, Alessia Saggese, and Mario Vento. 2013. Recognition of Human Actions from RGB-D Videos Using a Reject Option. In *New Trends in Image Analysis and Processing – ICIAP 2013*, Alfredo Petrosino, Lucia Maddalena, and Pietro Pala (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 436–445.

[24] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. 2018. A Short Note about Kinetics-600. arXiv:1808.01340 [cs.CV]

[25] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. 2019. A Short Note on the Kinetics-700 Human Action Dataset. arXiv:1907.06987 [cs.CV]

[26] Joao Carreira and Andrew Zisserman. 2017. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. arXiv:1705.07750 [cs.CV]

[27] Yu-Wei Chao, Sudheendra Vijayanarasimhan, Bryan Seybold, David A. Ross, Jia Deng, and Rahul Sukthankar. 2018. Rethinking the Faster R-CNN Architecture for Temporal Action Localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[28] Jose M. Chaquet, Enrique J. Carmona, and Antonio Fernández-Caballero. 2013. A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding* 117, 6 (2013), 633 – 659. https://doi.org/10.1016/j.cviu.2013.01.013

[29] Ken Chatfield, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2014. Return of the Devil in the Details: Delving Deep into Convolutional Nets. arXiv:1405.3531 [cs.CV]

[30] C. Chen and J. K. Aggarwal. 2009. Recognizing human action from a far field of view. In *2009 Workshop on Motion and Video Computing (WMVC)*. 1–7.

[31] Chia-Chih Chen, M. S. Ryoo, and J. K. Aggarwal. 2010. UT-Tower Dataset: Aerial View Activity Classification Challenge. http://cvrc.ece.utexas.edu/SDHA2010/Aerial_View_Activity.html.

[32] David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 190–200.

[33] Gang Chen. 2016. A Gentle Tutorial of Recurrent Neural Network with Error Backpropagation. arXiv:1610.02583 [cs.LG]

[34] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu. 2012. Sensor-Based Activity Recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42, 6 (2012), 790–808.

[35] Yunpeng Chen, Yannis Kalantidis, Jianshu Li, Shuicheng Yan, and Jiashi Feng. 2018. Multi-Fiber Networks for Video Recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[36] Guangchun Cheng, Yiwen Wan, Abdullah N. Saudagar, Kamesh Namuduri, and Bill P. Buckles. 2015. Advances in Human Action Recognition: A Survey. arXiv:1501.05964 [cs.CV]

[37] Zhongwei Cheng, Lei Qin, Yituo Ye, Qingming Huang, and Qi Tian. 2012. Human Daily Action Analysis with Multi-view and Color-Depth Data. In *Computer Vision – ECCV 2012. Workshops and Demonstrations*, Andrea Fusiello, Vittorio Murino, and Rita Cucchiara (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 52–61.

[38] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1724–1734.

[39] Liu Chunhui, Hu Yueyu, Li Yanghao, Song Sijie, and Liu Jiaying. 2017. PKU-MMD: A Large Scale Benchmark for Continuous Multi-Modal Human Action Understanding. *arXiv preprint arXiv:1703.07475* (2017).

[40] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. 2016. The Cityscapes Dataset for Semantic Urban Scene Understanding. arXiv:1604.01685 [cs.CV]

[41] Nieves Crasto, Philippe Weinzaepfel, Karteek Alahari, and Cordelia Schmid. 2019. MARS: Motion-Augmented RGB Stream for Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[42] Michel Crucianu. 2015. MEXaction2: action detection and localization dataset. http://mexculture.cnam.fr/Datasets/mex+action+dataset.html

[43] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[44] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Antonino Furnari, Evangelos Kazakos, Jian Ma, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2020. Rescaling Egocentric Vision. arXiv:2006.13256 [cs.CV]

[45] Roeland De Geest, Efstratios Gavves, Amir Ghodrati, Zhenyang Li, Cees Snoek, and Tinne Tuytelaars. 2016. Online Action Detection. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 269–284.

[46] R. De Geest and T. Tuytelaars. 2018. Modeling Temporal Structure with LSTM for Online Action Detection. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1549–1557.

[47] Giovanni Denina, Bir Bhanu, Hoang Thanh Nguyen, Chong Ding, Ahmed Kamal, Chinya Ravishankar, Amit Roy-Chowdhury, Allen Ivers, and Brenda Varda. 2011. *VideoWeb Dataset for Multi-camera Activities and Non-verbal Communication*. Springer London, London, 335–347. https://doi.org/10.1007/978-0-85729-127-1_23

[48] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jurgen Gall, Rainer Stiefelhagen, and Luc Van Gool. 2019. Large Scale Holistic Video Understanding. arXiv:1904.11451 [cs.CV]

[49] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[50] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. 2009. Automatic annotation of human actions in video. In *2009 IEEE 12th International Conference on Computer Vision*. 1491–1498.

[51] M. Edwards, J. Deng, and X. Xie. 2016. From pose to activity: Surveying datasets and introducing CONVERSE. *Computer Vision and Image Understanding* 144 (March 2016), 73 – 105. https://doi.org/10.1016/j.cviu.2015.10.010 Special Issue on Individual and Group Activities in Video Event Analysis.

[52] Victor Escorcia, Fabian Caba Heilbron, Juan Carlos Niebles, and Bernard Ghanem. 2016. DAPs: Deep Action Proposals for Action Understanding. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 768–784.

[53] Victor Escorcia, Cuong D. Dao, Mihir Jain, Bernard Ghanem, and Cees Snoek. 2020. Guess where? Actor-supervision for spatiotemporal action localization. *Computer Vision and Image Understanding* 192 (2020), 102886. https://doi.org/10.1016/j.cviu.2019.102886

[54] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2010. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision* 88, 2 (2010), 303–338. https://doi.org/10.1007/s11263-009-0275-4

[55] Gunnar Farnebäck. 2003. Two-Frame Motion Estimation Based on Polynomial Expansion. In *Image Analysis*, Josef Bigun and Tomas Gustavsson (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 363–370.

[56] Alireza Fathi, Yin Li, and James M Rehg. 2012. Learning to recognize daily actions using gaze. In *European Conference on Computer Vision*. Springer, 314–327.

[57] Alireza Fathi, Xiaofeng Ren, and James M Rehg. 2011. Learning to recognize objects in egocentric activities. In *CVPR 2011*. IEEE, 3281–3288.

[58] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. SlowFast Networks for Video Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[59] Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes. 2017. Spatiotemporal Multiplier Networks for Video Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[60] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. 2016. Convolutional Two-Stream Network Fusion for Video Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[61] Robert B. Fisher. 2004. The PETS04 Surveillance Ground-Truth Data Set. *Proeedings of the Sixth IEEE Workshop on Performance Evaluation of Tracking and Surveillance (PETS04)* 11 (05 2004).

[62] UCF Center for Research in Computer Vision. 2007. UCF Aerial Action Dataset. https://www.crcv.ucf.edu/research/data-sets/ucf-aerial-action/

[63] UCF Center for Research in Computer Vision. 2008. UCF ARG. https://www.crcv.ucf.edu/research/data-sets/ucf-arg/

[64] David F. Fouhey, Wei-cheng Kuo, Alexei A. Efros, and Jitendra Malik. 2018. From Lifestyle Vlogs to Everyday Interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[65] A. Furnari and G. M. Farinella. 2019. Egocentric Action Anticipation by Disentangling Encoding and Inference. In *2019 IEEE International Conference on Image Processing (ICIP)*. 3357–3361.

[66] Antonino Furnari and Giovanni Maria Farinella. 2019. What Would You Expect? Anticipating Egocentric Actions With Rolling-Unrolling LSTMs and Modality Attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[67] Jiyang Gao, Kan Chen, and Ram Nevatia. 2018. CTAP: Complementary Temporal Action Proposal Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[68] Jiyang Gao, Zhenheng Yang, Kan Chen, Chen Sun, and Ram Nevatia. 2017. TURN TAP: Temporal Unit Regression Network for Temporal Action Proposals. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[69] Jiyang Gao, Zhenheng Yang, and Ram Nevatia. 2017. RED: Reinforced Encoder-Decoder Networks for Action Anticipation. arXiv:1707.04818 [cs.CV]

[70] Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. 2019. Large-Scale Weakly-Supervised Pre-Training for Video Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[71] Bernard Ghanem, Juan Carlos Niebles, Cees Snoek, Fabian Caba Heilbron, Humam Alwassel, Victor Escorcia, Ranjay Krishna, Shyamal Buch, and Cuong Duc Dao. 2018. The ActivityNet Large-Scale Activity Recognition Challenge 2018 Summary. arXiv:1808.03766 [cs.CV]

[72] Bernard Ghanem, Juan Carlos Niebles, Cees Snoek, Fabian Caba Heilbron, Humam Alwassel, Ranjay Khrisna, Victor Escorcia, Kenji Hata, and Shyamal Buch. 2017. ActivityNet Challenge 2017 Summary. arXiv:1710.08011 [cs.CV]

[73] James J Gibson. 1950. The perception of the visual world. (1950).

[74] Rohit Girdhar, Joao Carreira, Carl Doersch, and Andrew Zisserman. 2019. Video Action Transformer Network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[75] Rohit Girdhar, Deva Ramanan, Abhinav Gupta, Josef Sivic, and Bryan Russell. 2017. ActionVLAD: Learning Spatio-Temporal Aggregation for Action Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[76] Ross Girshick. 2015. Fast R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[77] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[78] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas. 2009. The i3DPost Multi-View and 3D Human Action/Interaction Database. In *2009 Conference for Visual Media Production*. 159–168.

[79] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar. 2015. THUMOS Challenge: Action Recognition with a Large Number of Classes. http://www.thumos.info/.

[80] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. 2017. The "Something Something" Video Database for Learning and Evaluating Visual Common Sense. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[81] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. 2018. AVA: A Video Dataset of Spatio-Temporally Localized Atomic Visual Actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[82] P. Gujjar and R. Vaughan. 2019. Classifying Pedestrian Actions In Advance Using Predicted Video Of Urban Driving Scenes. In *2019 International Conference on Robotics and Automation (ICRA)*. 2097–2103.

[83] Guodong Guo and Alice Lai. 2014. A survey on still image based human action recognition. *Pattern Recognition* 47, 10 (2014), 3343 – 3361. https://doi.org/10.1016/j.patcog.2014.04.018

[84] Tengda Han, Weidi Xie, and Andrew Zisserman. 2019. Video Representation Learning by Dense Predictive Coding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

[85] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[86] Dongliang He, Zhichao Zhou, Chuang Gan, Fu Li, Xiao Liu, Yandong Li, Limin Wang, and Shilei Wen. 2019. Stnet: Local and global spatial-temporal modeling for action recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 8401–8408.

[87] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[88] Fabian Caba Heilbron and Juan Carlos Niebles. 2014. Collecting and Annotating Human Activities in Web Videos. In *Proceedings of International Conference on Multimedia Retrieval*. ACM, 377.

[89] Samitha Herath, Mehrtash Harandi, and Fatih Porikli. 2017. Going deeper into action recognition: A survey. *Image and Vision Computing* 60 (2017), 4 – 21. https://doi.org/10.1016/j.imavis.2017.01.010 Regularization Techniques for High-Dimensional Data Analysis.

[90] Sepp Hochreiter. 1998. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 06, 02 (1998), 107–116. https://doi.org/10.1142/S0218488598000094 arXiv:https://doi.org/10.1142/S0218488598000094

[91] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long Short-Term Memory. *Neural Computation* 9, 8 (1997), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735 arXiv:https://doi.org/10.1162/neco.1997.9.8.1735

[92] Omar Hommos, Silvia L. Pintea, Pascal S.M. Mettes, and Jan C. van Gemert. 2018. Using phase instead of optical flow for action recognition. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

[93] Berthold K.P. Horn and Brian G. Schunck. 1981. Determining Optical Flow. In *Techniques and Applications of Image Understanding*, James J. Pearson (Ed.), Vol. 0281. International Society for Optics and Photonics, SPIE, 319 – 331. https://doi.org/10.1117/12.965761

[94] Rui Hou, Chen Chen, and Mubarak Shah. 2017. Tube Convolutional Neural Network (T-CNN) for Action Detection in Videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[95] Jian-Fang Hu, Wei-Shi Zheng, Jianhuang Lai, and Jianguo Zhang. 2015. Jointly Learning Heterogeneous Features for RGB-D Activity Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[96] Dong Huang, Shitong Yao, Yi Wang, and Fernando De La Torre. 2014. Sequential Max-Margin Event Detectors. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 410–424.

[97] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. 2017. Densely Connected Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[98] Y. Huang, Q. Dai, and Y. Lu. 2019. Decoupling Localization and Classification in Single Shot Temporal Action Detection. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*. 1288–1293.

[99] Noureldien Hussein, Efstratios Gavves, and Arnold W.M. Smeulders. 2019. Timeception for Complex Action Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[100] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. 2016. A Hierarchical Deep Temporal Model for Group Activity Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[101] H. Idrees, A. R. Zamir, Y. Jiang, A. Gorban, I. Laptev, R. Sukthankar, and M. Shah. 2017. The THUMOS challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding* 155 (2017), 1–23.

[102] Haroon Idrees, Amir R. Zamir, Yu-Gang Jiang, Alex Gorban, Ivan Laptev, Rahul Sukthankar, and Mubarak Shah. 2017. The THUMOS challenge on action recognition for videos "in the wild". *Computer Vision and Image Understanding* 155 (Feb 2017), 1–23. https://doi.org/10.1016/j.cviu.2016.10.018

[103] Sergey Ioffe and Christian Szegedy. 2015. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. arXiv:1502.03167 [cs.LG]

[104] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena. 2016. Recurrent Neural Networks for driver activity anticipation via sensory-fusion architecture. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*. 3118–3125.

[105] Jenny Yuen, B. Russell, Ce Liu, and A. Torralba. 2009. LabelMe video: Building a video database with human annotations. In *2009 IEEE 12th International Conference on Computer Vision*. 1451–1458.

[106] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. 2013. Towards Understanding Action Recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[107] Jingwei Ji, Shyamal Buch, Alvaro Soto, and Juan Carlos Niebles. 2018. End-to-End Joint Semantic Segmentation of Actors and Actions in Video. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[108] S. Ji, W. Xu, M. Yang, and K. Yu. 2013. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (2013), 221–231.

[109] Yanli Ji, Feixiang Xu, Yang Yang, Fumin Shen, Heng Tao Shen, and Wei-Shi Zheng. 2019. A Large-scale Varying-view RGB-D Action Dataset for Arbitrary-view Human Action Recognition. arXiv:1904.10681 [cs.CV]

[110] Boyuan Jiang, MengMeng Wang, Weihao Gan, Wei Wu, and Junjie Yan. 2019. STM: SpatioTemporal and Motion Encoding for Action Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[111] Y. Jiang, Z. Wu, J. Wang, X. Xue, and S. Chang. 2018. Exploiting Feature and Class Relationships in Video Categorization with Regularized Deep Neural Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 2 (2018), 352–364.

[112] Y.-G. Jiang, J. Liu, A. Roshan Zamir, I. Laptev, M. Piccardi, M. Shah, and R. Sukthankar. 2013. THUMOS Challenge: Action Recognition with a Large Number of Classes. /ICCV13-Action-Workshop/.

[113] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. 2014. THUMOS Challenge: Action Recognition with a Large Number of Classes. http://crcv.ucf.edu/THUMOS14/.

[114] Yu-Gang Jiang, Guangnan Ye, S. Chang, Daniel Ellis, and Alexander Loui. 2011. Consumer video understanding: A benchmark database and an evaluation of human and machine performance. 29. https://doi.org/10.1145/1991996.1992025

[115] C. Jin, T. Zhang, W. Kong, T. Li, and G. Li. 2020. Regression Before Classification for Temporal Action Detection. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1–5.

[116] Y. Y. Joefrie and M. Aono. 2019. Action Recognition by Composite Deep Learning Architecture I3D-DenseLSTM. In *2019 International Conference of Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*. 1–6.

[117] Vicky Kalogeiton, Philippe Weinzaepfel, Vittorio Ferrari, and Cordelia Schmid. 2017. Action Tubelet Detector for Spatio-Temporal Action Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[118] Soo Min Kang and Richard P. Wildes. 2016. Review of Action Recognition and Detection Methods. arXiv:1610.06906 [cs.CV]

[119] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale Video Classification with Convolutional Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[120] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The Kinetics Human Action Video Dataset. arXiv:1705.06950 [cs.CV]

[121] Shian-Ru Ke, Hoang Thuc, Yong-Jin Lee, Jenq-Neng Hwang, Jang-Hee Yoo, and Kyoung-Ho Choi. 2013. A Review on Video-Based Human Activity Recognition. *Computers* 2, 2 (Jun 2013), 88–131. https://doi.org/10.3390/computers2020088

[122] T. Kim and R. Cipolla. 2009. Canonical Correlation Analysis of Video Volume Tensors for Action Categorization and Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 8 (2009), 1415–1428.

[123] T. Kim, S. Wong, and R. Cipolla. 2007. Tensor Canonical Correlation Analysis for Action Classification. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[124] Alexander Kläser, Marcin Marszałek, Cordelia Schmid, and Andrew Zisserman. 2012. Human Focused Action Localization in Video. In *Trends and Topics in Computer Vision*, Kiriakos N. Kutulakos (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 219–233.

[125] O. Kliper-Gross, T. Hassner, and L. Wolf. 2012. The Action Similarity Labeling Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 3 (2012), 615–621.

[126] Yu Kong and Yun Fu. 2018. Human Action Recognition and Prediction: A Survey. arXiv:1806.11230 [cs.CV]

[127] Yu Kong, Yunde Jia, and Yun Fu. 2012. Learning Human Interaction by Interactive Phrases. In *Computer Vision – ECCV 2012*, Andrew Fitzgibbon, Svetlana Lazebnik, Pietro Perona, Yoichi Sato, and Cordelia Schmid (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 300–313.

[128] Maryam koohzadi and Nasrollah Charkari. 2017. A Survey on Deep Learning Methods in Human Action Recognition. *IET Computer Vision* 11 (09 2017). https://doi.org/10.1049/iet-cvi.2016.0355

[129] Hema Swetha Koppula, Rudhir Gupta, and Ashutosh Saxena. 2013. Learning human activities and object affordances from RGB-D videos. *The International Journal of Robotics Research* 32, 8 (2013), 951–970. https://doi.org/10.1177/0278364913478446 arXiv:https://doi.org/10.1177/0278364913478446

[130] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-Captioning Events in Videos. In *International Conference on Computer Vision (ICCV)*.

[131] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105. http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

[132] Hilde Kuehne, Ali Arslan, and Thomas Serre. 2014. The Language of Actions: Recovering the Syntax and Semantics of Goal-Directed Human Activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[133] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. 2011. HMDB: A large video database for human motion recognition. In *2011 International Conference on Computer Vision*. 2556–2563.

[134] A. Kurakin, Z. Zhang, and Z. Liu. 2012. A real time system for dynamic hand gesture recognition with a depth sensor. In *2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO)*. 1975–1979.

[135] Jennifer R. Kwapisz, Gary M. Weiss, and Samuel A. Moore. 2011. Activity Recognition Using Cell Phone Accelerometers. *SIGKDD Explor. Newsl.* 12, 2 (March 2011), 74–82. https://doi.org/10.1145/1964897.1964918

[136] Okan Köpüklü, Xiangyu Wei, and Gerhard Rigoll. 2019. You Only Watch Once: A Unified CNN Architecture for Real-Time Spatiotemporal Action Localization. arXiv:1911.06644 [cs.CV]

[137] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. 2008. Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[138] I. Laptev and P. Perez. 2007. Retrieving actions in movies. In *2007 IEEE 11th International Conference on Computer Vision*. 1–8.

[139] Quoc V Le et al. 2015. A tutorial on deep learning part 2: Autoencoders, convolutional neural networks and recurrent neural networks. *Google Brain* (2015), 1–20.

[140] Yann LeCun and M Ranzato. 2013. Deep learning tutorial. In *Tutorials in International Conference on Machine Learning (ICML'13)*. Citeseer, 1–29.

[141] Yooyoung Lee, Jon Fiscus, Afzal Godil, Andrew Delgado, Jim Golden, Lukas Diduch, and Maxime Hubert. 2020. Summary of the 2019 Activity Detection in Extended Videos Prize Challenge. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) Workshops*.

[142] Ang Li, Meghana Thotakuri, David A. Ross, João Carreira, Alexander Vostrikov, and Andrew Zisserman. 2020. The AVA-Kinetics Localized Human Actions Video Dataset. arXiv:2005.00214 [cs.CV]

[143] Jun Li, Xianglong Liu, Mingyuan Zhang, and Deqing Wang. 2020. Spatio-temporal deformable 3D ConvNets with attention for action recognition. *Pattern Recognition* 98 (2020), 107037. https://doi.org/10.1016/j.patcog.2019.107037

[144] Luxuan Li, Tao Kong, Fuchun Sun, and Huaping Liu. 2019. Deep Point-Wise Prediction for Action Temporal Proposal. In *Neural Information Processing*, Tom Gedeon, Kok Wai Wong, and Minho Lee (Eds.). Springer International Publishing, Cham, 475–487.

[145] Minchen Li. 2016. A Tutorial On Backward Propagation Through Time (BPTT) In The Gated Recurrent Unit (GRU) RNN. https://doi.org/10.13140/RG.2.2.32858.98247

[146] S. Li, Z. Tao, K. Li, and Y. Fu. 2019. Visual to Text: Survey of Image and Video Captioning. *IEEE Transactions on Emerging Topics in Computational Intelligence* 3, 4 (2019), 297–312.

[147] W. Li and M. Fritz. 2016. Recognition of ongoing complex activities by sequence prediction over a hierarchical label space. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*. 1–9.

[148] W. Li, Z. Zhang, and Z. Liu. 2010. Action recognition based on a bag of 3D points. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*. 9–14.

[149] Yanghao Li, Cuiling Lan, Junliang Xing, Wenjun Zeng, Chunfeng Yuan, and Jiaying Liu. 2016. Online Human Action Detection Using Joint Classification-Regression Recurrent Neural Networks. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 203–220.

[150] Yingwei Li, Yi Li, and Nuno Vasconcelos. 2018. RESOUND: Towards Action Recognition without Representation Bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[151] Yin Li, Miao Liu, and James M. Rehg. 2018. In the Eye of Beholder: Joint Learning of Gaze and Actions in First Person Video. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[152] Yin Li, Zhefan Ye, and James M. Rehg. 2015. Delving Into Egocentric Actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[153] Yabei Li, Junge Zhang, Yanhua Cheng, Kaiqi Huang, and Tieniu Tan. 2017. Semantics-guided multi-level RGB-D feature fusion for indoor semantic segmentation. 1262–1266. https://doi.org/10.1109/ICIP.2017.8296484

[154] Ivan Lillo, Alvaro Soto, and Juan Carlos Niebles. 2014. Discriminative Hierarchical Modeling of Spatio-Temporally Composable Human Activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[155] Ji Lin, Chuang Gan, and Song Han. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[156] Tianwei Lin, Xu Zhao, and Zheng Shou. 2017. Single Shot Temporal Action Detection. In *Proceedings of the 25th ACM International Conference on Multimedia* (Mountain View, California, USA) *(MM '17)*. Association for Computing Machinery, New York, NY, USA, 988–996. https://doi.org/10.1145/3123266.3123343

[157] Tianwei Lin, Xu Zhao, Haisheng Su, Chongjing Wang, and Ming Yang. 2018. BSN: Boundary Sensitive Network for Temporal Action Proposal Generation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[158] Xiao Lin, Ying-lan Ma, Li-zhuang Ma, and Rui-ling Zhang. 2014. A survey for image resizing. *Journal of Zhejiang University SCIENCE C* 15, 9 (2014), 697–716. https://doi.org/10.1631/jzus.C1400102

[159] Yan-Ching Lin, Min-Chun Hu, Wen-Huang Cheng, Yung-Huan Hsieh, and Hong-Ming Chen. 2012. Human Action Recognition and Retrieval Using Sole Depth Information. In *Proceedings of the 20th ACM International Conference on Multimedia* (Nara, Japan) *(MM '12)*. Association for Computing Machinery, New York, NY, USA, 1053–1056. https://doi.org/10.1145/2393347.2396381

[160] An-An Liu, Wei-Zhi Nie, Yu-Ting Su, Li Ma, Tong Hao, and Zhao-Xuan Yang. 2015. Coupled hidden conditional random fields for RGB-D human action recognition. *Signal Processing* 112 (2015), 74 – 82. https://doi.org/10.1016/j.sigpro.2014.08.038 Signal Processing and Learning Methods for 3D Semantic Analysis.

[161] J. Liu, Jiebo Luo, and M. Shah. 2009. Recognizing realistic actions from videos "in the wild". In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 1996–2003.

[162] J. Liu, A. Shahroudy, M. L. Perez, G. Wang, L. Duan, and A. Kot Chichung. 2019. NTU RGB+D 120: A Large-Scale Benchmark for 3D Human Activity Understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), 1–1.

[163] Kun Liu, Wu Liu, Chuang Gan, Mingkui Tan, and Huadong Ma. 2018. T-c3d: Temporal convolutional 3d network for real-time action recognition. In *Thirty-second AAAI conference on artificial intelligence*.

[164] Yuan Liu, Lin Ma, Yifeng Zhang, Wei Liu, and Shih-Fu Chang. 2019. Multi-Granularity Generator for Temporal Action Proposal. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[165] Fuchen Long, Ting Yao, Zhaofan Qiu, Xinmei Tian, Jiebo Luo, and Tao Mei. 2019. Gaussian Temporal Awareness Networks for Action Localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[166] Bruce D Lucas, Takeo Kanade, et al. 1981. An iterative image registration technique with an application to stereo vision. (1981).

[167] Chenxu Luo and Alan L. Yuille. 2019. Grouped Spatial-Temporal Aggregation for Efficient Action Recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[168] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. 2016. Understanding the Effective Receptive Field in Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett (Eds.). Curran Associates, Inc., 4898–4906. http://papers.nips.cc/paper/6203-understanding-the-effective-receptive-field-in-deep-convolutional-neural-networks.pdf

[169] Chih-Yao Ma, Min-Hung Chen, Zsolt Kira, and Ghassan AlRegib. 2019. TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. *Signal Processing: Image Communication* 71 (2019), 76 – 87. https://doi.org/10.1016/j.image.2018.09.003

[170] Farzaneh Mahdisoltani, Guillaume Berger, Waseem Gharbieh, David Fleet, and Roland Memisevic. 2018. On the effectiveness of task granularity for transfer learning. arXiv:1804.09235 [cs.CV]

[171] A. Mansur, Y. Makihara, and Y. Yagi. 2013. Inverse Dynamics for Action Recognition. *IEEE Transactions on Cybernetics* 43, 4 (2013), 1226–1236.

[172] M. Marszalek, I. Laptev, and C. Schmid. 2009. Actions in context. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2929–2936.

[173] Joanna Materzynska, Guillaume Berger, Ingo Bax, and Roland Memisevic. 2019. The Jester Dataset: A Large-Scale Video Dataset of Human Gestures. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*.

[174] Merriam-Webster. [n.d.]. Action. Retrieved June 7, 2006 from https://www.merriam-webster.com/dictionary/action

[175] R. Messing, C. Pal, and H. Kautz. 2009. Activity recognition using the velocity histories of tracked keypoints. In *2009 IEEE 12th International Conference on Computer Vision*. 104–111.

[176] Media Integration & Communication Center (MICC). 2013. Florence 3D actions dataset.

[177] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfruend, Carl Vondrick, and Aude Oliva. 2018. Moments in Time Dataset: one million videos for event understanding. arXiv:1801.03150 [cs.CV]

[178] Mathew Monfort, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Bowen Pan, Quanfu Fan, Dan Gutfreund, Rogerio Feris, and Aude Oliva. 2019. Multi-Moments in Time: Learning and Interpreting Models for Multi-Action Video Understanding. arXiv:1911.00232 [cs.CV]

[179] Matteo Munaro, Gioia Ballin, Stefano Michieletto, and Emanuele Menegatti. 2013. 3D flow estimation for human action recognition from colored point clouds. *Biologically Inspired Cognitive Architectures* 5 (2013), 42 – 51. https://doi.org/10.1016/j.bica.2013.05.008 Extended versions of selected papers from the Third Annual Meeting of the BICA Society (BICA 2012).

[180] Matteo Munaro, Stefano Michieletto, and Emanuele Menegatti. 2013. An evaluation of 3d motion flow and 3d pose estimation for human action recognition. In *RSS Workshops: RGB-D: Advanced Reasoning with Depth Cameras*.

[181] Farhood Negin, Fırat Özdemir, Ceyhun Burak Akgül, Kamer Ali Yüksel, and Aytül Erçil. 2013. A Decision Forest Based Feature Selection Framework for Action Recognition from RGB-Depth Cameras. In *Image Analysis and Recognition*, Mohamed Kamel and Aurélio Campilho (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 648–657.

[182] Farhood Negin, Pau Rodriguez, Michal Koperski, Adlen Kerboua, Jordi Gonzàlez, Jeremy Bourgeois, Emmanuelle Chapoulie, Philippe Robert, and Francois Bremond. 2018. PRAXIS: Towards Automatic Cognitive Assessment Using Gesture Recognition. *Expert Systems with Applications* (2018).

[183] A. T. Nghiem, F. Bremond, M. Thonnat, and V. Valentin. 2007. ETISEO, performance evaluation for video surveillance systems. In *2007 IEEE Conference on Advanced Video and Signal Based Surveillance*. 476–481.

[184] Phuc Xuan Nguyen, Gregory Rogez, Charless Fowlkes, and Deva Ramanan. 2016. The Open World of Micro-Videos. arXiv:1603.09439 [cs.CV]

[185] B. Ni, Gang Wang, and P. Moulin. 2011. RGBD-HuDaAct: A color-depth video database for human daily activity recognition. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*. 1147–1153.

[186] J. C. Niebles and Li Fei-Fei. 2007. A Hierarchical Model of Shape and Appearance for Human Action Classification. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[187] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. 2013. Berkeley MHAD: A comprehensive Multimodal Human Action Database. In *2013 IEEE Workshop on Applications of Computer Vision (WACV)*. 53–60.

[188] S. Oh, A. Hoogs, A. Perera, N. Cuntoor, C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. Davis, E. Swears, X. Wang, Q. Ji, K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A.

Roy-Chowdhury, and M. Desai. 2011. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR 2011*. 3153–3160.

[189] Omar Oreifej and Zicheng Liu. 2013. HON4D: Histogram of Oriented 4D Normals for Activity Recognition from Depth Sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[190] Nicholas T Ouellette, Haitao Xu, and Eberhard Bodenschatz. 2006. A quantitative study of three-dimensional Lagrangian particle tracking algorithms. *Experiments in Fluids* 40, 2 (2006), 301–313.

[191] Alonso Patron-Perez, Marcin Marszalek, Andrew Zisserman, and Ian Reid. 2010. High Five: Recognising human interactions in TV shows.. In *BMVC*, Vol. 1. Citeseer, 33.

[192] Xiaojiang Peng and Cordelia Schmid. 2016. Multi-region Two-Stream R-CNN for Action Detection. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 744–759.

[193] Asanka G. Perera, Yee Wei Law, and Javaan Chahl. 2018. UAV-GESTURE: A Dataset for UAV Control and Gesture Recognition. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

[194] AJ Piergiovanni and Michael S. Ryoo. 2020. AViD Dataset: Anonymized Videos from Diverse Countries. arXiv:2007.05515 [cs.CV]

[195] H. Pirsiavash and D. Ramanan. 2012. Detecting activities of daily living in first-person camera views. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 2847–2854.

[196] Ronald Poppe. 2010. A survey on vision-based human action recognition. *Image and Vision Computing* 28, 6 (2010), 976 – 990. https://doi.org/10.1016/j.imavis.2009.11.014

[197] OED Online. Oxford Univesity Press. [n.d.]. Action. Retrieved June 7, 2006 from www.oed.com/view/Entry/1938

[198] Zhaofan Qiu, Ting Yao, and Tao Mei. 2017. Learning Spatio-Temporal Representation With Pseudo-3D Residual Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[199] M. A. Rahman and R. Laganière. 2020. Single-Stage End-to-End Temporal Activity Detection in Untrimmed Videos. In *2020 17th Conference on Computer and Robot Vision (CRV)*. 206–213.

[200] Hossein Rahmani, Arif Mahmood, Du Q Huynh, and Ajmal Mian. 2014. HOPC: Histogram of Oriented Principal Components of 3D Pointclouds for Action Recognition. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 742–757.

[201] Hossein Rahmani and Ajmal Mian. 2016. 3D Action Recognition From Novel Viewpoints. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[202] Amir Rasouli. 2020. Deep Learning for Vision-based Prediction: A Survey. arXiv:2007.00095 [cs.CV]

[203] Amir Rasouli, Iuliia Kotseruba, Toni Kunic, and John K. Tsotsos. 2019. PIE: A Large-Scale Dataset and Models for Pedestrian Intention Estimation and Trajectory Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[204] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. 2017. Are They Going to Cross? A Benchmark Dataset and Baseline for Pedestrian Crosswalk Behavior. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 206–213.

[205] Amir Rasouli, Iuliia Kotseruba, and John K Tsotsos. 2018. It's Not All About Size: On the Role of Data Properties in Pedestrian Detection. In *ECCVW*.

[206] Kishore K. Reddy and Mubarak Shah. 2013. Recognizing 50 human action categories of web videos. *Machine Vision and Applications* 24, 5 (2013), 971–981. https://doi.org/10.1007/s00138-012-0450-4

[207] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems 28*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett (Eds.). Curran Associates, Inc., 91–99. http://papers.nips.cc/paper/5638-faster-r-cnn-towards-real-time-object-detection-with-region-proposal-networks.pdf

[208] M. D. Rodriguez, J. Ahmed, and M. Shah. 2008. Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[209] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A Dataset for Movie Description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[210] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. 2012. A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 1194–1201.

[211] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2014. ImageNet Large Scale Visual Recognition Challenge. arXiv:1409.0575 [cs.CV]

[212] M. S. Ryoo and J. K. Aggarwal. 2009. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *2009 IEEE 12th International Conference on Computer Vision*. 1593–1600.

[213] M. S. Ryoo and J. K. Aggarwal. 2010. UT-Interaction Dataset, ICPR contest on Semantic Description of Human Activities (SDHA). http://cvrc.ece.utexas.edu/SDHA2010/Human_Interaction.html.

[214] Michael S. Ryoo and Larry Matthies. 2013. First-Person Activity Recognition: What Are They Doing to Me?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[215] Mohammad Sadegh Aliakbarian, Fatemeh Sadat Saleh, Mathieu Salzmann, Basura Fernando, Lars Petersson, and Lars Andersson. 2017. Encouraging LSTMs to Anticipate Actions Very Early. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[216] Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *WIREs Data Mining and Knowledge Discovery* 8, 4 (2018), e1249. https://doi.org/10.1002/widm.1249 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1249

[217] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip H. S. Torr, and Fabio Cuzzolin. 2017. Spatio-temporal Human Action Localisation and Instance Segmentation in Temporally Untrimmed Videos. arXiv:1707.07213 [cs.CV]

[218] Suman Saha, Gurkirt Singh, Michael Sapienza, Philip H. S. Torr, and Fabio Cuzzolin. 2020. *Spatio-Temporal Action Instance Segmentation and Localisation*. Springer International Publishing, Cham, 141–161. https://doi.org/10.1007/978-3-030-46732-6_8

[219] C. Schuldt, I. Laptev, and B. Caputo. 2004. Recognizing human actions: a local SVM approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, Vol. 3. 32–36 Vol.3.

[220] Lorenzo Seidenari, Vincenzo Varano, Stefano Berretti, Alberto Del Bimbo, and Pietro Pala. 2013. Recognizing Actions from Depth Cameras as Weakly Aligned Multi-part Bag-of-Poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[221] Laura Sevilla-Lara, Yiyi Liao, Fatma Güney, Varun Jampani, Andreas Geiger, and Michael J. Black. 2019. On the Integration of Optical Flow and Action Recognition. In *Pattern Recognition*, Thomas Brox, Andrés Bruhn, and Mario Fritz (Eds.). Springer International Publishing, Cham, 281–297.

[222] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. NTU RGB+D: A Large Scale Dataset for 3D Human Activity Analysis. arXiv:1604.02808 [cs.CV]

[223] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[224] Alex Sherstinsky. 2020. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. *Physica D: Nonlinear Phenomena* 404 (2020), 132306. https://doi.org/10.1016/j.physd.2019.132306

[225] Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data* 6, 1 (2019), 60. https://doi.org/10.1186/s40537-019-0197-0

[226] Zheng Shou, Jonathan Chan, Alireza Zareian, Kazuyuki Miyazawa, and Shih-Fu Chang. 2017. CDC: Convolutional-De-Convolutional Networks for Precise Temporal Action Localization in Untrimmed Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[227] Zheng Shou, Dongang Wang, and Shih-Fu Chang. 2016. Temporal Action Localization in Untrimmed Videos via Multi-Stage CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[228] Bernard Ghanem Shyamal Buch, Victor Escorcia and Juan Carlos Niebles. 2017. End-to-End, Single-Stream Temporal Action Detection in Untrimmed Videos. In *Proceedings of the British Machine Vision Conference (BMVC)*, Gabriel Brostow Tae-Kyun Kim, Stefanos Zafeiriou and Krystian Mikolajczyk (Eds.). BMVA Press, Article 93, 12 pages. https://doi.org/10.5244/C.31.93

[229] Gunnar A. Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. 2018. Charades-Ego: A Large-Scale Dataset of Paired Third and First Person Videos. arXiv:1804.09626 [cs.CV]

[230] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 510–526.

[231] Karen Simonyan and Andrew Zisserman. 2014. Two-Stream Convolutional Networks for Action Recognition in Videos. In *Advances in Neural Information Processing Systems 27*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 568–576. http://papers.nips.cc/paper/5353-two-stream-convolutional-networks-for-action-recognition-in-videos.pdf

[232] Gurkirt Singh, Suman Saha, Michael Sapienza, Philip H. S. Torr, and Fabio Cuzzolin. 2017. Online Real-Time Multiple Spatiotemporal Action Localisation and Prediction. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[233] Tej Singh and Dinesh Kumar Vishwakarma. 2019. Video benchmarks of human action datasets: a review. *Artificial Intelligence Review* 52, 2 (2019), 1107–1154. https://doi.org/10.1007/s10462-018-9651-1

[234] Cees Snoek, Juan Carlos Niebles, Bernard Ghanem, Fabian Caba Heilbron, Humam Alwassel, Victor Escorcia, Ranjay Krishna, Shyamal Buch, and Frost Xu. 2019. International Challenge on Activity Recognition. http://activity-net.org/challenges/2019/index.html

[235] Cees Snoek, Juan Carlos Niebles, Bernard Ghanem, Fabian Caba Heilbron, Humam Alwassel, Victor Escorcia, Ranjay Krishna, Shyamal Buch, and Frost Xu. 2020. International Challenge on Activity Recognition. http://activity-

net.org/challenges/2020/index.html
[236] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. *Information Processing & Management* 45, 4 (2009), 427 – 437. https://doi.org/10.1016/j.ipm.2009.03.002
[237] Y. Song, D. Demirdjian, and R. Davis. 2011. Tracking body and hands for gesture recognition: NATOPS aircraft handling signals database. In *Face and Gesture 2011*. 500–506.
[238] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. arXiv:1212.0402 [cs.CV]
[239] Ralf C. Staudemeyer and Eric Rothstein Morris. 2019. Understanding LSTM – a tutorial into Long Short-Term Memory Recurrent Neural Networks. arXiv:1909.09586 [cs.NE]
[240] Sebastian Stein and Stephen J. McKenna. 2013. Combining Embedded Accelerometers with Computer Vision for Recognizing Food Preparation Activities. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (Zurich, Switzerland) *(UbiComp '13)*. Association for Computing Machinery, New York, NY, USA, 729–738. https://doi.org/10.1145/2493432.2493482
[241] A. Stoian, M. Ferecatu, J. Benois-Pineau, and M. Crucianu. 2015. Scalable action localization with kernel-space hashing. In *2015 IEEE International Conference on Image Processing (ICIP)*. 257–261.
[242] A. Stoian, M. Ferecatu, J. Benois-Pineau, and M. Crucianu. 2016. Fast Action Localization in Large-Scale Video Archives. *IEEE Transactions on Circuits and Systems for Video Technology* 26, 10 (2016), 1917–1930.
[243] Jonathan Stroud, David Ross, Chen Sun, Jia Deng, and Rahul Sukthankar. 2020. D3D: Distilled 3D Networks for Video Action Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.
[244] Jaeyong Sung, Colin Ponce, Bart Selman, and Ashutosh Saxena. 2012. Unstructured human activity detection from rgbd images. In *2012 IEEE international conference on robotics and automation*. IEEE, 842–849.
[245] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper With Convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[246] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the Inception Architecture for Computer Vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
[247] A. t. NGHIEM, F. BREMOND, M. THONNAT, and R. MA. 2007. A New Evaluation Approach for Video Processing Algorithms. In *2007 IEEE Workshop on Motion and Video Computing (WMVC'07)*. 15–15.
[248] Graham W. Taylor, Rob Fergus, Yann LeCun, and Christoph Bregler. 2010. Convolutional Learning of Spatio-temporal Features. In *Computer Vision – ECCV 2010*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 140–153.
[249] L. Taylor and G. Nitschke. 2018. Improving Deep Learning with Generic Data Augmentation. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*. 1542–1547.
[250] Alaa Tharwat. 2018. Classification assessment methods. *Applied Computing and Informatics* (2018). https://doi.org/10.1016/j.aci.2018.08.003
[251] Matteo Tomei, Lorenzo Baraldi, Simone Calderara, Simone Bronzin, and Rita Cucchiara. 2019. STAGE: Spatio-Temporal Attention on Graph Entities for Video Action Detection. arXiv:1912.04316 [cs.CV]
[252] Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research. arXiv:1503.01070 [cs.CV]
[253] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Hervé Jégou. 2020. Fixing the train-test resolution discrepancy: FixEfficientNet. arXiv:2003.08237 [cs.CV]
[254] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning Spatiotemporal Features With 3D Convolutional Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
[255] Du Tran and Alexander Sorokin. 2008. Human Activity Recognition with Metric Learning. In *Computer Vision – ECCV 2008*, David Forsyth, Philip Torr, and Andrew Zisserman (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 548–561.
[256] Du Tran, Heng Wang, Lorenzo Torresani, and Matt Feiszli. 2019. Video Classification With Channel-Separated Convolutional Networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.
[257] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. 2018. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
[258] Grigorios Tsoumakas and Ioannis Katakis. 2007. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)* 3, 3 (2007), 1–13.
[259] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2016. Instance Normalization: The Missing Ingredient for Fast Stylization. arXiv:1607.08022 [cs.CV]

[260] G. Vaquette, A. Orcesi, L. Lucat, and C. Achard. 2017. The DAily Home LIfe Activity Dataset: A High Semantic Activity Dataset for Online Recognition. In *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*. 497–504.

[261] G. Varol, I. Laptev, and C. Schmid. 2018. Long-Term Temporal Convolutions for Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40, 6 (2018), 1510–1517.

[262] R. Vezzani and R. Cucchiara. 2008. Annotation Collection and Online Performance Evaluation for Video Surveillance: The ViSOR Project. In *2008 IEEE Fifth International Conference on Advanced Video and Signal Based Surveillance*. 227–234.

[263] R. Vezzani and R. Cucchiara. 2008. ViSOR: VIdeo Surveillance On-line Repository for annotation retrieval. In *2008 IEEE International Conference on Multimedia and Expo*. 1281–1284.

[264] Heng Wang and Cordelia Schmid. 2013. Action Recognition with Improved Trajectories. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[265] Jindong Wang, Yiqiang Chen, Shuji Hao, Xiaohui Peng, and Lisha Hu. 2019. Deep learning for sensor-based activity recognition: A survey. *Pattern Recognition Letters* 119 (2019), 3 – 11. https://doi.org/10.1016/j.patrec.2018.02.010 Deep Learning for Pattern Recognition.

[266] J. Wang, Z. Liu, Y. Wu, and J. Yuan. 2012. Mining actionlet ensemble for action recognition with depth cameras. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*. 1290–1297.

[267] Jiang Wang, Xiaohan Nie, Yin Xia, Ying Wu, and Song-Chun Zhu. 2014. Cross-view Action Modeling, Learning and Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[268] Keze Wang, Xiaolong Wang, Liang Lin, Meng Wang, and Wangmeng Zuo. 2014. 3D Human Activity Recognition with Reconfigurable Convolutional Neural Networks. In *Proceedings of the 22nd ACM International Conference on Multimedia* (Orlando, Florida, USA) *(MM '14)*. Association for Computing Machinery, New York, NY, USA, 97–106. https://doi.org/10.1145/2647868.2654912

[269] Liangliang Wang, Lianzheng Ge, Ruifeng Li, and Yajun Fang. 2017. Three-stream CNNs for action recognition. *Pattern Recognition Letters* 92 (2017), 33 – 40. https://doi.org/10.1016/j.patrec.2017.04.004

[270] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. 2018. Appearance-and-Relation Networks for Video Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[271] Limin Wang, Yu Qiao, and Xiaoou Tang. 2015. Action Recognition With Trajectory-Pooled Deep-Convolutional Descriptors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[272] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *Computer Vision – ECCV 2016*, Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (Eds.). Springer International Publishing, Cham, 20–36.

[273] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool. 2019. Temporal Segment Networks for Action Recognition in Videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 11 (2019), 2740–2755.

[274] Le Wang, Jinliang Zang, Qilin Zhang, Zhenxing Niu, Gang Hua, and Nanning Zheng. 2018. Action Recognition by an Attention-Aware Temporal Weighted Convolutional Neural Network. *Sensors* 18, 7 (Jun 2018), 1979. https://doi.org/10.3390/s18071979

[275] P. Wang, S. Lien, and M. Lee. 2019. A Learning-Based Prediction Model for Baby Accidents. In *2019 IEEE International Conference on Image Processing (ICIP)*. 629–633.

[276] T. Wang, Y. Chen, Z. Lin, A. Zhu, Y. Li, H. Snoussi, and H. Wang. 2020. RecapNet: Action Proposal Generation Mimicking Human Cognitive Process. *IEEE Transactions on Cybernetics* (2020), 1–12.

[277] Xianyuan Wang, Zhenjiang Miao, Ruyi Zhang, and Shanshan Hao. 2019. I3D-LSTM: A New Model for Human Action Recognition. *IOP Conference Series: Materials Science and Engineering* 569 (aug 2019), 032035. https://doi.org/10.1088/1757-899x/569/3/032035

[278] Y. Wang, K. Huang, and T. Tan. 2007. Human Activity Recognition Based on R Transform. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[279] Ping Wei, Yibiao Zhao, Nanning Zheng, and Song-Chun Zhu. 2013. Modeling 4D Human-Object Interactions for Event and Object Recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[280] Daniel Weinland, Remi Ronfard, and Edmond Boyer. 2006. Free viewpoint action recognition using motion history volumes. *Computer Vision and Image Understanding* 104, 2 (2006), 249 – 257. https://doi.org/10.1016/j.cviu.2006.07.013 Special Issue on Modeling People: Vision-based understanding of a person's shape, appearance, movement and behaviour.

[281] Daniel Weinland, Remi Ronfard, and Edmond Boyer. 2011. A survey of vision-based methods for action representation, segmentation and recognition. *Computer Vision and Image Understanding* 115, 2 (2011), 224 – 241. https://doi.org/10.1016/j.cviu.2010.10.002

[282] Philippe Weinzaepfel, Xavier Martin, and Cordelia Schmid. 2016. Human Action Localization with Sparse Spatial Supervision. arXiv:1605.05197 [cs.CV]

[283] Christian Wolf, Eric Lombardi, Julien Mille, Oya Celiktutan, Mingyuan Jiu, Emre Dogan, Gonen Eren, Moez Baccouche, Emmanuel Dellandréa, Charles-Edmond Bichot, Christophe Garcia, and Bülent Sankur. 2014. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding* 127 (2014), 14 – 30. https://doi.org/10.1016/j.cviu.2014.06.014

[284] Wongun Choi, K. Shahid, and S. Savarese. 2009. What are they doing? : Collective activity classification using spatio-temporal relationship among people. In *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops*. 1282–1289.

[285] Chenxia Wu, Jiemi Zhang, Silvio Savarese, and Ashutosh Saxena. 2015. Watch-n-Patch: Unsupervised Understanding of Actions and Relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[286] Hao-Yu Wu, Michael Rubinstein, Eugene Shih, John Guttag, Frédo Durand, and William Freeman. 2012. Eulerian Video Magnification for Revealing Subtle Changes in the World. *ACM Trans. Graph.* 31, 4, Article 65 (July 2012), 8 pages. https://doi.org/10.1145/2185520.2185561

[287] Xi-Zhu Wu and Zhi-Hua Zhou. 2017. A Unified View of Multi-Label Performance Measures. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, International Convention Centre, Sydney, Australia, 3780–3788. http://proceedings.mlr.press/v70/wu17a.html

[288] Yuxin Wu and Kaiming He. 2018. Group Normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[289] Zuxuan Wu, Xi Wang, Yu-Gang Jiang, Hao Ye, and Xiangyang Xue. 2015. Modeling Spatial-Temporal Clues in a Hybrid Deep Learning Framework for Video Classification. In *Proceedings of the 23rd ACM International Conference on Multimedia* (Brisbane, Australia) *(MM '15)*. Association for Computing Machinery, New York, NY, USA, 461–470. https://doi.org/10.1145/2733373.2806222

[290] H. Xia and Y. Zhan. 2020. A Survey on Temporal Action Localization. *IEEE Access* 8 (2020), 70477–70487.

[291] L. Xia, C. Chen, and J. K. Aggarwal. 2012. View invariant human action recognition using histograms of 3D joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 20–27.

[292] Huijuan Xu, Abir Das, and Kate Saenko. 2017. R-C3D: Region Convolutional 3D Network for Temporal Activity Detection. arXiv:1703.07814 [cs.CV]

[293] Huijuan Xu, Abir Das, and Kate Saenko. 2017. R-C3D: Region Convolutional 3D Network for Temporal Activity Detection. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[294] Ning Xu, Anan Liu, Weizhi Nie, Yongkang Wong, Fuwu Li, and Yuting Su. 2015. Multi-Modal & Multi-View & Interactive Benchmark Dataset for Human Action Recognition. In *Proceedings of the 23rd ACM International Conference on Multimedia* (Brisbane, Australia) *(MM '15)*. Association for Computing Machinery, New York, NY, USA, 1195–1198. https://doi.org/10.1145/2733373.2806315

[295] Yuecong Xu, Jianfei Yang, Haozhi Cao, Kezhi Mao, Jianxiong Yin, and Simon See. 2020. ARID: A New Dataset for Recognizing Action in the Dark. arXiv:2006.03876 [cs.CV]

[296] Xitong Yang, Xiaodong Yang, Ming-Yu Liu, Fanyi Xiao, Larry S. Davis, and Jan Kautz. 2019. STEP: Spatio-Temporal Progressive Learning for Video Action Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[297] Zhenheng Yang, Jiyang Gao, and Ram Nevatia. 2017. Spatio-Temporal Action Detection with Cascade Proposal and Location Anticipation. arXiv:1708.00042 [cs.CV]

[298] Z. Yang, L. Zicheng, and C. Hong. 2013. RGB-Depth feature for 3D human activity recognition. *China Communications* 10, 7 (2013), 93–103.

[299] Rui Yao, Guosheng Lin, Shixiong Xia, Jiaqi Zhao, and Yong Zhou. 2019. Video Object Segmentation and Tracking: A Survey. arXiv:1904.09172 [cs.CV]

[300] Yuancheng Ye, Xiaodong Yang, and YingLi Tian. 2019. Discovering spatio-temporal action tubes. *Journal of Visual Communication and Image Representation* 58 (2019), 515 – 524. https://doi.org/10.1016/j.jvcir.2018.12.019

[301] Serena Yeung, Olga Russakovsky, Ning Jin, Mykhaylo Andriluka, Greg Mori, and Li Fei-Fei. 2018. Every Moment Counts: Dense Detailed Labeling of Actions in Complex Videos. *International Journal of Computer Vision* 126, 2 (2018), 375–389.

[302] Gang Yu, Zicheng Liu, and Junsong Yuan. 2015. Discriminative Orderlet Mining for Real-Time Recognition of Human-Object Interaction. In *Computer Vision – ACCV 2014*, Daniel Cremers, Ian Reid, Hideo Saito, and Ming-Hsuan Yang (Eds.). Springer International Publishing, Cham, 50–65.

[303] J. Yuan, Z. Liu, and Y. Wu. 2011. Discriminative Video Pattern Search for Efficient Action Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 9 (2011), 1728–1743.

[304] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. 2015. Beyond Short Snippets: Deep Networks for Video Classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[305] K. Yun, J. Honorio, D. Chattopadhyay, T. L. Berg, and D. Samaras. 2012. Two-person interaction detection using body-pose features and multiple instance learning. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. 28–35.

[306] C. Zach, T. Pock, and H. Bischof. 2007. A Duality Based Approach for Realtime TV-L1 Optical Flow. In *Pattern Recognition*, Fred A. Hamprecht, Christoph Schnörr, and Bernd Jähne (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 214–223.

[307] Kuo-Hao Zeng, William B. Shen, De-An Huang, Min Sun, and Juan Carlos Niebles. 2017. Visual Forecasting by Imitating Dynamics in Natural Sequences. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[308] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. 2019. Graph Convolutional Networks for Temporal Action Localization. arXiv:1909.03252 [cs.CV]

[309] Chenyang Zhang and Yingli Tian. 2012. RGB-D camera-based daily living activity recognition. *Journal of computer vision and image processing* 2, 4 (2012), 12.

[310] Hong-Bo Zhang, Yi-Xiang Zhang, Bineng Zhong, Qing Lei, Lijie Yang, Ji-Xiang Du, and Duan-Sheng Chen. 2019. A Comprehensive Survey of Vision-Based Human Action Recognition Methods. *Sensors* 19, 5 (2019). https://doi.org/10.3390/s19051005

[311] Jing Zhang, Wanqing Li, Philip O. Ogunbona, Pichao Wang, and Chang Tang. 2016. RGB-D-based action recognition datasets: A survey. *Pattern Recognition* 60 (2016), 86 – 105. https://doi.org/10.1016/j.patcog.2016.05.019

[312] Yubo Zhang, Pavel Tokmakov, Martial Hebert, and Cordelia Schmid. 2019. A Structured Model for Action Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[313] Zhang Zhang, Kaiqi Huang, and Tieniu Tan. 2008. Multi-thread Parsing for Recognizing Complex Events in Videos. In *Computer Vision – ECCV 2008*, David Forsyth, Philip Torr, and Andrew Zisserman (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 738–751.

[314] Z. Zhang, K. Huang, T. Tan, and L. Wang. 2007. Trajectory Series Analysis based Event Rule Induction for Visual Surveillance. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8.

[315] Hang Zhao, Antonio Torralba, Lorenzo Torresani, and Zhicheng Yan. 2017. HACS: Human Action Clips and Segments Dataset for Recognition and Temporal Localization. arXiv:1712.09374 [cs.CV]

[316] He Zhao and Richard P. Wildes. 2019. Spatiotemporal Feature Residual Propagation for Action Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[317] Hang Zhao, Zhicheng Yan, Heng Wang, Lorenzo Torresani, and Antonio Torralba. 2017. SLAC: A Sparsely Labeled Dataset for Action Classification and Localization. (12 2017).

[318] Yue Zhao, Yuanjun Xiong, and Dahua Lin. 2018. Trajectory Convolution for Action Recognition. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates, Inc., 2204–2215. http://papers.nips.cc/paper/7489-trajectory-convolution-for-action-recognition.pdf

[319] Yue Zhao, Yuanjun Xiong, Limin Wang, Zhirong Wu, Xiaoou Tang, and Dahua Lin. 2017. Temporal Action Detection With Structured Segment Networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[320] Zhe Lin, Zhuolin Jiang, and L. S. Davis. 2009. Recognizing actions by shape-motion prototype trees. In *2009 IEEE 12th International Conference on Computer Vision*. 444–451.

[321] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. 2018. Temporal Relational Reasoning in Videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[322] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2017).

[323] Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[324] Yipin Zhou and Tamara L. Berg. 2015. Temporal Perception and Prediction in Ego-Centric Video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

[325] Fan Zhu, Ling Shao, Jin Xie, and Yi Fang. 2016. From handcrafted to learned representations for human action recognition: A survey. *Image and Vision Computing* 55 (2016), 42 – 52. https://doi.org/10.1016/j.imavis.2016.06.007 Handcrafted vs. Learned Representations for Human Action Recognition.

[326] Yi Zhu, Zhenzhong Lan, Shawn Newsam, and Alexander Hauptmann. 2019. Hidden Two-Stream Convolutional Networks for Action Recognition. In *Computer Vision – ACCV 2018*, C. V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler (Eds.). Springer International Publishing, Cham, 363–378.

[327] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. 2018. ECO: Efficient Convolutional Network for Online Video Understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

A ADDITIONAL TABLES

Table 3. 137 video action datasets are sorted by release year. Tabular information includes dataset name, year of publication, citations on Google Scholar as of August 2020, number of action classes, number of action instances, actors: human (H) and/or non-human (N), annotations: action class (C), temporal markers (T), spatiotemporal bounding boxes/masks (S), and theme/purpose..

			Action		Actors		Annotations			
Video Dataset	Year	Cited	Classes	Instances	H	N	C	T	S	Theme/Purpose
KTH [219]	2004	3,853	6	2,391	✓		✓			B/W, static background
CAVIAR [61]	2004	49	9	>28	✓		✓		✓	surveillance
Weizmann [15]	2005	1,890	10	90	✓		✓			human motions
ViSOR [262, 263]	2005	47	n/a	n/a	✓		✓			surveillance
ETISEO [183, 247]	2005	183	15	n/a	✓		✓			human motions
IXMAS [280]	2006	977	13	390	✓		✓			B/W, partial occlusion
UCF Aerial [62]	2007	n/a	9	n/a	✓		✓		✓	aerial-view
CASIA Action [278, 314]	2007	242	15	1,446	✓		✓			multi-view, outdoors
Coffee & Cigarettes [138]	2007	491	2	246	✓		✓	✓	✓	movies and TV
UIUC Action [255]	2008	378	14	532	✓		✓			action repetition
UCF Sports [208]	2008	1,269	10	150	✓		✓		✓	sports
UCF ARG [63]	2008	n/a	10	480	✓		✓		✓	multi-view, aerial-view
Hollywood (HOHA) [137]	2008	3,727	8	n/a	✓		✓			movies
Cambridge-Gesture [122]	2008	298	9	900	✓		✓			gestures
BEHAVE [18]	2009	134	10	163	✓		✓		✓	human-human interaction
URADL [175]	2009	574	10	150	✓		✓			daily activities
UCF11 [161]	2009	1,183	11	3,040	✓		✓			web videos
MSR-I [303]	2009	181	3	n/a	✓		✓		✓	activities
i3DPost MuHAVi [78]	2009	179	12	>1,000	✓		✓			multi-view, studio
Hollywood2 [172]	2009	1,312	12	3,669	✓		✓			movies
Collective Activity [284]	2009	201	5	44	✓		✓		✓	group activities
LabelMe [105]	2009	203	70	>300	✓	✓	✓	✓	✓	actor-object interactions
Keck Gesture [320]	2009	349	14	126	✓		✓			gestures
DLSPB [50]	2009	307	2	89	✓		✓	✓		movies and TV
Hollywood-Localization [124]	2010	159	2	408	✓		✓		✓	movies
VideoWeb [47]	2010	55	51	368	✓		✓			multi-view, tasks
UT-Tower [30, 31]	2010	61	9	648	✓		✓		✓	aerial-view, human motions
UT-Interaction [212, 213]	2010	593	6	60	✓		✓			human-human interaction
UCF50 [206]	2010	537	50	>5,000	✓		✓			web videos, expand UCF11
TV-Human Interaction [191]	2010	176	4	300	✓		✓		✓	TV, human-human interaction
Olympic Sports [313]	2010	745	16	800	✓		✓			sports
MSR-II [22]	2010	232	3	n/a	✓		✓		✓	activities
MSR-Action3D [148]	2010	1,285	20	4,020	✓		✓			RGB-D, gestures and motions
CMU MoCap [1]	2010	n/a	n/a	2,605	✓		✓			RGB-D, human motions
VIRAT [188]	2011	536	23	~10,000	✓		✓	✓	✓	surveillance, aerial-view
HMDB51 [133]	2011	1,928	51	~7,000	✓		✓			human motions
CAD-60 [244]	2011	549	12	60	✓		✓		✓	RGB-D, daily activities
GTEA [57, 152]	2011	492	71	526	✓		✓			egocentric, kitchen
CCV [114]	2011	288	*20	9,317	✓		✓			web videos
ChaLearn [2]	2011	n/a	86	50,000	✓		✓			RGB-D, gestures and motions
RGBD-HuDaAct [185]	2011	393	12	1,189	✓		✓			RGB-D, daily activities
NATOPS [237]	2011	111	24	400	✓		✓			aircraft hand signaling
GTEA Gaze [56]	2012	331	40	331	✓		✓	✓		egocentric, kitchen
GTEA Gaze+ [56, 152]	2012	165	44	1,958	✓		✓	✓		egocentric, kitchen
BIT-Interaction [127]	2012	109	8	400	✓		✓			human-human interaction
LIRIS [283]	2012	60	10	n/a	✓		✓		✓	RGB-D, office environment
MSR-DailyActivity3D [266]	2012	1,339	16	320	✓		✓			RGB-D, gestures
UCF101 [238]	2012	2,470	101	13,320	✓		✓			web videos, expand UCF50
UTKinect-A [291]	2012	1,216	10	200	✓		✓			RGB-D, indoors
MSR-Gesture3D [134]	2012	317	12	n/a	✓		✓			RGB-D, gestures
ASLAN [125]	2012	106	432	3,631	✓		✓			web videos, action similarity
ADL [195]	2012	619	18	~1,200	✓		✓		✓	egocentric, daily activities
ACT4^2 [37]	2012	122	14	6,844	✓		✓			RGB-D, multi-view
SBU-Kinect-Interaction [305]	2012	339	8	~170	✓		✓	✓		RGB-D, human-human inter.
MPII-Cooking [210]	2012	436	65	5,609	✓		✓	✓		kitchen, fine-grained actions
Osaka Kinect [171]	2012	31	10	80	✓		✓			RGB-D, gestures

*Only a portion of classes are actions. Some are objects or visual tags.

(continued on next page)

(continued from previous page)

			Action		Actors		Annotations			
Video Dataset	Year	Cited	Classes	Instances	H	N	C	T	S	Theme/Focus
DHA [159]	2012	66	23	483	✓		✓			RGB-D, gestures and motions
Falling Event [309]	2012	138	5	200	✓		✓			RGB-D, daily activities
G3D [16, 17]	2012	207	20	659	✓		✓	✓		RGB-D, gaming gestures
MSR-3DActionPairs [189]	2013	866	12	360	✓		✓			RGB-D, gestures
Multiview 3D Event [279]	2013	119	8	3,815	✓		✓			RGB-D, multi-view
RGBD-SAR [298]	2013	29	12	810	✓		✓			RGB-D, monitoring seniors
CAD-120 [129]	2013	587	10	120	✓		✓		✓	RGB-D, daily activities
JPL Interaction [214]	2013	253	7	∼85	✓		✓	✓		egocentric, human-human inter.
MHAD [187]	2013	336	11	∼650	✓		✓			RGB-D, multi-view, gestures
Florence3D [176, 220]	2013	189	9	213	✓		✓			RGB-D , gestures
THUMOS'13 [101, 112, 238]	2013	146	**101	13,320	✓		✓		✓	web videos, extend UCF101
J-HMDB-21 [106]	2013	458	51	928	✓		✓		✓	re-annotate HMDB51 subset
Mivia [23]	2013	21	7	490	✓		✓			RGB-D, daily activities
IAS-lab [179, 180]	2013	31	15	540	✓		✓			RGB-D, human motions
WorkoutSU-10 [181]	2013	66	10	1,200	✓		✓			RGB-D, group activities
50Salads [240]	2013	177	17	966	✓		✓	✓		RGB-D, kitchen
UWA3D-I [200]	2014	141	30	∼900	✓		✓			RGB-D, multi-view
MANIAC [6]	2014	43	8	120	✓		✓	✓		RGB-D, ego., manipulations
Breakfast Action [132]	2014	203	48	11,267	✓		✓	✓		kitchen
Northwester-UCLA [267]	2014	222	10	1,475	✓		✓			RGB-D, multi-view
Sports-1M [119]	2014	4,361	487	1,000,000	✓		✓			multi-label, sports
ORGBD (3D Online) [302]	2014	136	7	336	✓		✓			RGB-D, human-object inter.
THUMOS'14 [101, 113]	2014	146	***101	15,904	✓		✓	✓		extends THUMOS'13
Office Activity [268]	2014	94	20	1,180	✓		✓			RGB-D, office environment
Composable [154]	2014	81	16	693	✓		✓			RGB-D, gestures and motions
CMU-MAD [96]	2014	80	20	1,400	✓		✓	✓		RGB-D, gestures and motions
FFPA [324]	2015	48	5	591	✓		✓			egocentric, daily activities
TJU [160]	2015	69	15	1,200	✓		✓			RGB-D, static background
M^2I [294]	2015	23	22	1,760	✓		✓			RGB-D, multi-view
FCVID [111]	2015	219	*239	91,223	✓		✓			web videos, diverse categories
ActivityNet100 (v1.2) [21]	2015	797	100	10,733	✓		✓	✓		untrimmed web videos
THUMOS'15 [79, 101]	2015	146	***101	21,037	✓		✓	✓		extends THUMOS'14
MEXaction [241, 242]	2015	16/3	2	1,108	✓		✓	✓		culturally relevant actions
MEXaction2 [42]	2015	n/a	2	1,975	✓		✓	✓		extends MEXaction
Watch-n-Patch [285]	2015	119	21	∼2,000	✓		✓	✓		RGB-D, daily activities
TVSeries [45]	2016	109	30	6,231	✓		✓	✓		TV
OAD [149]	2016	109	10	n/a	✓		✓	✓		RGB-D, daily activities
CONVERSE [51]	2016	20	7	n/a	✓		✓	✓		RGB-D, human-human inter.
OA [147]	2016	11	48	480	✓		✓			action semantic hierarchy
Volleyball [100]	2016	215	6	1,643	✓		✓			sports (volleyball motions)
UWA3D-II [201]	2016	117	30	1,075	✓		✓			RGB-D, multi-view
ActivityNet200 (v2.3) [21]	2016	797	200	23,064	✓		✓	✓		untrimmed web videos
YouTube-8M [4]	2016	607	*n/a	n/a	✓		✓			multi-label
Charades [230]	2016	343	157	66,500	✓		✓	✓		crowd-sourced, daily activities
NTU RGB-D [222]	2016	792	60	56,880	✓		✓			RGB-D, multi-view
Micro-Videos [184]	2016	27	*n/a	n/a	✓	✓	✓			micro-videos (e.g. Vine, Tik-Tok)
JAAD [204, 205]	2017	53	n/a	654	✓		✓		✓	pedestrians
DAHLIA [260]	2017	9	7	51	✓		✓			RGB-D, daily activities
PKU-MMD [39]	2017	67	51	3,366	✓		✓			RGB-D, multi-view
SYSU 3DHOI [95]	2017	302	12	480	✓		✓			RGB-D, human-object inter.
DALY [282]	2017	26	10	3,600	✓		✓		✓	daily activities
Okutama Action [12]	2017	55	12	4,700	✓		✓		✓	aerial view
Kinetics-400 [120]	2017	810	400	306,245	✓		✓			diverse web videos
AVA [81]	2017	270	80	>392,416	✓		✓		✓	atomic visual actions
Something-Something [80]	2017	182	174	108,499	✓		✓			human-object inter.
SLAC [317]	2017	19	200	∼1,750,000	✓		✓	✓		sparse-labelled web videos
Moments in Time (MiT) [177]	2017	137	339	836,144	✓	✓	✓			intra-class variation, web videos
MultiTHUMOS [301]	2017	231	65	∼16,000	✓		✓	✓		multi-label, extends THUMOS
$VIENA^2$ [7]	2018	7	25	15,000	✓	✓	✓	✓		pedestrians and vehicles
PRAXIS Gesture [182]	2018	16	29	∼4,600	✓		✓			RGB-D, gestures
UAV-GESTURE [193]	2018	10	13	119	✓		✓		✓	aerial-view, gestures
Diving48 [150]	2018	25	48	18,404	✓		✓			diving motions (sports)
EPIC-KITCHENS-55 [43]	2018	209	125	39,594	✓		✓	✓	✓	egocentric, kitchen
YouCook2 [323]	2018	96	n/a	∼15,400	✓		✓	✓		web videos, kitchen

*Only a portion of classes are actions. Some are objects or visual tags.
**Only 24 classes have temporal annotations. This subset is known as UCF101-24.
***Only 20 classes have temporal annotations.

(continued on next page)

(continued from previous page)

			Action		Actors		Annotations			
Video Dataset	Year	Cited	Classes	Instances	H	N	C	T	S	Theme/Focus
Kinetics-600 [24]	2018	52	600	495,547	✓		✓			extends Kinetics-400
VLOG [64]	2018	41	30	~122,000	✓		✓	✓		web videos, human-object inter.
EGTEA Gaze+ [151]	2018	52	106	10,325	✓		✓	✓	✓	egocentric, kitchen
Something-Something-v2 [170]	2018	5	174	220,847	✓		✓			extends Something-Something
Charades-Ego [229]	2018	19	157	68,536	✓		✓	✓		egocentric, daily activities
Youtube-8M Segments [4]	2019	n/a	*n/a	n/a	✓		✓	✓		multi-label, extends YT-8M
Jester [173]	2019	12	27	148,092	✓		✓			crowd-sourced, gestures
LSVV-HRI [109]	2019	4	83	25,600	✓		✓			RGB-D, human-robot inter.
PIE [203]	2019	10	6	~1,800	✓		✓		✓	pedestrians
Kinetics-700 [25]	2019	33	700	~650,000	✓		✓			extends Kinetics-600
Multi-MiT [178]	2019	1	313	~1,020,000	✓	✓	✓			multi-label, extends MiT
HACS Clips [315]	2019	31	200	~1,500,000	✓		✓			trimmed web videos
HACS Segments [315]	2019	31	200	~139,000	✓		✓	✓		extends and improves SLAC
NTU RGB-D 120 [162]	2019	55	120	114,480	✓		✓			extends NTU RGB-D 60
EPIC-KITCHENS-100 [44]	2020	6	97	~90,000	✓		✓	✓	✓	extends EPIC-KITCHENS-55
AVA-Kinetics [142]	2020	5	80	>238,000	✓		✓		✓	adds annotations, AVA+Kinetics
ARID [295]	2020	0	11	3,784	✓		✓			dark (low-lighting) videos
AViD [194]	2020	0	887	~450,000	✓	✓	✓			diverse, anonymized faces

*Only a portion of classes are actions. Some are objects or visual tags.

Table 4. Prominent Video Action Understanding Challenges 2013-2020.

Workshop	Year	Conf.	Problem	Dataset(s)	Metric(s)	#Teams
THUMOS [112]	2013	ICCV	AR	UCF101	average accuracy	17
			SAL/D	UCF101-24	ROC AUC sIoU@0.2	n/a
THUMOS [113]	2014	ECCV	AR	UCF101+	mAP	11
			TAL/D	UCF101-20	mAP tIoU@{0.1,0.2,0.3,0.4,0.5}	3
THUMOS [79, 102]	2015	CVPR	AR	UCF101+1	mAP	11
			TAL/D	UCF101-20	mAP tIoU@{0.1,0.2,0.3,0.4,0.5}	1
ActivityNet [21]	2016	CVPR	AR	ActivityNet 1.3	mAP, Top-1 accuracy, Top-3 accuracy	26
			TAL/D	ActivityNet 1.3	mAP-50, mAP-75, average-mAP	6
ActivityNet [72]	2017	CVPR	AR	ActivityNet 1.3	Top-1 error	n/a
			AR	Kinetics-400	average(Top-1 error, Top-5 error)	31
			TAP	ActivityNet 1.3	AR-AN AUC	17
			TAL/D	ActivityNet 1.3	mAP tIoU@0.5:0.05:0.95	17
ActivityNet [71]	2018	CVPR	TAP	ActivityNet 1.3	AR-AN AUC	55
			TAL/D	ActivityNet 1.3	mAP tIoU@0.5:0.05:0.95	43
			AR	Kinetics-600	average(Top-1 error, Top-5 error)	13
			SAL/D	AVA	frame-mAP sIoU@0.5	23
			AR	MiT (full-track)	average(Top-1 acc, Top-5 acc)	29
			AR	MiT (mini-track)	average(Top-1 acc, Top-5 acc)	12
ActivityNet [141, 234]	2019	CVPR	TAP	ActivityNet 1.3	AR-AN AUC	72
			TAL/D	ActivityNet 1.3	mAP tIoU@0.5:0.05:0.95	23
			AR	Kinetics-700	average(Top-1 error, Top-5 error)	15
			SAL/D	AVA	frame mAP sIoU@0.5	32
			AR	EPIC-KITCHENS-55	micro-avg Top-1,5 acc, macro-AP,AR	39
			AP	EPIC-KITCHENS-55	micro-avg Top-1,5 acc, macro-AP,AR	19
			TAL/D	VIRAT	P_{rate}@$miss_{FA}$	42
Multi-modal [3]	2019	ICCV	AR	Multi-MiT	mAP	10
			TAL/D	HACS Segments	mAP tIoU@0.5:0.05:0.95	5
ActivityNet [235]	2020	CVPR	TAL/D	ActivityNet 1.3	mAP tIoU@0.5:0.05:0.95	n/a
			AR	Kinetics-700	average(Top-1 error, Top-5 error)	n/a
			SAL/D	AVA	frame mAP sIoU@0.5	n/a
			TAL/D	VIRAT	P_{rate}@$miss_{FA}$	11
			TAL/D	HACS Segments	mAP tIoU@0.5:0.05:0.95	22
			TAL/D	HACS Clips+Seg.	mAP tIoU@0.5:0.05:0.95	13

AR = Action Recognition
TAP = Temporal Action Proposal
TAL/D = Temporal Action Localization/Detection
SAL/D = Spatiotemporal Action Localization/Detection