

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE JUNE 2023	2. REPORT TYPE JOURNAL ARTICLE (POST PRINT)	3. DATES COVERED <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 50%;">START DATE JANUARY 2020</td> <td style="width: 50%;">END DATE OCTOBER 2021</td> </tr> </table>		START DATE JANUARY 2020	END DATE OCTOBER 2021	
START DATE JANUARY 2020	END DATE OCTOBER 2021					
4. TITLE AND SUBTITLE VIDEO ACTION UNDERSTANDING						
5a. CONTRACT NUMBER FA8750-19-2-1000		5b. GRANT NUMBER				
5d. PROJECT NUMBER		5c. PROGRAM ELEMENT NUMBER 62788F				
5e. TASK NUMBER		5f. WORK UNIT NUMBER P27U				
6. AUTHOR(S) Matthew S. Hutchinson, Vijay N. Gadepally						
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Massachusetts Institute of Technology (MIT) 77 Massachusetts Ave Cambridge MA 02139			8. PERFORMING ORGANIZATION REPORT NUMBER			
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) Air Force Research Laboratory/RIEA 525 Brooks Road Rome NY 13441-4505		10. SPONSOR/MONITOR'S ACRONYM(S) AFRL/RI	11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-RI-RS-TP-2023-031			
12. DISTRIBUTION/AVAILABILITY STATEMENT Distribution Approved for Public Release; Distribution Unlimited. This report is the result of contracted fundamental research deemed exempt from public affairs security and policy review in accordance with SAF/AQR memorandum dated 10 Dec 08 and AFRL/CA policy clarification memorandum dated 16 Jan 09.						
13. SUPPLEMENTARY NOTES This work was funded in whole or in part by Department of the Air Force contract number FA8750-19-2-1000. The U.S. Government has for itself and others acting on its behalf an unlimited, paid-up, nonexclusive, irrevocable worldwide license to use, modify, reproduce, release, perform, display, or disclose the work by or on behalf of the Government. All other rights are reserved by the copyright owner.						
14. ABSTRACT Many believe that the successes of deep learning on image understanding problems can be replicated in the realm of video understanding. However, due to the scale and temporal nature of video, the span of video understanding 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 this emerging and rapidly evolving field. With a pedagogical emphasis, this tutorial introduces and systematizes fundamental topics, basic concepts, and notable examples in supervised video action understanding. Specifically, we clarify a taxonomy of action problems, catalog and highlight video datasets, describe common video data preparation methods, present the building blocks of state-of-the-art deep learning model architectures, and formalize domain-specific metrics to baseline proposed solutions. This tutorial is intended to be accessible to a general computer science audience and assumes a conceptual understanding of supervised learning.						
15. SUBJECT TERMS Action detection, action localization, action prediction, action proposal, action recognition, action understanding, video understanding						
16. SECURITY CLASSIFICATION OF: <table border="1" style="width: 100%; border-collapse: collapse;"> <tr> <td style="width: 33%;">a. REPORT U</td> <td style="width: 33%;">b. ABSTRACT U</td> <td style="width: 33%;">c. THIS PAGE U</td> </tr> </table>			a. REPORT U	b. ABSTRACT U	c. THIS PAGE U	17. LIMITATION OF ABSTRACT SAR
a. REPORT U	b. ABSTRACT U	c. THIS PAGE U				
18. NUMBER OF PAGES 28						
19a. NAME OF RESPONSIBLE PERSON ROSEANNE PELLI			19b. PHONE NUMBER (Include area code) N/A			

Received September 3, 2021, accepted September 18, 2021, date of publication September 24, 2021, date of current version October 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3115476

Video Action Understanding

MATTHEW S. HUTCHINSON^{ID} AND VIJAY N. GADEPALLY^{ID}, (Senior Member, IEEE)

Massachusetts Institute of Technology Lincoln Laboratory Supercomputing Center, Lexington, MA 02421, USA

Corresponding author: Matthew S. Hutchinson (hutchinson@alum.mit.edu)

This work was supported by the United States Air Force Research Laboratory and the United States Air Force Artificial Intelligence Accelerator and was accomplished under Cooperative Agreement FA8750-19-2-1000.

ABSTRACT Many believe that the successes of deep learning on image understanding problems can be replicated in the realm of video understanding. However, due to the scale and temporal nature of video, the span of video understanding 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 this emerging and rapidly evolving field. With a pedagogical emphasis, this tutorial introduces and systematizes fundamental topics, basic concepts, and notable examples in supervised video action understanding. Specifically, we clarify a taxonomy of action problems, catalog and highlight video datasets, describe common video data preparation methods, present the building blocks of state-of-the-art deep learning model architectures, and formalize domain-specific metrics to baseline proposed solutions. This tutorial is intended to be accessible to a general computer science audience and assumes a conceptual understanding of supervised learning.

INDEX TERMS Action detection, action localization, action prediction, action proposal, action recognition, action understanding, video understanding.

I. INTRODUCTION

Video understanding is a natural extension of deep learning research efforts in computer vision. The image understanding field greatly benefited 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 [1], [2]. Hikvision Model D scores 90.99% Top-5 accuracy on the Places2 scene classification task [2], [3]. High-resolution network - object contextual representations (HRNet-OCR) with hierarchical multi-scale attention (HMS) yields a mean intersection over union (IoU) of 85.1% on the Cityscapes semantic segmentation task [4], [5]. Naturally, many hope that deep learning methods can achieve similar levels of success on video understanding problems.

Drawing from Diba *et al.* [6], *semantic video understanding* is a combination of understanding the scene/environment, objects, actions, events, attributes, and concepts. This article focuses on the action component and is presented as a tutorial

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott^{ID}.

that introduces a common set of terms and tools, explains basic and fundamental concepts, and provides 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.

A. ACTION UNDERSTANDING

While the literature often uses the terms *action* and *activity* synonymously [7]–[9], we prefer to use action in this article for a few reasons. First, action is the dominant term used across the field, so we would need a significant reason to divert from that term. Second, the use of activity is generally biased towards human actors rather than non-human actors and phenomenon. Examples of non-human actors and phenomenon performing actions include a dog running, a cloud floating, and a wheel turning. We prefer action over activity for its broader human and non-human applicability. Third, activity recognition is a term already used in several non-video domains [10]–[12]. Meanwhile, action recognition is primarily a computer vision and video-based term.

But what is an action? Kang and Wildes [13] consider an action to be “a motion created by the human body, which may or may not be cyclic.” Zhu *et al.* [14] define action

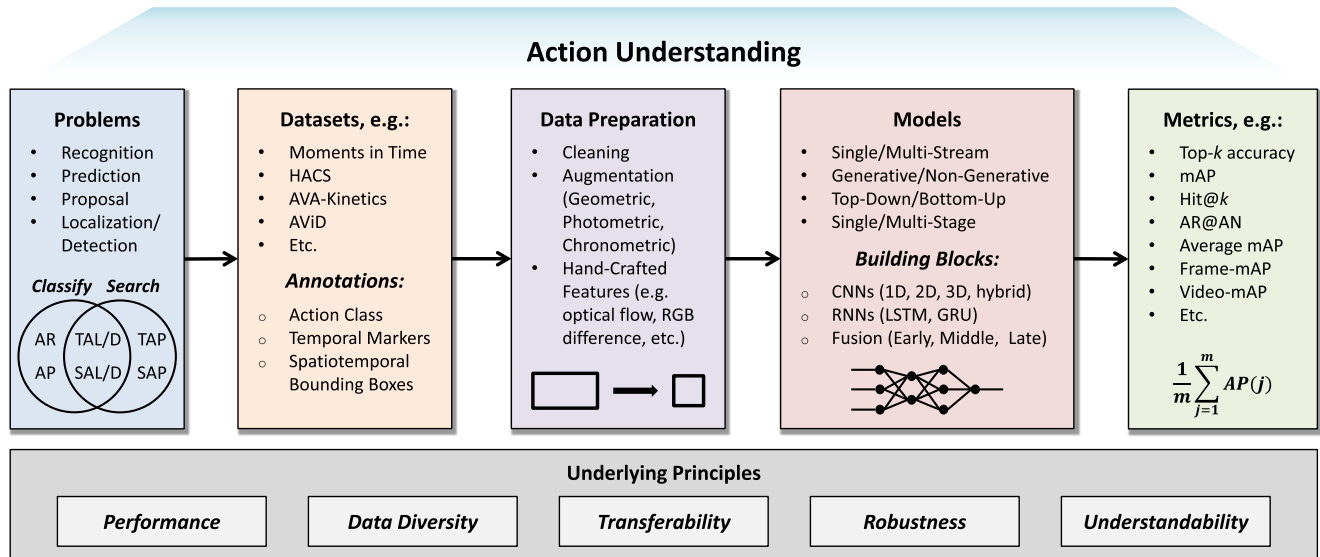


FIGURE 1. Overview of action understanding steps (problem formulation, dataset selection, data preparation, model development, and metric-based evaluation) and underlying principles (computational performance, data diversity, transferability, robustness, and understandability. This serves as the framework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

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 [9], [14], [15]. Unlike these surveys, we 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 [16], [17]; 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 [18], [19]. 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 Fig. 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.

B. RELATED WORK AND OUR CONTRIBUTION

Table 1 shows a selection of surveys and tutorials on action understanding written in the last decade. Yet, few focus on more than one or two action problems or present more than a narrow coverage of datasets. Additionally, the majority only consider a narrow (human) definition of actions and have little or no discussion of metrics. Of the more recent examples, Kong and Fu [27], Xia and Zhan [31], and Rasouli [32] are

TABLE 1. Coverage of surveys and tutorials on action understanding. Tabular information includes year of publication, action coverage: human (H) and/or non-human (N), topic coverage: datasets (Ds), metrics (Mc), models/methods (Md), and problem coverage: 1) action recognition; 2) action proposal; 3) temporal action proposal; 4) temporal action localization/detection; and 5) spatiotemporal action localization/detection.

Survey/Tutorial	Year	Actions		Topics			Problems				
		H	N	Ds	Mc	Md	1	2	3	4	5
Poppe [20]	2010	✓		✓	✓	✓	✓				
Weinland et al. [21]	2011	✓		✓	✓	✓	✓				
Ahad et al. [22]	2011	✓		✓	✓	✓			✓		
Chaquet et al. [8]	2013	✓		✓	✓	✓	✓				
Guo and Lai [15]	2014	✓		✓	✓	✓	✓	✓			
Cheng et al. [9]	2015	✓		✓	✓	✓	✓	✓			
Zhu et al. [14]	2016	✓		✓	✓	✓	✓				
Kang and Wildes [13]	2016	✓		✓	✓	✓	✓	✓		✓	
Zhang et al. [23]	2016	✓		✓	✓	✓	✓			✓	
Herath et al. [24]	2017	✓		✓	✓	✓	✓				
Koohzadi and Charkari [25]	2017	✓		✓	✓	✓	✓				
Asadi-Aghbolaghi et al. [26]	2017	✓		✓	✓	✓	✓				
Kong and Fu [27]	2018	✓		✓	✓	✓	✓	✓			
Zhang et al. [28]	2019	✓		✓	✓	✓	✓	✓		✓	
Bhoi [29]	2019	✓	✓	✓	✓	✓					✓
Singh and Vishwakarma [30]	2019	✓		✓	✓	✓	✓			✓	
Xia and Zhan [31]	2020	✓	✓	✓	✓	✓			✓	✓	
Rasouli [32]	2020	✓	✓	✓	✓	✓	✓				
Yadav et al. [33]	2021	✓		✓	✓	✓	✓	✓		✓	
Ours	2021	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

the most thorough in their independent directions. We recommend these works for a more advanced analysis and comparison of deep learning models. Relative to the literature noted above, we present this article as a tutorial to introduce and systematize a balance of topics across datasets, methods, and metrics rather than a deeply technical methods-heavy survey. To that goal, this tutorial contributes the following:

- Clear definitions of recognition, prediction, proposal, and localization/detection video action problems.

- An extensive catalog of video action datasets and discussion of those most relevant to each action problem.
- Descriptions of the oft neglected, yet important methods of video data preparation and feature extraction.
- Explanations of common deep learning model building blocks with domain-specific examples.
- Groupings of state-of-the-art model architectures.
- Formal definitions of evaluation metrics across the span of video action problems.

This article is organized in the following way. Section II defines and organizes action understanding problems. Section III catalogs video action datasets by annotation type which directly relates to the problems for which they are applicable. Section IV provides an introduction to video data and data preparation techniques. Section V presents basic model building blocks and organizes state-of-the-art methods into families. Section VI 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 VII summarizes and concludes the tutorial. Note that because the breadth of topics covered is large, we chose to confine the scope of this tutorial to supervised learning as this is the dominant paradigm employed for video action understanding.

II. PROBLEMS

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

A. TAXONOMY

We organize six main action understanding problems into overlapping classification and search bins. Classification involves labeling videos by their action class. Search involves temporally or spatiotemporally finding action instances.

1) DEFINITIONS

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 needs to complete the action classification task while disregarding non-action background segments of the input.

Action Prediction (AP) is the process of classifying an incomplete input by the action yet to be observed. One subproblem is *action anticipation (AA)* in which no portion of the action is yet 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 is 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/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) 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., [34]–[36]). 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 incorrectly equates recognition and prediction. Second, localization and detection are often used interchangeably (e.g., [37]–[39]). However, in this case, because the task involves 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., [40]–[42]). We chose to use the former because proposal can be defined as the act of generating a proposal. Proposal generation is therefore redundant. An important takeaway is that the literature contains many examples where different terms refer to the same action problem (e.g., [29] and [43]). Similarly, there are many examples where the same terms refer to different problems (e.g., [44] and [45]). To compound the issue, many video action datasets can be applied to more than one of these problems. We encourage readers to pay careful attention to these terms when venturing into this field.

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 does not refer to SAP independently of SAL/D.

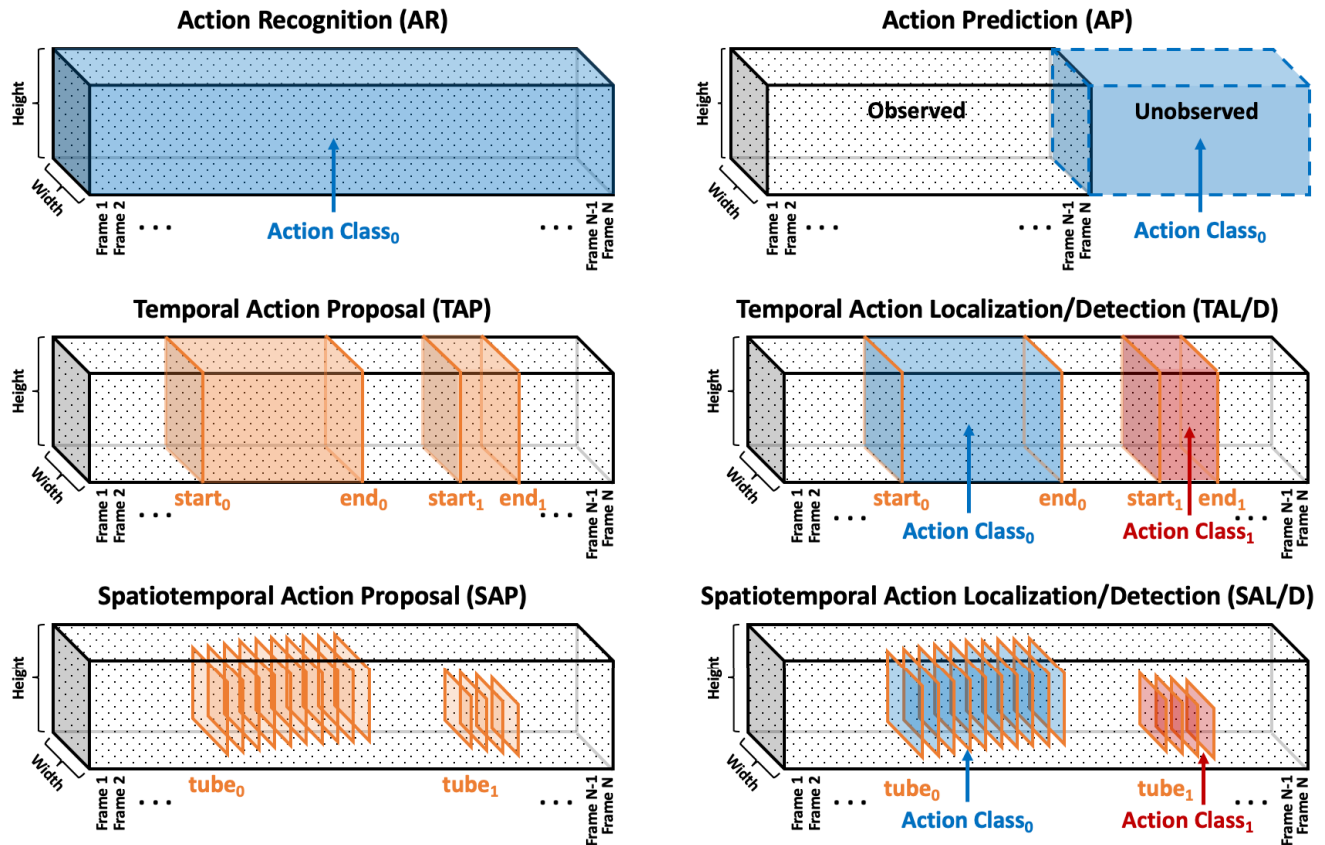


FIGURE 2. An overview of the main action understanding problems. Video is depicted as a 3D volume where N frames are densely stacked along a temporal dimension (left-to-right). Action recognition (upper left) shows how an action class label is assigned to the entirety of the video or video clip. Action prediction (upper right) shows how an action class label is assigned to a yet unobserved or only partial observed portion of a video. Temporal action proposal (middle left) shows how temporal regions of likely action are bounded by start and end frames. Temporal action localization/detection (middle right) shows how action class labels are assigned to temporal regions of likely action that are bounded by start and end frames. Spatiotemporal action proposal (bottom left) shows how “tubes” or “tubelets” are formed from bounding boxes across frames indicating spatiotemporal regions of likely action. Spatiotemporal action localization/detection (bottom right) shows how action class labels are assigned to spatiotemporal tubes of likely action.

B. RELATED PROBLEMS

Here, we define a few problems related to, but not included in, our main taxonomy and cite literature for further reading.

Action instance segmentation (AIS) is the labeling of individual instances or examples of an action within the same video 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 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.* [21], Saha *et al.* [46], Ji *et al.* [47], and Saha *et al.* [48].

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 [49]–[52]. Similarly, video captioning datasets (such as MSVD [53], MVAD [54], MPII-MD [55] and ActivityNet Captions [56]) will sometimes be included in video action understanding dataset lists. For more on video captioning, Li *et al.* [57] present a survey on methods, datasets, difficulties, and trends.

Action spotting (AS), proposed by Alwassel *et al.* [58], 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 finding only 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 [59], [60]. We recommend Yao *et al.* [61] for a recent and broad survey on video object segmentation and tracking.

III. DATASETS

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

A. VIDEO ACTION DATASET CATALOG

The last two decades saw huge growth in available video action datasets. To the best of our knowledge, we organized the most comprehensive collection of these datasets. We catalog over 130 video action datasets sorted by release year. Due to the scale of this catalog, Table 2 shows a selection of thirty datasets that are historically influential, current state-of-the-art, or emerging benchmarks. The extended catalog can be found in our online repository.¹

1) CRITERIA

To be included in our catalog, a dataset must meet each of the following criteria:

- 1) Released between 2004 and 2020.
- 2) Comprised of single-channel or multi-channel videos.
- 3) Includes full-video or video segment annotations.
- 4) Captures at least two action classes.
- 5) Utilizes 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.

2) TRENDS

Several trends emerge from this catalog. First, these datasets grew 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 occurred over several orders of magnitude. Larger datasets are essential for training deep learning models with often millions of parameters. Second, datasets useful only for classification (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 generally precedes localization/detection research.

B. FOUNDATIONAL AND EMERGING BENCHMARKS

We organize datasets in three groups, those with:

- 1) only action class annotations primarily for AR;
- 2) temporal annotations most useful for TAP, TAL/D, and sometimes AP; and
- 3) spatiotemporal annotations most useful for SAL/D.

Because many of the earlier influential video action datasets such as KTH and Weizmann are described at length in previous survey papers [8], [22], [27], we focus on the current largest and highest quality datasets.

1) ACTION RECOGNITION DATASETS

Table 3 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 [72] was produced as a benchmark for comparing convolutional neural networks (CNNs). Examples of the 487 sports action classes include “cycling,” “snowboarding,” and “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 one million videos are divided with a 70/20/10 training/validation/test split. On average, videos are ~5.5 minutes long, and about 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 [90] (a.k.a. 20BN-SOMETHING-SOMETHING) was produced 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). 108,499 videos are divided with an 80/10/10 training/validation/test split. Each single-instance lasts for 2–6 seconds. A second and larger version [79] was released in 2018. It also added object annotations, reducing label noise, and improving video resolution. These 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 and human-human interactions. Class examples from *Kinetics-400* [74] include “hugging,” “mowing lawn,” and “washing dishes.” Clips were collected from YouTube and annotated by AMT crowdworkers. The dataset consists of 306,245 videos, and within each class, 50 and 100 are reserved for validation and testing, respectively. Each single-instance video lasts for ~10 seconds. Additional videos and classes were added in 2018 [77] and 2019 [82]. 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 [91] was produced 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. It was improved in 2019 [84] with additional

¹<https://github.com/hutch-matt/vau-tutorial/blob/master/VideoActionDatasetCatalog-updated.pdf>

TABLE 2. Thirty 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 May 2021, 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.

Video Dataset	Year	Cited	Action		Actors		Annotations			Theme/Purpose
			Classes	Instances	H	N	C	T	S	
KTH [62]	2004	4,246	6	2,391	✓		✓			B/W, static background
Weizmann [63]	2005	2,068	10	90	✓		✓			human motions
Coffee & Cigarettes [64]	2007	526	2	246	✓		✓	✓	✓	movies and TV
Hollywood2 [65]	2009	1,488	12	3,669	✓		✓			movies
VIRAT [66]	2011	634	23	~10,000	✓		✓	✓	✓	surveillance, aerial-view
HMDB51 [67]	2011	2,428	51	~7,000	✓		✓			human motions
UCF101 [68]	2012	3,183	101	13,320	✓		✓			web videos, expand UCF50
THUMOS'13 [68]–[70]	2013	191	*101	13,320	✓		✓		✓	web videos, extend UCF101
J-HMDB-21 [71]	2013	567	51	928	✓		✓		✓	re-annotate HMDB51 subset
Sports-1M [72]	2014	5,667	487	1,000,000	✓		✓			multi-label, sports
ActivityNet200 (v2.3) [73]	2016	1,118	200	23,064	✓		✓	✓		untrimmed web videos
Kinetics-400 [74]	2017	1,380	400	306,245	✓		✓			diverse web videos
AVA [75]	2017	404	80	>392,416	✓		✓		✓	atomic visual actions
Moments in Time (MiT) [16]	2017	212	339	836,144	✓	✓	✓			intra-class variation, web videos
MultiTHUMOS [76]	2017	305	65	~16,000	✓		✓	✓		multi-label, extends THUMOS
Kinetics-600 [77]	2018	115	600	495,547	✓		✓			extends Kinetics-400
EGTEA Gaze+ [78]	2018	94	106	10,325	✓		✓	✓	✓	egocentric, kitchen
Something-Something-v2 [79]	2018	12	174	220,847	✓		✓			extends Something-Something
Charades-Ego [80]	2018	39	157	68,536	✓		✓	✓		egocentric, daily activities
Jester [81]	2019	37	27	148,092	✓		✓			crowd-sourced, gestures
Kinetics-700 [82]	2019	96	700	~650,000	✓		✓			extends Kinetics-600
Multi-MiT [17]	2019	10	313	~1,020,000	✓	✓	✓			multi-label, extends MiT
HACS Clips [83]	2019	64	200	~1,500,000	✓		✓			trimmed web videos
HACS Segments [83]	2019	64	200	~139,000	✓		✓	✓		extends and improves SLAC
NTU RGB-D 120 [84]	2019	168	120	114,480	✓		✓			extends NTU RGB-D 60
EPIC-KITCHENS-100 [85]	2020	8	97	~90,000	✓		✓	✓	✓	extends EPIC-KITCHENS-55
AVA-Kinetics [86]	2020	17	80	>238,000	✓		✓		✓	adds annotations, AVA+Kinetics
AViD [87]	2020	5	887	~450,000	✓	✓	✓			diverse peoples, anonymized faces
FineGym [88]	2020	25	10	4,883	✓		✓	✓		gymnastics w/ sub-actions
HAA500 [89]	2020	2	500	~10,000	✓		✓	✓	✓	course-grained atomic actions

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

classes, videos, and views. This serves as a state-of-the-art benchmark for human AR with non-RGB modalities.

Moments in Time (MiT) [16] was produced with a focus on broadening action understanding to include people, objects, animals, and natural phenomenon. Examples of the 339 human and non-human action classes include “running,” “opening,” and “picking.” Clips were collected from a variety of internet sources and crowd-annotated by AMT. Nearly one million single-instance 3-second videos are divided with a roughly 89/4/7 training/validation/test split. The dataset was improved to *Multi-Moments in Time (M-MiT)* [17] in 2019, increasing the number of videos, 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 [81] (a.k.a. 20BN-JESTER) was produced 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 and organized with a 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) [87] uniquely 1) avoids western bias by providing human actions (and some non-human actions) from a diverse set of people

and cultures; 2) anonymizes all faces to protect privacy; and 3) ensures that all videos are static with a creative commons license. Most of the 887 classes are drawn from Kinetics [82], Charades [92], and MiT [16] 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. Approximately 450,000 videos were organized 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.

2) TEMPORALLY ANNOTATED DATASETS

Table 4 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 [73], [94] family was produced for both action recognition and detection. Example human action classes include “Drinking coffee,” “Getting a tattoo,” and “Ironing clothes.” *ActivityNet 100 (v1.2)* is a 100-class dataset 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. It was expanded to *ActivityNet 200 (v1.3)* with 200-classes 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.

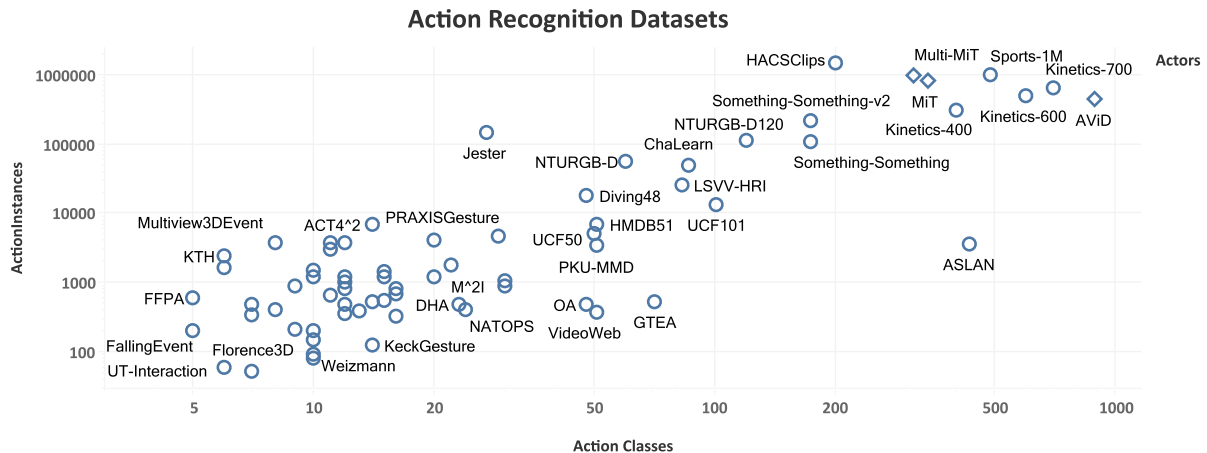


FIGURE 3. Datasets with only action class annotations mainly useful for action recognition (AR). Note that the plot is log-scaled in both dimensions. Datasets in the upper right (e.g., Kinetics-700, Sport-1M, AVID) have the greatest number of action classes as well as action instances. Many of the earlier influential datasets were much smaller and can be found in the lower left (e.g., KTH and Weizmann). Because space on the plot is limited, not all AR datasets are labeled. We recommend the catalog described in Section III-A for more details.

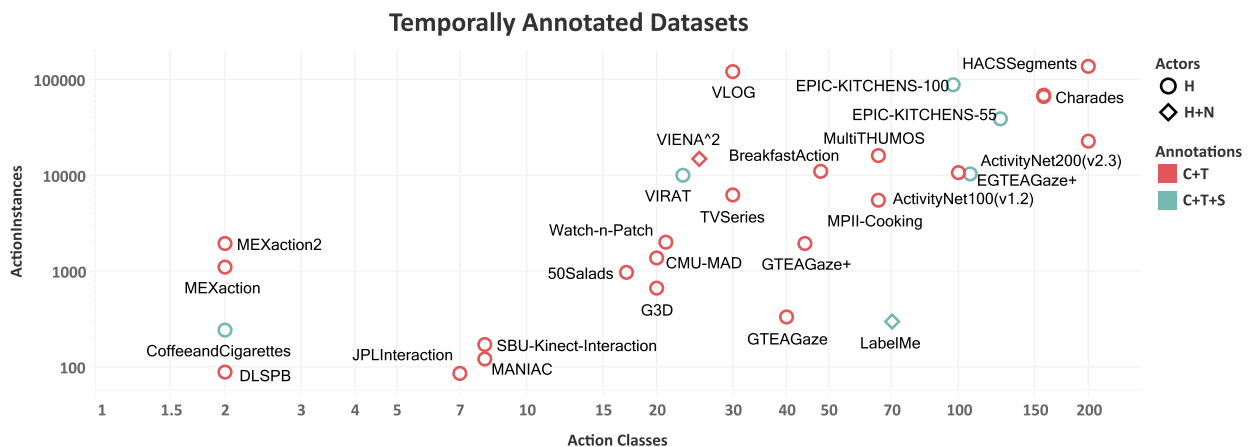


FIGURE 4. Datasets with temporal annotations useful for temporal action proposal (TAP), temporal action localization/detection (TAL/D), and possibly action prediction (AP). Note that the plot is log-scaled in both action instances and action classes dimensions. The largest datasets with the greatest number of action classes and action instances (e.g., HACSSegments, Charades, and EPIC-Kitchens-100) can be found in the upper right. Even these “large” temporally annotated datasets are an order of magnitude smaller in both classes and instances than the largest action recognition datasets. Also to note, the SLAC dataset [93] is excluded from this plot because while it has a very large number of temporally annotated action instances, the dataset was of poor quality. HACSSegments was developed out of SLAC and has significantly fewer temporally annotated action instances.

On average, action instances are 51.4 seconds long. Web videos were temporal annotated by AMT crowd-workers. ActivityNet remains 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 [92] was produced as a crowd-sourced dataset of daily human activities (e.g., “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* used 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* serves as a TAL/D benchmark along with ActivityNet, but it also useful 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 [76] was produced as an extension of the dataset used in the 2014 THUMOS Challenge [95]. Examples of the 65 human action classes include “throw,” “hug,” and “talkToCamera.” The 413 video (30 hours) dataset

has 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 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 Multi-THUMOS dataset offers a benchmark for multi-label TAP and TAL/D. It stands out due to its dense multi-labeling scheme.

VLOG [96] was produced as an implicitly gathered large-scale daily human actions dataset. Unlike previous daily human action datasets [90], [92], [97] 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 [83] was produced as a larger AR and TAL/D web-video dataset. Both *HACS Segments* and *HACS Clips* (the AR portion) are improvements on the *SLAC* dataset produced in the 2017 [93]. *HACS* uses the same 200 human action classes as ActivityNet 200 (1.3). Videos were collected from YouTube and temporally annotated by crowd-workers. In *HACS Segments* 50,000 videos are divided 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) SPATIOTEMPORALLY ANNOTATED DATASETS

Table 5 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 [66] was created 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 distance across 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 [51], [52].

UCF101-24, the spatiotemporally labelled data subset of *THUMOS’13* [69], was produced 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 [68]. Each video contains one or more spatiotemporally annotated action instances. While multiple instances within a video have separate spatial and temporal boundaries, they 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 [71] was produced 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 [67] 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 segmentations were created by 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 an early foundational benchmark for SAL/D.

EPIC-KITCHENS-55 [98] was produced 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 432 videos (55 hours) are 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 correspond to 28,561, 8,064, and 2,939 action instances, respectively. The dataset was improved to *EPIC-KITCHENS-100* [85] increasing the number of videos, action instances, participants, and environments. Annotation quality was also improved. This dataset serves as a state-of-the-art egocentric kitchen activities benchmark.

Atomic Visual Actions (AVA) [75] was produced 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

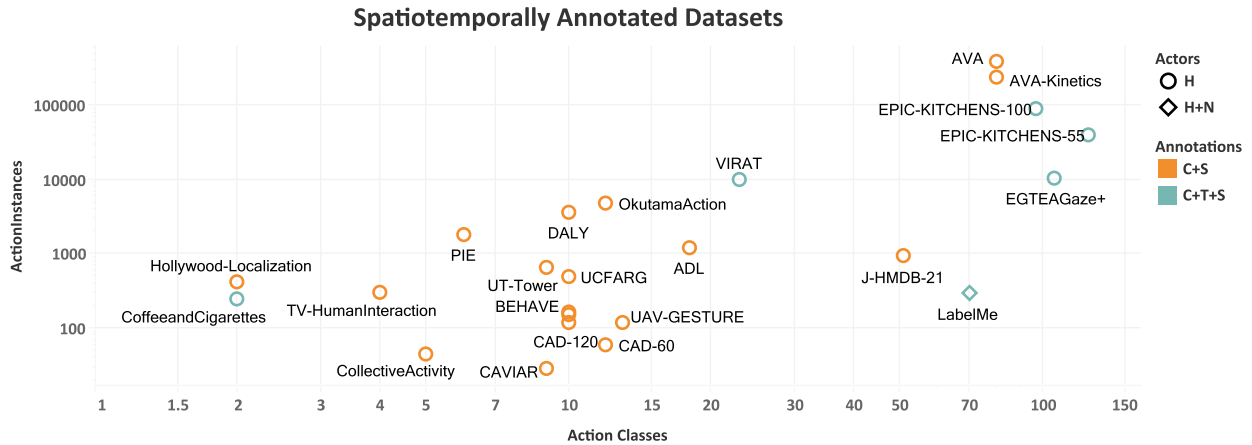


FIGURE 5. Datasets with spatiotemporal annotations useful for spatiotemporal action proposal (SAP) and spatiotemporal action localization/detection (SAL/D). Those with class labels and only spatial annotations have actions which span the entirety of the video or video clip. Note the plot is log-scaled in both instances and classes dimensions. The largest of these datasets can be found in the upper right (e.g., AVA, AVA-Kinetics, and EPIC-KITCHENS-100). Even these “large” spatiotemporally annotated datasets are an order of magnitude smaller than the largest action recognition datasets.

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 as a large-scale multi-label benchmark for TAL/D.

The *AVA-Kinetics* dataset [86] was produced by using an existing large-scale human action recognition dataset and a spatiotemporal atomic action annotation schema. The dataset combines a subset of videos from Kinetics-700 [82] and all videos from AVA [75] for a total of 238,906 videos with a roughly 59/14/27 training/validation/test split. For each 10-second video from Kinetics, 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.

C. COMPETITIONS

Several competitions introduced state-of-the-art datasets, galvanized model development, and standardized metrics. THUMOS Challenges [69], [95], [102] were 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 [49]–[52], [73] were held at CVPR from 2016 through 2020 and have slowly expanded in scope encompassing trimmed AR, untrimmed AR, TAP, TAL/D, and SAL/D competitions. Other challenges were modeled off THUMOS and ActivityNet such as the Workshop on Multi-modal Video Analysis and Moments

in Time Challenge² held at ICCV in 2019. We provide an overview of these highly visible competitions in Table 3. We anticipate these competitions will continue to grow in popularity in future years. Note that the metrics specified in the table will be defined and described in Section VI. These competitions are a useful place for finding the current state-of-the-art models and methods. Check the “Top Results” column of Table 3 for papers describing the best-performing models.

IV. DATA PREPARATION

While some datasets are available in pre-processed forms, others are presented *raw*—using the original frame rate, frame dimensions, duration, and/or formatting. *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 detecting and removing incomplete or irrelevant portions of the dataset. For datasets that simply link to YouTube or other web videos (e.g., [74], [77], [82], [83]), 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

²<https://sites.google.com/view/multimodalvideo/home>

TABLE 3. Prominent video action understanding challenges 2013-2021.

Workshop	Year	Conf.	Problem	Dataset	Metric(s)	Top Result	#Teams
THUMOS [69]	2013	ICCV	AR	UCF101	average accuracy	85.90 [99]	17
			SAL/D	UCF101-24	ROC AUC sIoU@0.2	n/a	n/a
THUMOS [95]	2014	ECCV	AR	UCF101+	mAP	0.71 [100]	11
			TAL/D	UCF101-20	mAP tIoU@{0.1,0.2,0.3,0.4,0.5}	.4,.3,.3,.2,.1 [101]	3
THUMOS [70], [102]	2015	CVPR	AR	UCF101+1	mAP	0.74 [103]	11
			TAL/D	UCF101-20	mAP tIoU@{0.1,0.2,0.3,0.4,0.5}	.4,.4,.3,.2,.2 [104]	1
ActivityNet [73]	2016	CVPR	AR	ActivityNet 1.3	mAP, Top-1 accuracy, Top-3 accuracy	93.2, 88.1 [105]	26
			TAL/D	ActivityNet 1.3	mAP-50, mAP-75, average-mAP	42.5	6
ActivityNet [49]	2017	CVPR	AR	ActivityNet 1.3	Top-1 error	8.8 [49]	n/a
			AR	Kinetics-400	average(Top-1 error, Top-5 error)	12.4 [106]	31
			TAP	ActivityNet 1.3	AR-AN AUC	64.80 [107]	17
			TAL/D	ActivityNet 1.3	mAP tIoU@0.5:0.05:0.95	33.40 [107]	17
ActivityNet [50]	2018	CVPR	AR	Kinetics-600	average(Top-1 error, Top-5 error)	10.99 [108]	13
			AR	MiT (full-track)	average(Top-1 acc, Top-5 acc)	52.91 [109]	29
			AR	MiT (mini-track)	average(Top-1 acc, Top-5 acc)	47.72 [110]	12
			TAP	ActivityNet 1.3	AR-AN AUC	71.0 [50]	55
			TAL/D	ActivityNet 1.3	mAP tIoU@0.5:0.05:0.95	38.53 [40]	43
			SAL/D	AVA	frame-mAP sIoU@0.5	20.99 [111]	23
ActivityNet [51], [112]	2019	CVPR	AR	Kinetics-700	average(Top-1 error, Top-5 error)	17.88 [113]	15
			AR	EPIC-KITCHENS-55	micro-avg Top-1,5 acc, macro-AP,AR	41.4, 25.1 [114]	25
			AP	EPIC-KITCHENS-55	micro-avg Top-1,5 acc, macro-AP,AR	13.2, 8.5 [115]	8
			TAP	ActivityNet 1.3	AR-AN AUC	72.99	61
			TAL/D	ActivityNet 1.3	mAP tIoU@0.5:0.05:0.95	39.7	23
			TAL/D	VIRAT	$P_{rate}@miss_{FA}$	0.605	42
			SAL/D	AVA	frame mAP sIoU@0.5	34.25 [116]	32
Multi-modal [117]	2019	ICCV	AR	Multi-MiT	mAP	0.608 [118]	10
			TAL/D	HACS Segments	mAP tIoU@0.5:0.05:0.95	23.49 [119]	5
ActivityNet [52]	2020	CVPR	AR	Kinetics-700	average(Top-1 error, Top-5 error)	14.9	4
			TAL/D	ActivityNet 1.3	mAP tIoU@0.5:0.05:0.95	42.79 [120]	55
			TAL/D	MEVA [121]	avg($1-P_{miss}$) across TFA from 0-20%	0.350	11
			TAL/D	HACS Segments	mAP tIoU@0.5:0.05:0.95	40.53 [122]	22
			TAL/D	HACS Clips+Seg.	mAP tIoU@0.5:0.05:0.95	39.29 [123]	13
			SAL/D	AVA-Kinetics	frame mAP sIoU@0.5	39.62 [124]	11
ActivityNet	2021	CVPR	AR	Kinetics-700-2020 [125]	average(Top-1 error, Top-5 error)	14.0	27
			AR	TinyVIRAT [126]	F1-score	0.478 [127]	3+
			TAL/D	ActivityNet 1.3	mAP tIoU@0.5:0.05:0.95	44.67 [128]	3+
			TAL/D	HACS Segments	mAP tIoU@0.5:0.05:0.95	44.29 [129]	22
			TAL/D	HACS Clips+Seg.	mAP tIoU@0.5:0.05:0.95	22.45 [129]	13
			TAL/D	SoccerNet-V2 [130]	average mAP per class	74.84 [131]	7
			TAL/D	MEVA	avg($1-P_{miss}$) across TFA from 0-20%	0.425	9
			SAL/D	AVA-Kinetics	frame mAP sIoU@0.5	40.67 [132]	11

independently learn. With video data, motion representations are the most common extracted features.

While data cleaning is certainly important, this section focuses primarily on augmentation and hand-crafted feature extraction because of the many video domain-specific terms.

A. VIDEO DATA

Video is composed of a series of still-image frames where each frame is made of rows and columns of *pixels*. Pixels are 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.³

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., Theano⁴ and MXNet⁵ versus CNTK⁶ and TensorFlow⁷).

B. DATA AUGMENTATION

1) GEOMETRIC AUGMENTATION METHODS

In the context of video, *geometric augmentation methods* are transformations that alter the geometry of frames [133]. To be effective, these must be applied equally across all frames. If separate geometric transformations are applied on different

³<https://developer.microsoft.com/en-us/windows/kinect/>

⁴<https://deeplearning.net/software/theano/>

⁵<https://mxnet.apache.org/versions/1.6/>

⁶<https://docs.microsoft.com/en-us/cognitive-toolkit/>

⁷<https://www.tensorflow.org/>

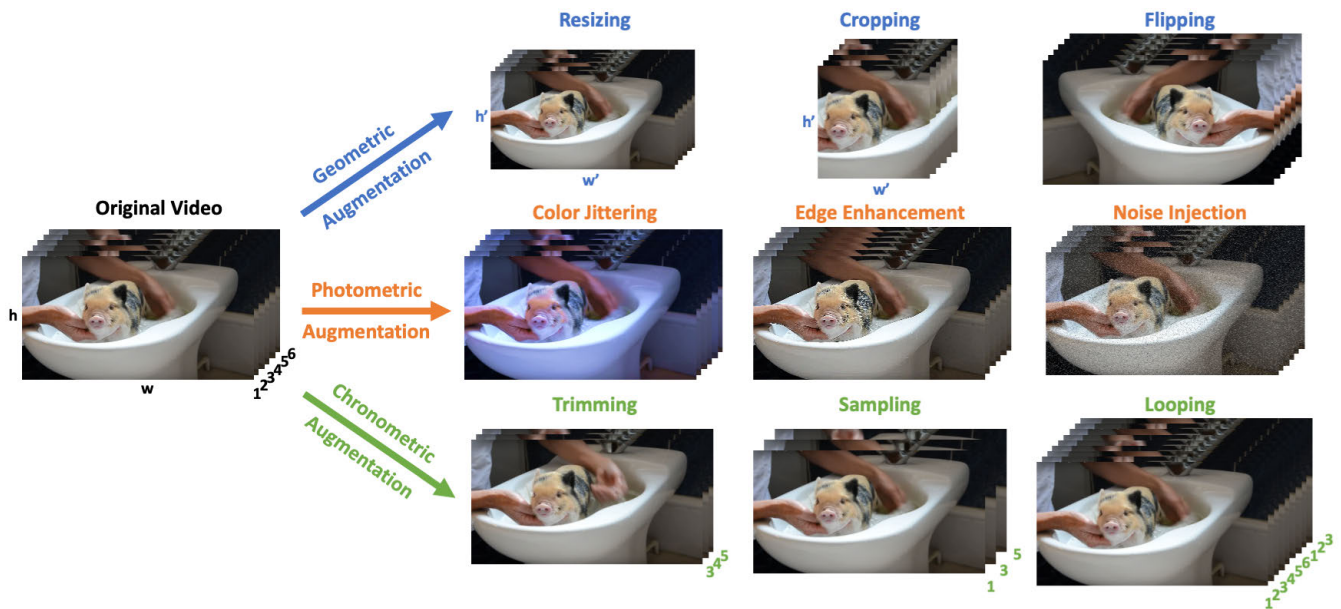


FIGURE 6. Common video augmentations. The top row shows three geometric augmentations that affect geometry of the frames. These include resizing which changes the height and width while either preserving or permuting the aspect ratio, cropping which removes pixels, and horizontal flipping which mirrors across the vertical center axis. The middle row shows three photometric augmentations that affect the color-space of the frames. These include color jittering which adjusts pixel color in an orchestrated way, edge enhancement which increases the definition of contours, and noise injection which adjusts pixel color in a random way. The bottom row shows three chronometric augmentations that affect the temporal aspect of the video. These include trimming which removes consecutive sections of frames, sampling which removes periodic frames, and looping which duplicates frame sequences and adds that to the beginning or end of the original frame sequence. Frames are taken from the Moments in Time dataset [16] “washing” class.

frames, a video could quickly lose 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 [134]. Ratio jittering [135] 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 [136]–[138] and corner cropping [139].
- *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 [136], [139]–[141].

Other geometric augmentation methods that are less popular for video include *vertical flipping*, *shearing*, *piecewise affine transforming*, and *rotating*. Shorten and Khoshgoftaar [138] present a survey on image augmentation which describes some of these alternative techniques that could easily be applied to video. 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.

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 [133]. 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 [140], [142], [143] or via a specific adjustments [136], [144].
- *Edge enhancement*—the process of increasing the appearance of contours in a video’s frames. In some settings, this speeds up the learning process since it is shown that the first few layers in convolutional neural networks learn to detect edges and gradients [136].

Other photometric augmentation methods that may be useful in future settings are *superpixelization*, *random gray* [143], *random erasing* [138], and *vignetting* [143]. However, these are not only absent from the action understanding literature but also uncommon in image understanding.

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 augmentations* 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 [141], [145] or randomly selected frame indices [139], [140].
- *Looping*—the process of repeating a video's frames to increase the duration—essentially temporal padding [141]. This might be necessary when a video segment has fewer frames than the model's input specifies.

C. HAND-CRAFTED FEATURE EXTRACTION

While shallow learning is less common since the deep learning revolution, several hand-crafted motion features have found their way into state-of-the-art deep learning models [139]–[141], [145]. These motion representations generally fall under two classical field theories: Lagrangian flow [149] and Eulerian flow [150]. Motion representations are one way of capturing temporal information for a video. This can be important because temporal or casual learning remains a significant challenge in the deep learning field.

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)* [151]. Many methods exist for computing this feature: the Lucas–Kanade method [152], the Horn–Schunck method [153], the TV-L1 approach [154], the Farneback method [146], and others [155]. It is also possible to employ a WarpFlow technique [156] to attempt to reduce background or camera motion. This technique requires computing the *homography*, a transformation between two planes, between frames. OF is noted for its usefulness in action understanding because it is invariant to appearance [157].

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)* [135], [139], [148] and *phase difference/derivative (dPhase)* [148]. RGB derivative is the difference between pixel color intensities at equivalent spatial locations in adjacent frames. To compute this, one frame is subtracted from another. Phase difference requires converting each frame into the frequency domain before taking the difference and converting back to the time domain.

V. MODELS

The past decade of action understanding research saw 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 a variety of models into groups based on similar underlying architectures. For brevity, we do not cover loss functions or methods of supervising model training. We recommend exploring these methods by examining the papers referenced throughout this section.

A. MODEL BUILDING BLOCKS

1) CONVOLUTIONAL NEURAL NETWORKS

No deep learning architecture component has impacted action understanding (and computer vision at large) greater 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., [158], [159]). 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 [160]. However, deep models can suffer from problems like exploding or vanishing gradients [161].

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. 3D-convolution allows for a temporal receptive field in addition to the standard spatial one. Single-channel examples of 2D and 3D convolutions are shown in Fig. 8. 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. We briefly note a few influential developments not unique to, but consistently employed throughout, the action understanding literature:

- *Residual networks (ResNets)* [162]—utilize skip connections to avoid vanishing gradients.
- *Inception blocks* [163], [164]—utilize multi-size filters for computational efficiency.
- *Dense connections (DenseNet)* [165]—utilize skip connections between each layer and every subsequent layer for strengthening feature propagation.

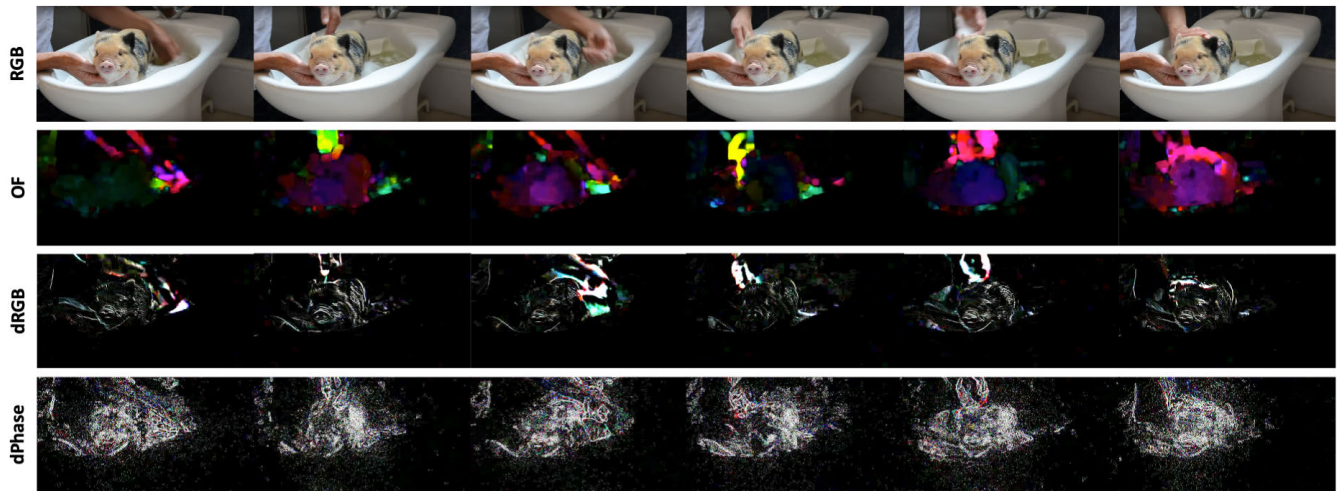


FIGURE 7. Examples of hand-crafted video features. The first row shows the original RGB sampled video frames. The second row shows dense optical flow (OF) computed using the Farneback method [146] and OpenCV packages [147] (color indicates direction). The third row shows RGB difference/derivative (dRGB) between consecutive frames. The fourth row shows phase difference/derivative (dPhase) between consecutive frames computed using the approach described in [148]. The video frames are from the Moments in Time dataset [16] “washing” class.

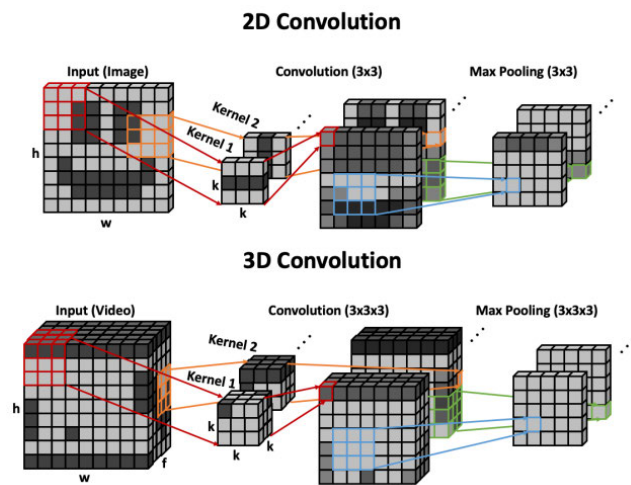


FIGURE 8. Examples of 2D and 3D convolutional layers and max pooling on single-channel image and video inputs. In the 2D (image) example, each small “cube” in the input represents a 3-channel pixel. In the 3D (video) example, frames (rows and columns of pixels) are stacked from front to back. Note that these filter kernels were chosen randomly and do not necessarily lead to good embedded features.

- *Inflated networks* [141]—expand lower dimensional networks into a higher dimension in a way that benefits from lower dimensional pretrained weights (e.g., I3D).
- *Normalization* [166]—methods of suppressing the undesired effects of random initialization and random internal distribution shifts. These include *batch normalization (BN)* [166], *layer normalization (LN)* [167], *instance normalization (IN)* [168], and *group normalization (GN)* [169].

Recently, many *hybrid CNNs* introduced new convolutional blocks, layers, and modules. Some focus on reducing the large computational costs of C3D: P3D [170],

R(2+1)D [171], [172], ARTNet [173], MFNet [174], GST [175], and CSN [176]. Others focus on recognizing long-range temporal dependencies: LTC-CNN [177], NL [173], Timeception [178], and STDA [179]. Some unique modules include TSM [180] which shifts individual channels along the temporal dimension for improved C2D performance, TrajectoryNet [181] which introduces a TDD-like [182] trajectory convolution to replace temporal convolutions, and GSM [183] which introduces a gate-shift module.

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)* [184]. 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)* [185]. 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 usage and underlying principles (e.g., [186]–[189]).

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 [72], [190], [191]. The simplest and most naïve form is averaging. However, recently *attention mechanisms*, processes that allow a model to focus on the most relevant

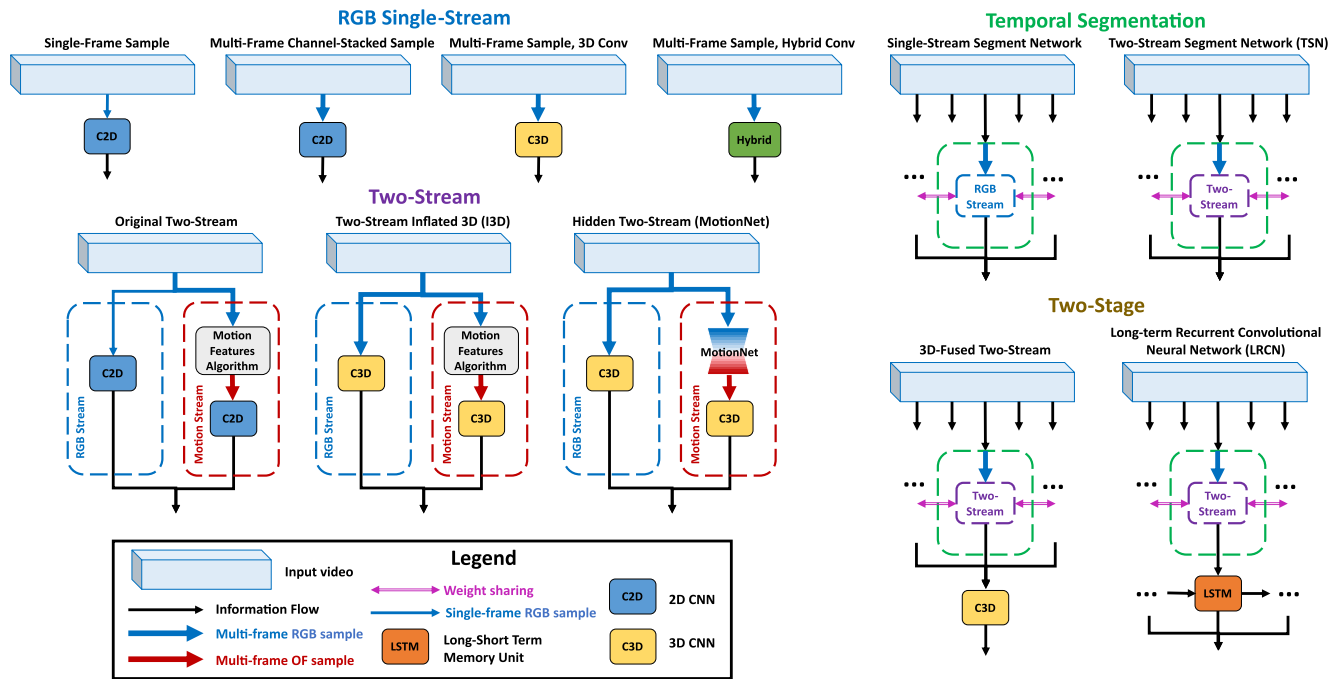


FIGURE 9. Action recognition model examples. RGB and Motion Single-Stream architectures train a 2D, 3D, or Hybrid CNN on one sampled feature. Two-stream architectures fuse RGB and Motion streams. Temporal Segmentation architectures divide a video into segments, process each segment on a single-stream 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. Please note that this is a limited selection of many action recognition model architectures. For example, models that use recent vision transformers for video are not included above.

information and disregard the least relevant information, have gained popularity.

B. MODEL ARCHITECTURE FAMILIES

We focus here on grouping these methods into architecture families under each action problem and pointing to useful examples. These lists are neither exhaustive nor intended to critique the field. As this article is cast as a tutorial, we hope the reader gains a sense of the many varying directions of ongoing study. Because of the rapidly evolving nature of the field, we recommend checking online scoreboards⁸ for up-to-date performances on benchmark datasets.

1) ACTION RECOGNITION MODELS

As shown in Fig. 9, we broadly group AR architectures into families of varying complexity. The first is *single-stream architectures*, which sample or extract one 2D [72], [192], [193] or 3D [194]–[197] 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 [16], [72], 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 V-A1.

The second family is *two-stream architectures* with one stream for RGB learning and one stream for motion feature learning [140], [141]. 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. These include MotionNet [198] which operates similarly to standard two-stream methods, and MARS [199] and D3D [200] which perform middle fusion between the streams. Feichtenhofer *et al.* [201] explores gating techniques between the streams. While these models are generally computationally constrained to two streams, more streams for additional modalities are possible [202], [203].

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 [135], [139] 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 [204], TRN [205], ECO [206], and Slow-Fast [116] build on temporal segmentation by performing multi-resolution segmentation and/or fusion. Temporal Pyramid Networks (TPN) [207] also perform segmentation across multiple temporal levels with improved performance and robustness over standard TSN and TRN implementations.

The fourth family, a higher level of complexity, 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 [145] and CNN+LSTM approaches [208]–[212].

⁸<https://paperswithcode.com/area/computer-vision>

Ma *et al.* conducted a side-by-side comparison of C3D and CNN+LSTM performance [213]. Temporal Excitation and Aggregation (TEA) [214] also uses a two-stage-like approach extracting frame-level features with a C2D prior to motion excitation and multiple temporal aggregation.

Arguably, a fifth family of action recognition models that employs *vision transformers* coalesced in 2020 and 2021. Transformers, which made their debut in the natural language processing (NLP) field in 2017 [215], are an encoder-decoder sequence-to-sequence modeling schema that uses self-attention rather than recurrent neural networks or convolution. Examples include TimeSformer [216] which uses embeddings of frame patches augmented with positional information as a sequence of tokens for the transformer, VTN [217] which is based off a transformer model that processes long sequences of tokens, ViViT [218] which uses another pure-transformer architecture, and MViT [219] which introduce resolution and channel scaling in combination with the vision transformer.

2) ACTION PREDICTION MODELS

Rasouli [32] 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 [220] which uses a reinforcement learning module to improve an encoder-decoder, IRL [221] which uses a C2D inverse-reinforcement learning strategy to predict future frames, Conv3D [222] which uses a C3D to generate unseen features for prediction, RGN [223] which uses a recursive generation and prediction scheme with a Kalman filter during training, and RU-LSTM [115], [224] 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 [225] which uses an exponential loss to bias a multi-modal CNN+LSTM fusion strategy towards the most recent predictions, MS-LSTM [226] which uses two LSTM stages for action-aware and context-aware learning, MM-LSTM [227] which extends MS-LSTM to arbitrarily many modalities, FN [228] which uses a three-stage LSTM approach, and TP-LSTM [229] 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 was observed). Additionally, action recognition models described in Section V-B1 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.

3) TEMPORAL ACTION PROPOSAL MODELS

As shown in Fig. 11, TAP approaches can be grouped into three families. The first family is *top-down architectures* which use sliding windows to derive segment-level proposals. Examples include DAP [230] and SST [231] which use CNN feature extractors and recurrent networks, S-CNN [37] which uses multi-scale sliding windows, and TURN TAP [232] 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 [38] which uses a flooding algorithm to convert these predictions into multi-scale groupings, BSN [40] and BMN [233] which use additional “startness” and “endness” features for different proposal generation and proposal evaluation techniques, and RecapNet [234] which uses a residual causal network rather than a generic 1D CNN to compute confidence predictions. R-C3D [45] and TAL-Net [39] use region-based methods to adapt 2D object proposals in images to 1D action proposals in videos. 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 architectures generally create segment proposals and actionness scores in parallel and then use actionness to refine the proposals. Examples include CDC [235], CTAP [41], MGG [42], and DPP [236].

4) TEMPORAL ACTION LOCALIZATION/DETECTION MODELS

As shown in Fig. 12 and introduced in Xia and Zhan [31], there are two main families of TAL/D methods. The first family is *two-stage architectures* in which the first stage creates proposals and the second stage classifies them. Therefore, you can pair any TAP model described in Section V-B3 with an AR model described in Section V-B1. 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. This family of architectures spans a wide variety of implementations. Examples include SSAD [237] which creates a snippet-level action score sequence from which a 1D CNN extracts multi-scale detections, SS-TAD [238] in which parallel recurrent memory cells create proposals and classifications, Decouple-SSAD [239] which builds on SSAD with a three-stream decoupled-anchor network, GTAN [240] which uses multi-scale Gaussian kernels, Two-stream SSD [241] which fuses RGB detections with OF detections, and RBC [242] which completes boundary refinement prior to classification. Recently, models that use graph convolutional networks have shown promise. Examples include those proposed by Zeng *et al.* [44] and Huang *et al.* [243].

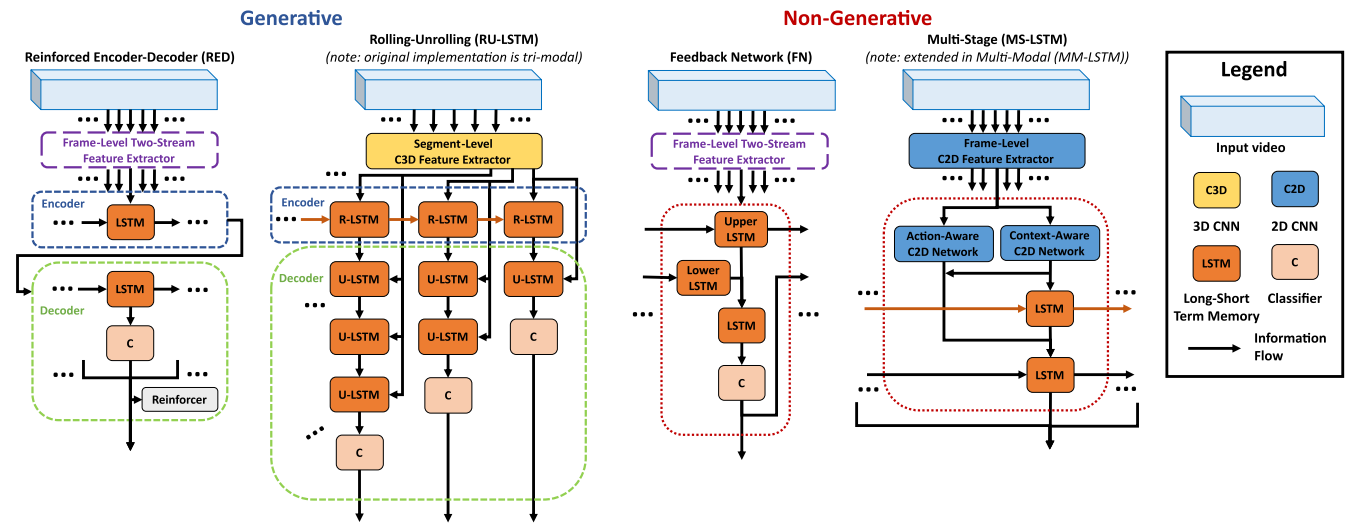


FIGURE 10. Action prediction model examples. Generative models create representations of future timesteps for prediction (typically via an encoder-decoder scheme). Non-generative models are a broad-sweeping category for those which create predictions directly from observed sections of the input.

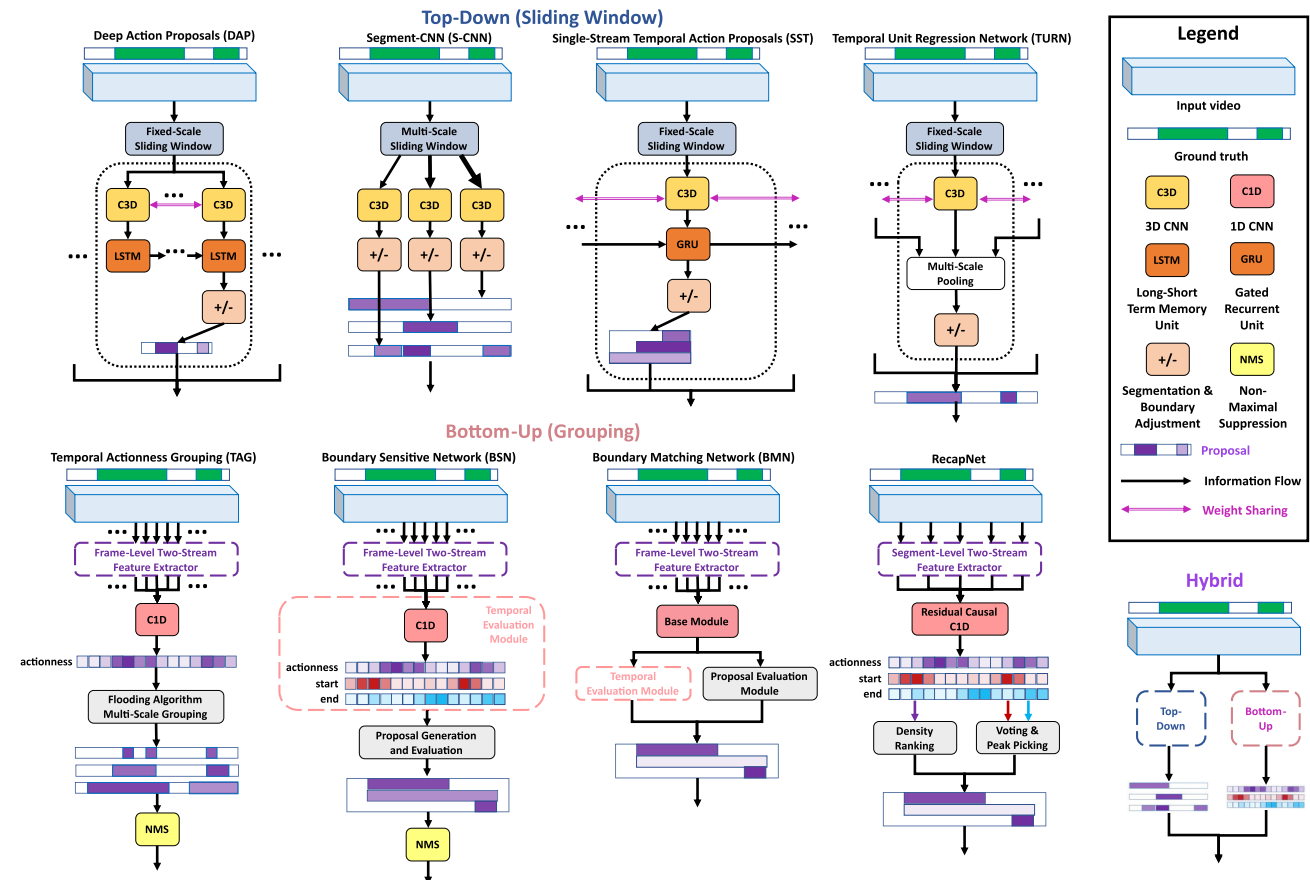


FIGURE 11. Temporal action proposal model examples. Top-down models use a sliding window approach to create segment-level proposals. Bottom-up models use frame or short-segment level actionness score predictions with grouping strategies to produce proposals. Hybrid models use both top-down and bottom-up strategies in parallel.

5) SPATIOTEMPORAL ACTION LOCALIZATION/DETECTION MODELS

As shown in Fig. 13, 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 [244], Fast R-CNN [245], Faster R-CNN [246], early+late fusion Faster R-CNN [247]) to derive bounding boxes from frame then apply a

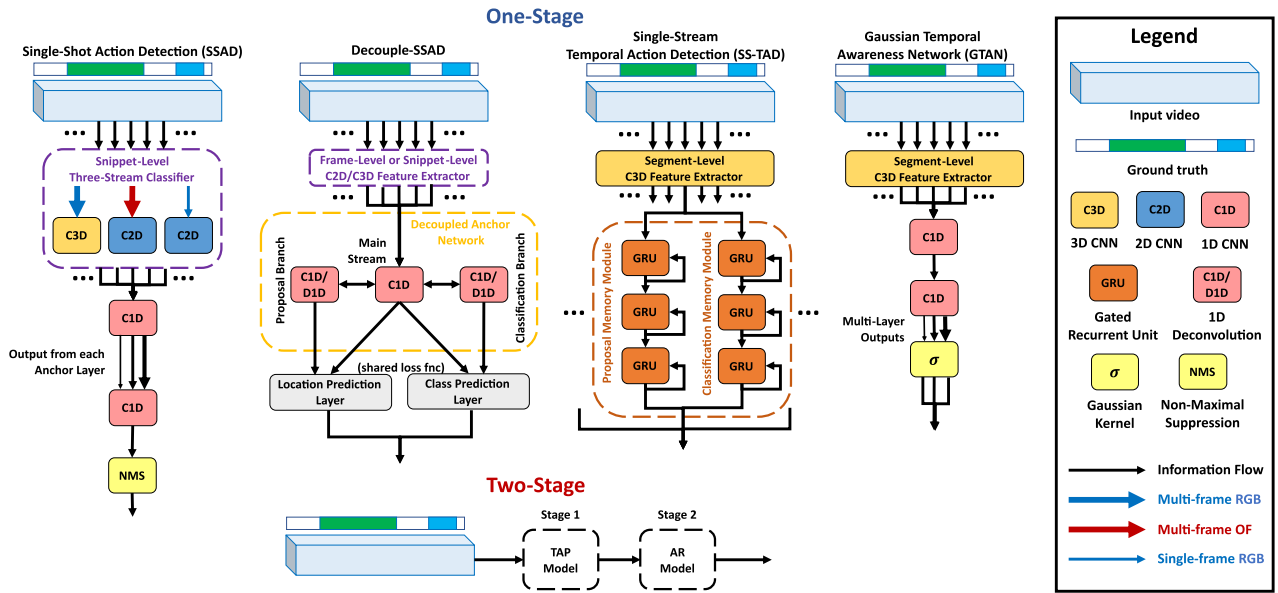


FIGURE 12. Temporal action localization/detection model examples. One-stage architectures conduct proposal and classification together while two-stage architectures create proposals and then use an action recognition model to classify each proposal.

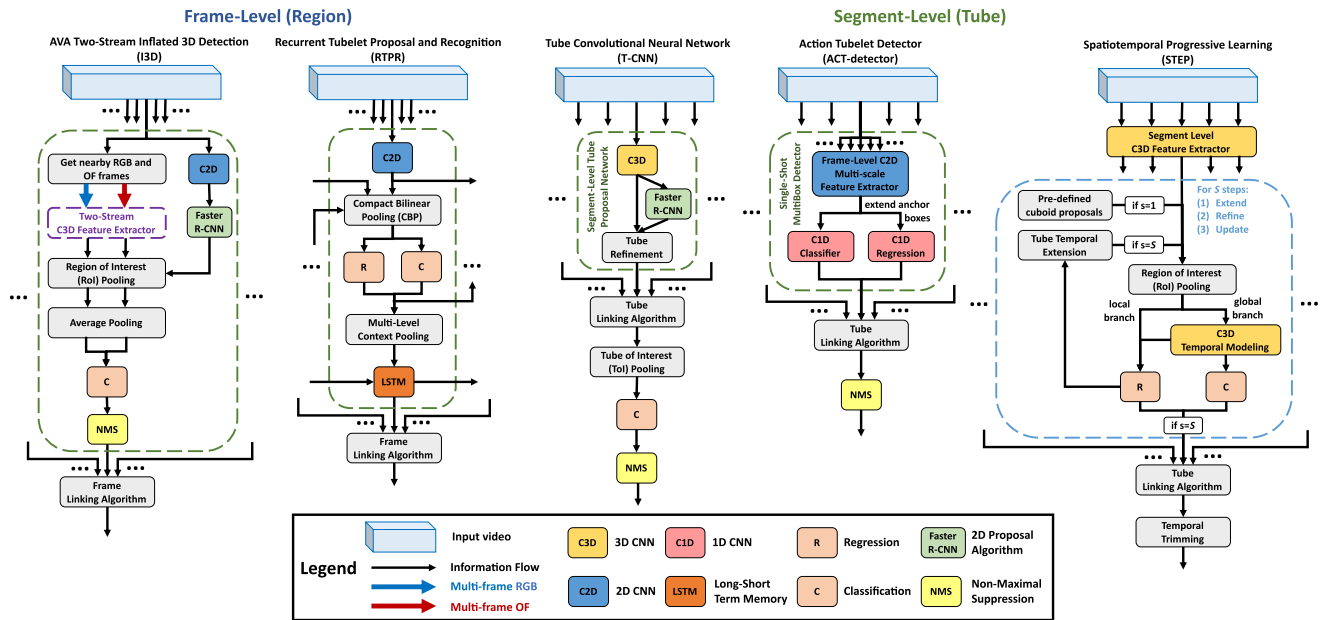


FIGURE 13. Spatiotemporal action localization/detection model examples. Frame-level (region) proposal models link frame-level detections together while segment-level (tube) proposal models create small “tubelets” for short segments and link the tubelets into longer tubes.

frame linking algorithm. Examples include MR-TS [248], CPLA [249], ROAD [250], AVA I3D [75], RTPR [251], and PntMatch [247].

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 [252], ACT-detector [253], and STEP [254].

A few state-of-the-art models do not fit nicely in either of these families but are worth noting. Zhang *et al.* [255] use a tracking network and graph convolutional network

to derive person-object detections. VATX [256] augments I3D approaches with a multi-head, multi-layer transformer. STAGE [257] introduces a temporal graph attention method.

VI. 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 do not cover binary classification metrics as the action datasets we 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:

- $X = \{x^{(1)}, \dots, x^{(n)}\}$ as the set of n input videos,
- $Y = \{y^{(1)}, \dots, y^{(n)}\}$ as the set of n ground truth annotations for the input videos,
- $M : X \rightarrow \hat{Y}$ as a function (a.k.a. model) mapping input videos to prediction annotations,
- $\hat{Y} = \{\hat{y}^{(1)}, \dots, \hat{y}^{(n)}\}$ as the set of n model outputs,
- $C = \{1, \dots, m\}$ as the set of m action classes, and
- $\text{TP}_j : \mathbb{N} \rightarrow \{0, 1\}$ as 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. Fig. 14 depicts spatial IoU (sIoU), temporal IoU (tIoU), and spatiotemporal IoU (stIoU).

A. 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 one. Formally, $\forall i \in \{1, \dots, n\}$:

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

We define the two common metrics below. Other metrics that we do not cover include F_1 score (micro-averaged and macro-averaged), Cohen's Kappa, PR-AUC, ROC-AUC, partial AUC (pAUC), or two-way pAUC. Sokolava and Lapalme [258] and Tharwat [259] present thorough evaluations of these and other multi-class classification metrics.

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 [49]–[52], [73]. 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 (a_k) over the entire input set, where $\mathbb{1}$ is a 0-1 indicator function, is defined as

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

2) MULTI-CLASS MEAN AVERAGE PRECISION (mAP)

This metric is the arithmetic mean of the interpolated average precision of each class, and it was used in multiple THUMOS and ActivityNet challenges [49], [69], [73], [95], [102]. To calculate interpolated average precision for a particular class j , the model outputs must be

ranked in decreasing confidence of that class. Formally, $\forall j \in \{1, \dots, m\}$, L_j is a ranked list of outputs such that $\forall a, b \in \{1, \dots, n\}$, $p_j^{(a)} \geq p_j^{(b)}$. 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., $\text{TP}_j(r) = 1$ if $y^{(i)} = j$). Using these lists $L_1, \dots, L_m$, precision (P) up to rank k in a given list L_j is defined as

$$P_j(k) = \frac{1}{k} \sum_{r=1}^k \text{TP}_j(r). \quad (2)$$

Interpolated average precision (AP) over all ranks with unique recall values for a given class j is calculated as

$$\text{AP}(j) = \frac{\sum_{k=1}^n P_j(k) * \text{TP}_j(k)}{\sum_{k=1}^n \text{TP}_j(k)}. \quad (3)$$

Therefore, mean average precision (mAP) is calculated as

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

B. MULTI-LABEL CLASSIFICATION METRICS

In the context of action understanding, multi-label classification consists of action recognition or prediction 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 one. Formally, $\forall i \in \{1, \dots, n\}$:

- $y^{(i)} \subseteq C$ because the ground truth annotation for input $x^{(i)}$ is a subset of action classes and
- $\hat{y}^{(i)} = \{p_1^{(i)}, \dots, p_m^{(i)}\}$ 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 [260] and Wu and Zhou [261] which present surveys of multi-label classification metrics.

1) MULTI-LABEL MEAN AVERAGE PRECISION (mAP)

This is the same metric as described in Section VI-A2, 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., $\text{TP}_j(r) = 1$ 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 (2), (3), and (4). This metric is used for MultiTHUMOS [76], ActivityNet 1.3 [73] when applied as an untrimmed action recognition problem, and Multi-Moments in Time [17]. 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., $t = 0.5$).

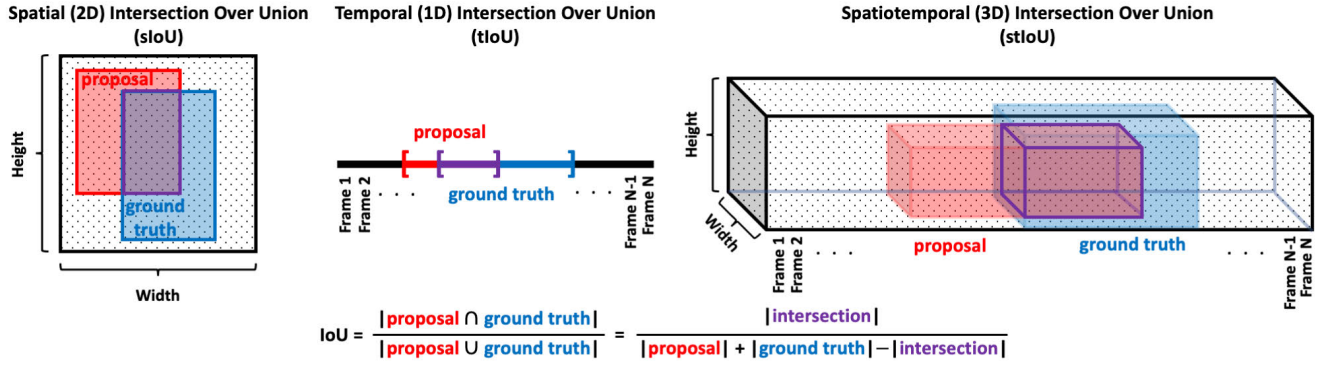


FIGURE 14. Illustration of types of intersection over union: spatial (left), temporal (center), and spatiotemporal (right). Below the diagrams is an equation for intersection over union (IoU). IoU is also known as the *Jaccard index* or *similarity coefficient*.

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 [72]. 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. Formally,

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

C. TEMPORAL PROPOSAL METRICS

Metrics for temporal action proposal are less varied than those for classification. Below, we define the two main metrics 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, $\forall i \in \{1, \dots, n\}$:

- $y^{(i)} = \{s_1^{(i)}, \dots\}$ is the ground truth annotation set of temporal segments where $s_j^{(i)}$ consists of start and end markers for input video $x^{(i)}$,
- $\hat{y}^{(i)} = \{(\hat{s}_1^{(i)}, c_1^{(i)}), \dots\}$ 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 $x^{(i)}$, and
- $\text{tIoU}(s_a^{(i)}, \hat{s}_b^{(i)})$ is the temporal intersection over union between the ground truth and a proposal.

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

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

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., $\text{TP}_a^{(i)}(t) = 1$ if $\text{tIoU}(s_a^{(i)}, \hat{s}_b^{(i)}) \geq t$). Recall is the proportion of all ground truth temporal segments for which there is a true positive prediction. Formally, recall (R) at a particular threshold t and average number of proposals (AN) is calculated as

$$\text{R}(t)@AN = \frac{1}{\sum_{i=1}^n |y^{(i)}|} \sum_{i=1}^n \sum_{j=\{1, \dots\}} \text{TP}_j^{(i)}(t). \quad (7)$$

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 $\eta = 0.05$ [49]–[51]. Formally, average recall (AR) over thresholds from 0.5 to $t_{\max}$ is calculated as

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

2) AREA UNDER THE AR-AN CURVE (AUC)

Another metric for temporal action proposal 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 [49]–[51]. Note that at an AN of 0 where no proposals are given, AR is trivially 0. Using AR@AN from (8), AR-AN AUC is calculated as

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

D. TEMPORAL LOCALIZATION/DETECTION METRICS

Like temporal proposal, there are two main metrics for TAL/D and both are used across many challenges [49]–[52], [73], [95], [102]. 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, $\forall i \in \{1, \dots, n\}$:

- $y^{(i)} = \{(s_1^{(i)}, l_1^{(i)}), \dots\}$ is 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 $l_j^{(i)} \in C$,

- $\hat{y}^{(i)} = \{(\hat{s}_1^{(i)}, \hat{l}_1^{(i)}, c_1^{(i)}), \dots\}$ 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 $x^{(i)}$, and
- $\text{tIoU}(s_a^{(i)}, \hat{s}_b^{(i)})$ is the temporal IoU between a ground truth segment and a proposal.

1) MEAN AVERAGE PRECISION AT tIoU THRESHOLD (mAP tIoU@t)

This metric is the arithmetic mean of the interpolated average precision 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 VI-A2 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., $\text{TP}_j(r) = 1$ if $\text{tIoU}(\hat{s}_a^{(i)}, s_b^{(i)}) \geq t$ 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 (2), (3), and (4). Note that in this case, n in (3) must be replaced with the number of prediction tuples for the class j .

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 $\eta = 0.05$ [49]–[52], [73]. Therefore, average mAP is computed as

average mAP

$$\begin{aligned} &= \text{mAP tIoU}@0.5:t_{\max}:\eta \\ &= \frac{1}{\frac{t_{\max}-0.5}{\eta} + 1} \sum_{j=0}^{(t_{\max}-0.5)/\eta} \text{mAP tIoU}@ (0.5 + j\eta). \quad (10) \end{aligned}$$

E. 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, $\forall i \in \{1, \dots, n\}$:

- $y^{(i)} = \{(f_1^{(i)}, b_1^{(i)}, l_1^{(i)}), \dots\}$ is 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 $l_j^{(i)} \in C$,
- $\hat{y}^{(i)} = \{(\hat{f}_1^{(i)}, \hat{b}_1^{(i)}, \hat{l}_1^{(i)}, c_1^{(i)}), \dots\}$ 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 $\hat{l}_j^{(i)}$,

- $\text{tube}_j^{(i)}$ is a spatiotemporal tube in video $x^{(i)}$ defined as a linked set of bounding boxes $b_k^{(i)}, b_l^{(i)}, b_m^{(i)}, \dots$ with the same class label ($l_k^{(i)} = l_l^{(i)} = l_m^{(i)} = \dots$) in adjacent frames ($k = l - 1 = m - 2 = \dots$),
- $\text{sIoU}(b_a^{(i)}, \hat{b}_b^{(i)})$ is the spatial IoU between a ground truth bounding box and a proposed bounding box (note: this requires $f_a^{(i)} = \hat{f}_b^{(i)}$), and
- $\text{stIoU}(\text{tube}_a^{(i)}, \widehat{\text{tube}}_b^{(i)})$ is the spatiotemporal IoU between a ground truth and proposed tubes.

1) FRAME-LEVEL MEAN AVERAGE PRECISION (FRAME-mAP)

This metric 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 [50]–[52]. 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 a ground truth box $b_b^{(i)}$ exists on the same frame with the same class label that has a sIoU greater than or equal to a given threshold t that was not already detected by another proposed box higher in the ranked list (i.e., $\text{TP}_j(r) = 1$ if $\text{sIoU}(\hat{b}_a^{(i)}, b_b^{(i)}) \geq t$ 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 (2), (3) and (4). Note that n in (3) must be replaced with the number of prediction tuples for the class j (i.e., the length of ranked list L_j).

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 take up more frames and weigh 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 was employed for use with both AVA and J-HMDB-21 datasets [71], [75]. 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 $\widehat{\text{tube}}_a^{(i)}$ at rank r in list L_j is counted as a true positive if a ground truth tube $\text{tube}_b^{(i)}$ exists with the same class label that has a stIoU greater than or equal to a given threshold t that was not already detected by another proposed tube higher in the ranked list (i.e., $\text{TP}_j(r) = 1$ if $\text{stIoU}(\widehat{\text{tube}}_a^{(i)}, \text{tube}_b^{(i)}) \geq t$ 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 (2), (3) and (4). Note that in this case, n in (3) must be replaced with the number of prediction tubes for the class j .

VII. 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 model architecture families, and defined common metrics for assessing models. We hope that this tutorial clarifies terminology and expands your understanding of these problems 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.

ACKNOWLEDGMENT

The authors would like to thank the following individuals who have provided feedback on this article: Jeremy Kepner, Andrew Kirby, Alex Knapp, Alison Louthain, Judy Marchese, 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] H. Touvron, A. Vedaldi, M. Douze, and H. Jegou, "Fixing the train-test resolution discrepancy," in *Advances in Neural Information Processing Systems*, vol. 32, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2019, pp. 8252–8262. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/file/d03a857a23b5285736c4d55e0bb067c8-Paper.pdf>
- [2] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
- [3] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba, "Places: A 10 million image database for scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1452–1464, Jun. 2018.
- [4] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.
- [5] S. Borse, Y. Wang, Y. Zhang, and F. Porikli, "InverseForm: A loss function for structured boundary-aware segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5901–5911.
- [6] A. Diba, M. Fayyaz, V. Sharma, M. Paluri, J. Gall, R. Stiefelhagen, and L. Van Gool, "Large scale holistic video understanding," in *Computer Vision—ECCV 2020*, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham, Switzerland: Springer, 2020, pp. 593–610.
- [7] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, Jun. 2013, doi: [10.3390/computers2020088](https://doi.org/10.3390/computers2020088).
- [8] J. M. Chaquet, E. J. Carmona, and A. Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Comput. Vis. Image Understand.*, vol. 117, no. 6, pp. 633–659, Jun. 2013. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1077314213000295>
- [9] G. Cheng, Y. Wan, A. N. Saudagar, K. Namuduri, and B. P. Buckles, "Advances in human action recognition: A survey," 2015, *arXiv:1501.05964*. [Online]. Available: <http://arxiv.org/abs/1501.05964>
- [10] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 6, pp. 790–808, Nov. 2012.
- [11] J. R. Kwapisz, G. M. Weiss, and S. A. Moore, "Activity recognition using cell phone accelerometers," *ACM SIGKDD Explor. Newslett.*, vol. 12, no. 2, pp. 74–82, Mar. 2011. 10.1145/1964897.1964918
- [12] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S016786551830045X>
- [13] S. Min Kang and R. P. Wildes, "Review of action recognition and detection methods," 2016, *arXiv:1610.06906*. [Online]. Available: <http://arxiv.org/abs/1610.06906>
- [14] F. Zhu, L. Shao, J. Xie, and Y. Fang, "From handcrafted to learned representations for human action recognition: A survey," *Image Vis. Comput.*, vol. 55, pp. 42–52, Nov. 2016. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0262885616301081>
- [15] G. Guo and A. Lai, "A survey on still image based human action recognition," *Pattern Recognit.*, vol. 47, no. 10, pp. 3343–3361, Oct. 2014. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0031320314001642>
- [16] M. Monfort, C. Vondrick, A. Oliva, A. Andonian, B. Zhou, K. Ramakrishnan, S. A. Bargal, T. Yan, L. Brown, Q. Fan, and D. Gutfreund, "Moments in time dataset: One million videos for event understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 2, pp. 502–508, Feb. 2020.
- [17] M. Monfort, K. Ramakrishnan, A. Andonian, B. A. McNamara, A. Lascelles, B. Pan, Q. Fan, D. Gutfreund, R. Feris, and A. Oliva, "Multi-moments in time: Learning and interpreting models for multi-action video understanding," 2019, *arXiv:1911.00232*. [Online]. Available: <https://arxiv.org/abs/1911.00232>
- [18] Merriam-Webster. *Action*. Accessed: Aug. 18, 2020. [Online]. Available: <https://www.merriam-webster.com/dictionary/action>
- [19] Oxford University Press. *Action*. Accessed: Aug. 18, 2020. [Online]. Available: <http://www.oed.com/view/Entry/1938>
- [20] R. Poppe, "A survey on vision-based human action recognition," *Image Vis. Comput.*, vol. 28, no. 6, pp. 976–990, 2010. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0262885609002704>
- [21] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Comput. Vis. Image Understand.*, vol. 115, no. 2, pp. 224–241, Feb. 2011. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1077314210002171>
- [22] M. A. R. Ahad, J. Tan, H. Kim, and S. Ishikawa, "Action dataset—A survey," in *Proc. SICE Annu. Conf.*, 2011, pp. 1650–1655.
- [23] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang, "RGB-D-based action recognition datasets: A survey," *Pattern Recognit.*, vol. 60, pp. 86–105, Dec. 2016. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0031320316301029>
- [24] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image Vis. Comput.*, vol. 60, pp. 4–21, Apr. 2017. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0262885617300343>
- [25] M. Koohzadi and N. M. Charkari, "Survey on deep learning methods in human action recognition," *IET Comput. Vis.*, vol. 11, no. 8, pp. 623–632, Dec. 2017.
- [26] M. Asadi-Aghbolaghi, A. Clapés, M. Bellantonio, H. J. Escalante, V. Ponce-López, X. Baró, I. Guyon, S. Kasaei, and S. Escalera, "A survey on deep learning based approaches for action and gesture recognition in image sequences," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2017, pp. 476–483.
- [27] Y. Kong and Y. Fu, "Human action recognition and prediction: A survey," 2018, *arXiv:1806.11230*. [Online]. Available: <http://arxiv.org/abs/1806.11230>

- [28] H.-B. Zhang, Y.-X. Zhang, B. Zhong, Q. Lei, L. Yang, J.-X. Du, and D.-S. Chen, "A comprehensive survey of vision-based human action recognition methods," *Sensors*, vol. 19, no. 5, p. 1005, Feb. 2019. [Online]. Available: <https://www.mdpi.com/1424-8220/19/5/1005>
- [29] A. Bhoi, "Spatio-temporal action recognition: A survey," 2019, *arXiv:1901.09403*. [Online]. Available: <http://arxiv.org/abs/1901.09403>
- [30] T. Singh and D. K. Vishwakarma, "Video benchmarks of human action datasets: A review," *Artif. Intell. Rev.*, vol. 52, no. 2, pp. 1107–1154, Aug. 2019, doi: [10.1007/s10462-018-9651-1](https://doi.org/10.1007/s10462-018-9651-1).
- [31] H. Xia and Y. Zhan, "A survey on temporal action localization," *IEEE Access*, vol. 8, pp. 70477–70487, 2020.
- [32] A. Rasouli, "Deep learning for vision-based prediction: A survey," 2020, *arXiv:2007.00095*. [Online]. Available: <http://arxiv.org/abs/2007.00095>
- [33] S. K. Yadav, K. Tiwari, H. M. Pandey, and S. A. Akbar, "A review of multimodal human activity recognition with special emphasis on classification, applications, challenges and future directions," *Knowl.-Based Syst.*, vol. 223, Jul. 2021, Art. no. 106970. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0950705121002331>
- [34] T.-K. Kim, S.-F. Wong, and R. Cipolla, "Tensor canonical correlation analysis for action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8, doi: [10.1109/CVPR.2007.383137](https://doi.org/10.1109/CVPR.2007.383137).
- [35] J. C. Niebles and L. Fei-Fei, "A hierarchical model of shape and appearance for human action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2007, pp. 1–8, doi: [10.1109/CVPR.2007.383132](https://doi.org/10.1109/CVPR.2007.383132).
- [36] R. Girdhar, D. Ramanan, A. Gupta, J. Sivic, and B. Russell, "Action-VLAD: Learning spatio-temporal aggregation for action classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 971–980.
- [37] Z. Shou, D. Wang, and S. Chang, "Temporal action localization in untrimmed videos via multi-stage CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1049–1058.
- [38] Y. Zhao, Y. Xiong, L. Wang, Z. Wu, X. Tang, and D. Lin, "Temporal action detection with structured segment networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2914–2923.
- [39] Y. Chao, S. Vijayanarasimhan, B. Seybold, D. A. Ross, J. Deng, and R. Sukthankar, "Rethinking the faster R-CNN architecture for temporal action localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1130–1139.
- [40] T. Lin, X. Zhao, H. Su, C. Wang, and M. Yang, "BSN: Boundary sensitive network for temporal action proposal generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.
- [41] J. Gao, K. Chen, and R. Nevatia, "CTAP: Complementary temporal action proposal generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 68–83.
- [42] Y. Liu, L. Ma, Y. Zhang, W. Liu, and S.-F. Chang, "Multi-granularity generator for temporal action proposal," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3604–3613.
- [43] V. Escorcia, C. D. Dao, M. Jain, B. Ghanem, and C. Snoek, "Guess where? Actor-supervision for spatiotemporal action localization," *Comput. Vis. Image Understand.*, vol. 192, Mar. 2020, Art. no. 102886. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1077314219301687>
- [44] R. Zeng, W. Huang, C. Gan, M. Tan, Y. Rong, P. Zhao, and J. Huang, "Graph convolutional networks for temporal action localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7094–7103.
- [45] H. Xu, A. Das, and K. Saenko, "R-C3D: Region convolutional 3D network for temporal activity detection," in *Proc. ICCV*, Oct. 2017, pp. 5783–5792.
- [46] S. Saha, G. Singh, M. Sapienza, P. H. S. Torr, and F. Cuzzolin, "Spatio-temporal human action localisation and instance segmentation in temporally untrimmed videos," 2017, *arXiv:1707.07213*. [Online]. Available: <http://arxiv.org/abs/1707.07213>
- [47] J. Ji, S. Buch, A. Soto, and J. C. Niebles, "End-to-end joint semantic segmentation of actors and actions in video," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 702–717.
- [48] S. Saha, G. Singh, M. Sapienza, P. H. S. Torr, and F. Cuzzolin, *Spatio-Temporal Action Instance Segmentation Localisation*. Cham, Switzerland: Springer, 2020, pp. 141–161, doi: [10.1007/978-3-030-46732-6_8](https://doi.org/10.1007/978-3-030-46732-6_8).
- [49] B. Ghanem, J. Carlos Niebles, C. Snoek, F. Caba Heilbron, H. Alwassel, R. Khrisna, V. Escorcia, K. Hata, and S. Buch, "ActivityNet challenge 2017 summary," 2017, *arXiv:1710.08011*. [Online]. Available: <http://arxiv.org/abs/1710.08011>
- [50] B. Ghanem, J. C. Niebles, C. Snoek, F. C. Heilbron, H. Alwassel, V. Escorcia, R. Krishna, S. Buch, and C. D. Dao, "The ActivityNet large-scale activity recognition challenge 2018 summary," 2018, *arXiv:1808.03766*. [Online]. Available: <http://arxiv.org/abs/1808.03766>
- [51] C. Snoek, J. C. Niebles, B. Ghanem, F. C. Heilbron, H. Alwassel, V. Escorcia, R. Krishna, S. Buch, and F. Xu, "International challenge on activity recognition," in *Proc. Comput. Vis. Pattern Recognit. (CVPR) Conf.*, Long Beach, CA, USA, 2019. [Online]. Available: <http://activity-net.org/challenges/2019/index.html>
- [52] (2020). *International Challenge on Activity Recognition*. [Online]. Available: <http://activity-net.org/challenges/2020/index.html>
- [53] D. Chen and W. B. Dolan, "Collecting highly parallel data for paraphrase evaluation," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2011, pp. 190–200.
- [54] A. Torabi, C. Pal, H. Larochelle, and A. Courville, "Using descriptive video services to create a large data source for video annotation research," 2015, *arXiv:1503.01070*. [Online]. Available: <http://arxiv.org/abs/1503.01070>
- [55] A. Rohrbach, M. Rohrbach, N. Tandon, and B. Schiele, "A dataset for movie description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3202–3212.
- [56] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 706–715.
- [57] S. Li, Z. Tao, K. Li, and Y. Fu, "Visual to text: Survey of image and video captioning," *IEEE Trans. Emerg. Topics Comput. Intell.*, vol. 3, no. 4, pp. 297–312, Aug. 2019.
- [58] H. Alwassel, F. Caba Heilbron, and B. Ghanem, "Action search: Spotting actions in videos and its application to temporal action localization," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 251–266.
- [59] O. Köpüklü, X. Wei, and G. Rigoll, "You only watch once: A unified CNN architecture for real-time spatiotemporal action localization," 2019, *arXiv:1911.06644*. [Online]. Available: <http://arxiv.org/abs/1911.06644>
- [60] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Sep. 2009, doi: [10.1007/s11263-009-0275-4](https://doi.org/10.1007/s11263-009-0275-4).
- [61] R. Yao, G. Lin, S. Xia, J. Zhao, and Y. Zhou, "Video object segmentation and tracking: A survey," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 4, pp. 1–47, 2020.
- [62] C. Schudt, I. Laptev, and B. Caputo, "Recognizing human actions: A local SVM approach," in *Proc. 17th Int. Conf. Pattern Recognit.*, vol. 3, Aug. 2004, pp. 32–36.
- [63] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Proc. Int. Conf. Comput. Vis. (ICCV) Volume 1*, Oct. 2005, vol. 2, pp. 1395–1402.
- [64] I. Laptev and P. Perez, "Retrieving actions in movies," in *Proc. IEEE 11th Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8, doi: [10.1109/ICCV.2007.4409105](https://doi.org/10.1109/ICCV.2007.4409105).
- [65] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 2929–2936.
- [66] S. Oh *et al.*, "A large-scale benchmark dataset for event recognition in surveillance video," in *Proc. CVPR*, Jun. 2011, pp. 3153–3160.
- [67] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "HMDB: A large video database for human motion recognition," in *Proc. Int. Conf. Comput. Vis.*, Nov. 2011, pp. 2556–2563.
- [68] K. Soomro, A. R. Zamir, and M. Shah, "UCF101: A dataset of 101 human actions classes from videos in the wild," 2012, *arXiv:1212.0402*. [Online]. Available: <http://arxiv.org/abs/1212.0402>
- [69] Y.-G. Jiang, J. Liu, A. R. Zamir, I. Laptev, M. Piccardi, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," in *Proc. Int. Conf. Comput. Vis. (ICCV)*, Sydney, NSW, Australia, 2013. [Online]. Available: <http://www.crcv.ucf.edu/ICCV13-Action-Workshop/>
- [70] H. Idrees, A. R. Zamir, Y.-G. Jiang, A. Ghorban, I. Laptev, R. Sukthankar, and M. Shah, "The THUMOS challenge on action recognition for videos 'in the wild,'" *Comput. Vis. Image Understand.*, vol. 155, pp. 1–23, Feb. 2017.
- [71] H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black, "Towards understanding action recognition," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3192–3199.

- [72] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "Large-scale video classification with convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Sep. 2014, pp. 1725–1732.
- [73] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Nibbles, "ActivityNet: A large-scale video benchmark for human activity understanding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 961–970.
- [74] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017, *arXiv:1705.06950*. [Online]. Available: <http://arxiv.org/abs/1705.06950>
- [75] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, "AVA: A video dataset of spatio-temporally localized atomic visual actions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 6047–6056.
- [76] S. Yeung, O. Russakovsky, N. Jin, M. Andriluka, G. Mori, and L. Fei-Fei, "Every moment counts: Dense detailed labeling of actions in complex videos," *Int. J. Comput. Vis.*, vol. 126, nos. 2–4, pp. 375–389, 2018.
- [77] J. Carreira, E. Noland, A. Banki-Horvath, C. Hillier, and A. Zisserman, "A short note about kinetics-600," 2018, *arXiv:1808.01340*. [Online]. Available: <http://arxiv.org/abs/1808.01340>
- [78] Y. Li, M. Liu, and J. M. Reh, "In the eye of beholder: Joint learning of gaze and actions in first person video," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 619–635.
- [79] F. Mahdisoltani, G. Berger, W. Gharbieh, D. Fleet, and R. Memisevic, "On the effectiveness of task granularity for transfer learning," 2018, *arXiv:1804.09235*. [Online]. Available: <http://arxiv.org/abs/1804.09235>
- [80] G. A. Sigurdsson, A. Gupta, C. Schmid, A. Farhadi, and K. Alahari, "Charades-ego: A large-scale dataset of paired third and first person videos," 2018, *arXiv:1804.09626*. [Online]. Available: <http://arxiv.org/abs/1804.09626>
- [81] J. Materzynska, G. Berger, I. Bax, and R. Memisevic, "The jester dataset: A large-scale video dataset of human gestures," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 2874–2882.
- [82] J. Carreira, E. Noland, C. Hillier, and A. Zisserman, "A short note on the kinetics-700 human action dataset," 2019, *arXiv:1907.06987*. [Online]. Available: <http://arxiv.org/abs/1907.06987>
- [83] H. Zhao, A. Torralba, L. Torresani, and Z. Yan, "HACS: Human action clips and segments dataset for recognition and temporal localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8668–8678.
- [84] J. Liu, A. Shahroudy, M. Perez, G. Wang, L.-Y. Duan, and A. C. Kot, "NTU RGB+D 120: A large-scale benchmark for 3D human activity understanding," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, Oct. 2020.
- [85] D. Damen, H. Doughty, G. M. Farinella, A. Furnari, E. Kazakos, J. Ma, D. Molitanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Rescaling egocentric vision," 2020, *arXiv:2006.13256*. [Online]. Available: <http://arxiv.org/abs/2006.13256>
- [86] A. Li, M. Thotakuri, D. A. Ross, J. Carreira, A. Vostroikov, and A. Zisserman, "The AVA-kinetics localized human actions video dataset," 2020, *arXiv:2005.00214*. [Online]. Available: <http://arxiv.org/abs/2005.00214>
- [87] A. Piergiovanni and M. S. Ryoo, "AViD dataset: Anonymized videos from diverse countries," 2020, *arXiv:2007.05515*. [Online]. Available: <http://arxiv.org/abs/2007.05515>
- [88] D. Shao, Y. Zhao, B. Dai, and D. Lin, "FineGym: A hierarchical video dataset for fine-grained action understanding," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2616–2625.
- [89] J. Chung, C.-H. Wu, H.-R. Yang, Y.-W. Tai, and C.-K. Tang, "HAA500: Human-centric atomic action dataset with curated videos," 2020, *arXiv:2009.05224*. [Online]. Available: <http://arxiv.org/abs/2009.05224>
- [90] R. Goyal, S. E. Kahou, V. Michalski, J. Materzynska, S. Westphal, H. Kim, V. Haenel, I. Freund, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thureau, I. Bax, and R. Memisevic, "The 'something something' video database for learning and evaluating visual common sense," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 5842–5850.
- [91] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large scale dataset for 3D human activity analysis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1010–1019.
- [92] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, "Hollywood in homes: Crowdsourcing data collection for activity understanding," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer, 2016, pp. 510–526.
- [93] H. Zhao, Z. Yan, H. Wang, L. Torresani, and A. Torralba, "SLAC: A sparsely labeled dataset for action classification and localization," MIT, Facebook, Tech. Rep., Dec. 2017. [Online]. Available: <https://www.arxiv-vanity.com/papers/1712.09374/>
- [94] F. C. Heilbron and J. C. Nibbles, "Collecting and annotating human activities in web videos," in *Proc. Int. Conf. Multimedia Retr.*, Apr. 2014, p. 377.
- [95] Y.-G. Jiang, J. Liu, A. R. Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Zürich, Switzerland, 2014. [Online]. Available: <http://csrcv.ucf.edu/THUMOS14/>
- [96] D. F. Fouhey, W.-C. Kuo, A. A. Efros, and J. Malik, "From lifestyle vlogs to everyday interactions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4991–5000.
- [97] H. S. Koppula, R. Gupta, and A. Saxena, "Learning human activities and object affordances from RGB-D videos," *Int. J. Robot. Res.*, vol. 32, no. 8, pp. 951–970, 2013, doi: [10.1177/0278364913478446](https://doi.org/10.1177/0278364913478446).
- [98] D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Molitanti, J. Munro, T. Perrett, W. Price, and M. Wray, "Scaling egocentric vision: The epic-kitchens dataset," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 720–736.
- [99] H. Wang. (2013). *Lear-Inria Submission for the Thumos Workshop*. [Online]. Available: [http://csrcv.ucf.edu/THUMOS14/index.files/NotebookPapers13/Paper203920\(Supplementary\).pdf](http://csrcv.ucf.edu/THUMOS14/index.files/NotebookPapers13/Paper203920(Supplementary).pdf)
- [100] M. Jain, J. V. Gemert, and C. G. M. Snoek, "University of Amsterdam at thumos challenge 2014," ISLA, Inform. Inst., Univ. Amsterdam, Amsterdam, The Netherlands, Tech. Rep., 2014. [Online]. Available: <http://csrcv.ucf.edu/THUMOS14/papers/University20of20Amsterdam.pdf>
- [101] D. Oneata, J. Verbeek, and C. Schmid. (2014). *The Lear Submission at Thumos 2014*. [Online]. Available: <http://csrcv.ucf.edu/THUMOS14/papers/INRIA20LEAR.pdf>
- [102] A. Gorban, H. Idrees, Y.-G. Jiang, A. Roshan Zamir, I. Laptev, M. Shah, and R. Sukthankar, "THUMOS challenge: Action recognition with a large number of classes," in *Proc. Comput. Vis. Pattern Recognit. Conf. (CVPR)*, Boston, MA, USA, 2015. [Online]. Available: <http://www.thumos.info/>
- [103] Z. Xu, L. Zhu, Y. Yang, and A. Hauptmann. (Jun. 2015). *UTS-CMU at Thumos 2015*. [Online]. Available: http://storage.googleapis.com/www.thumos.info/thumos15_notebooks/TH15_UTS&CMU.pdf
- [104] J. Yuan, Y. Pei, B. Ni, P. Moulin, and A. Kassim. (2015). *ADSC Submission at Thumos Challenge 2015*. [Online]. Available: http://storage.googleapis.com/www.thumos.info/thumos15_notebooks/TH15_A%DSC.pdf
- [105] Y. Xiong, L. Wang, Z. Wang, B. Zhang, H. Song, W. Li, D. Lin, Y. Qiao, L. Van Gool, and X. Tang, "CUHK & ETHZ & SIAT submission to ActivityNet challenge 2016," 2016, *arXiv:1608.00797*. [Online]. Available: <http://arxiv.org/abs/1608.00797>
- [106] Y. Bian, C. Gan, X. Liu, F. Li, X. Long, Y. Li, H. Qi, J. Zhou, S. Wen, and Y. Lin, "Revisiting the effectiveness of off-the-shelf temporal modeling approaches for large-scale video classification," 2017, *arXiv:1708.03805*. [Online]. Available: <http://arxiv.org/abs/1708.03805>
- [107] T. Lin, X. Zhao, and Z. Shou, "Temporal convolution based action proposal: Submission to ActivityNet 2017," 2017, *arXiv:1707.06750*. [Online]. Available: <http://arxiv.org/abs/1707.06750>
- [108] D. He, F. Li, Q. Zhao, X. Long, Y. Fu, and S. Wen, "Exploiting spatial-temporal modelling and multi-modal fusion for human action recognition," 2018, *arXiv:1806.10319*. [Online]. Available: <http://arxiv.org/abs/1806.10319>
- [109] C. Li, Z. Hou, J. Chen, Y. Bu, J. Zhou, Q. Zhong, D. Xie, and S. Pu, "Team deep-hri moments in time challenge 2018 technical report," Hikvision Res. Inst., Hangzhou, China, Tech. Rep., 2018. [Online]. Available: http://moments.csail.mit.edu/challenge2018/DEEP_HRI.pdf
- [110] S. Guan and H. Li, "SYSU ISEE submission to moments in time challenge 2018," School Data Comput. Sci., Sun Yat-Sen Univ., Guangzhou, China, Tech. Rep. [Online]. Available: http://moments.csail.mit.edu/challenge2018/SYSU_isee.pdf

- [111] J. Jiang, Y. Cao, L. Song, S. Zhang, Y. Li, Z. Xu, Q. Wu, C. Gan, C. Zhang, and G. Yu, "Human centric spatio-temporal action localization," in *Proc. ActivityNet Workshop (CVPR)*, Salt Lake City, UT, USA, 2018. [Online]. Available: https://www.skicyu.org/AVA/AVA_report.pdf
- [112] Y. Lee, J. Fiscus, A. Godil, A. Delgado, J. Golden, L. Diduch, and M. Hubert, "Summary of the 2019 activity detection in extended videos prize challenge," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops (WACVW)*, Mar. 2020, pp. 148–154.
- [113] Z. Qiu, D. Li, Y. Li, Q. Cai, Y. Pan, and T. Yao, "Trimmed action recognition, dense-captioning events in videos, and spatio-temporal action localization with focus on ActivityNet challenge 2019," 2019, *arXiv:1906.07016*. [Online]. Available: <http://arxiv.org/abs/1906.07016>
- [114] X. Wang, Y. Wu, L. Zhu, and Y. Yang, "Baidu-UTS submission to the EPIC-kitchens action recognition challenge 2019," 2019, *arXiv:1906.09383*. [Online]. Available: <http://arxiv.org/abs/1906.09383>
- [115] A. Furnari and G. Farinella, "What would you expect? Anticipating egocentric actions with rolling-unrolling LSTMs and modality attention," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6252–6261.
- [116] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 6202–6211.
- [117] (2019). *Workshop on Multi-Modal Video Analysis and Moments in Time Challenge*. [Online]. Available: <https://sites.google.com/view/multimodalvideo/home>
- [118] M. Zhang, H. Shao, G. Song, Y. Liu, and J. Yan, "Top-1 solution of multi-moments in time challenge 2019," 2020, *arXiv:2003.05837*. [Online]. Available: <http://arxiv.org/abs/2003.05837>
- [119] S. Zhang, H. Peng, L. Yang, J. Fu, and J. Luo, "Learning sparse 2D temporal adjacent networks for temporal action localization," 2019, *arXiv:1912.03612*. [Online]. Available: <http://arxiv.org/abs/1912.03612>
- [120] X. Wang, B. Ma, Z. Qing, Y. Sang, C. Gao, S. Zhang, and N. Sang, "CBR-Net: Cascade boundary refinement network for action detection: Submission to activitynet challenge 2020 (task 1)," 2020, *arXiv:2006.07526*. [Online]. Available: <http://arxiv.org/abs/2006.07526>
- [121] K. Corona, K. Osterdahl, R. Collins, and A. Hoogs, "MEVA: A large-scale multiview, multimodal video dataset for activity detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1060–1068.
- [122] Z. Qing, X. Wang, Y. Sang, C. Gao, S. Zhang, and N. Sang, "Temporal fusion network for temporal action localization: Submission to ActivityNet challenge 2020 (task E)," 2020, *arXiv:2006.07520*. [Online]. Available: <http://arxiv.org/abs/2006.07520>
- [123] J. Gao, T. Lin, X. Long, D. He, F. Li, X. Li, S. Wen, and E. Ding, "Multi-modal fusion network based on relation-aware pyramid network for temporal action localization," Shanghai Jiao Tong Univ., Shanghai, China. [Online]. Available: http://hacs.csail.mit.edu/challenge/report_cvpr20/hacs_report_baidu_vis%.pdf
- [124] S. Chen, J. Pan, G. Song, M. Zhang, H. Shao, Z. Lin, J. Shao, H. Li, and Y. Liu, "1st place solution for AVA-kinetics crossover in ActivityNet challenge 2020," 2020, *arXiv:2006.09116*. [Online]. Available: <http://arxiv.org/abs/2006.09116>
- [125] L. Smaira, J. Carreira, E. Noland, E. Clancy, A. Wu, and A. Zisserman, "A short note on the Kinetics-700-2020 human action dataset," 2020, *arXiv:2010.10864*. [Online]. Available: <http://arxiv.org/abs/2010.10864>
- [126] U. Demir, Y. S. Rawat, and M. Shah, "TinyVIRAT: Low-resolution video action recognition," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 7387–7394.
- [127] P. Tirupattur, A. J. Rana, T. Sangam, S. Vyas, Y. S. Rawat, and M. Shah, "TinyAction challenge: Recognizing real-world low-resolution activities in videos," 2021, *arXiv:2107.11494*. [Online]. Available: <http://arxiv.org/abs/2107.11494>
- [128] X. Wang, Z. Qing, Z. Huang, Y. Feng, S. Zhang, J. Jiang, M. Tang, C. Gao, and N. Sang, "Proposal relation network for temporal action detection," 2021, *arXiv:2106.11812*. [Online]. Available: <http://arxiv.org/abs/2106.11812>
- [129] Z. Qing, X. Wang, Z. Huang, Y. Feng, S. Zhang, J. Jiang, M. Tang, C. Gao, and N. Sang, "Exploring stronger feature for temporal action localization," 2021, *arXiv:2106.13014*. [Online]. Available: <http://arxiv.org/abs/2106.13014>
- [130] A. Deliege, A. Cioppa, S. Giancola, M. J. Seikavandi, J. V. Dueholm, K. Nasrollahi, B. Ghanem, T. B. Moeslund, and M. Van Droogenbroeck, "SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 4508–4519.
- [131] X. Zhou, L. Kang, Z. Cheng, B. He, and J. Xin, "Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection," 2021, *arXiv:2106.14447*. [Online]. Available: <http://arxiv.org/abs/2106.14447>
- [132] Y. Feng, J. Jiang, Z. Huang, Z. Qing, X. Wang, S. Zhang, M. Tang, and Y. Gao, "Relation modeling in spatio-temporal action localization," 2021, *arXiv:2106.08061*. [Online]. Available: <http://arxiv.org/abs/2106.08061>
- [133] L. Taylor and G. Nitschke, "Improving deep learning with generic data augmentation," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Nov. 2018, pp. 1542–1547.
- [134] X. Lin, Y.-L. Ma, L.-Z. Ma, and R.-L. Zhang, "A survey for image resizing," *J. Zhejiang Univ. Sci. C*, vol. 15, no. 9, pp. 697–716, Sep. 2014, doi: [10.1631/jzus.C1400102](https://doi.org/10.1631/jzus.C1400102).
- [135] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks: Towards good practices for deep action recognition," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer, 2016, pp. 20–36.
- [136] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "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. Red Hook, NY, USA: Curran Associates, 2012, pp. 1097–1105. [Online]. Available: <http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf>
- [137] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," 2014, *arXiv:1405.3531*. [Online]. Available: <http://arxiv.org/abs/1405.3531>
- [138] C. Shorten and T. M. Khoshgofaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, p. 60, 2019, doi: [10.1186/s40537-019-0197-0](https://doi.org/10.1186/s40537-019-0197-0).
- [139] L. Wang, Y. Xiong, Z. Wang, Y. Qiao, D. Lin, X. Tang, and L. Van Gool, "Temporal segment networks for action recognition in videos," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2740–2755, Nov. 2019.
- [140] K. Simonyan and A. Zisserman, "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. Red Hook, NY, USA: Curran Associates, 2014, pp. 568–576.
- [141] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jul. 2017, pp. 6299–6308.
- [142] C. Wu, J. Zhang, S. Savarese, and A. Saxena, "Watch-n-patch: Unsupervised understanding of actions and relations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4362–4370.
- [143] T. Han, W. Xie, and A. Zisserman, "Video representation learning by dense predictive coding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1483–1492.
- [144] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recognit. Workshop*, Jun. 2014, pp. 806–813.
- [145] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 1933–1941.
- [146] G. Farnéback, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, J. Bigun and T. Gustavsson, Eds. Berlin, Germany: Springer, 2003, pp. 363–370.
- [147] G. Bradski, "The OpenCV Library," *Dr. Dobbs's J. Softw. Tools*, vol. 25, no. 11, pp. 120–123, 2000. [Online]. Available: <https://opencv.org/>
- [148] O. Hommos, S. L. Pintea, P. S. Mettes, and J. C. van Gemert, "Using phase instead of optical flow for action recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, Sep. 2018, pp. 678–691.
- [149] N. T. Ouellette, H. Xu, and E. Bodenschatz, "A quantitative study of three-dimensional Lagrangian particle tracking algorithms," *Exp. Fluids*, vol. 40, no. 2, pp. 301–313, 2006.
- [150] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–8, Aug. 2012, doi: [10.1145/2185520.2185561](https://doi.org/10.1145/2185520.2185561).
- [151] J. J. Gibson, *The Perception of the Visual World*. Boston, MA, USA: Houghton Mifflin, 1950.
- [152] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. Imag. Understand. Workshop*, 1981, pp. 121–130.

- [153] B. K. Horn and B. G. Schunck, "Determining optical flow," in *Techniques and Applications of Image Understanding*, J. J. Pearson, Ed., vol. 281. Bellingham, WA, USA: SPIE, 1981, pp. 319–331, doi: [10.1117/12.965761](https://doi.org/10.1117/12.965761).
- [154] C. Zach, T. Pock, and H. Bischof, "A duality based approach for realtime tv-l1 optical flow," in *Pattern Recognition*, F. A. Hamprecht, C. Schnörr, and B. Jähne, Eds. Berlin, Germany: Springer, 2007, pp. 214–223.
- [155] S. S. Beauchemin and J. L. Barron, "The computation of optical flow," *ACM Comput. Surv.*, vol. 27, no. 3, pp. 433–466, Sep. 1995, doi: [10.1145/212094.212141](https://doi.org/10.1145/212094.212141).
- [156] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3551–3558.
- [157] L. Sevilla-Lara, Y. Liao, F. Güney, V. Jampani, A. Geiger, and M. J. Black, "On the integration of optical flow and action recognition," in *Pattern Recognition*, T. Brox, A. Bruhn, and M. Fritz, Eds. Cham, Switzerland: Springer, 2019, pp. 281–297.
- [158] Y. LeCun and M. Ranzato, "Deep learning tutorial," in *Proc. Tuts. Int. Conf. Mach. Learn. (ICML)*, 2013, pp. 1–29, doi: [10.1.1.366.4088](https://doi.org/10.1.1.366.4088).
- [159] Q. V. Le, "A tutorial on deep learning part 2: Autoencoders, convolutional neural networks and recurrent neural networks," *Google Brain*, vol. 20, pp. 1–20, Oct. 2015, doi: [10.1.1.703.5244](https://doi.org/10.1.1.703.5244).
- [160] W. Luo, Y. Li, R. Urtasun, and R. Zemel, "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. Red Hook, NY, USA: Curran Associates, 2016, pp. 4898–4906.
- [161] S. Hochreiter, "The vanishing gradient problem during learning recurrent neural nets and problem solutions," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 6, no. 2, pp. 107–116, Apr. 1998, doi: [10.1142/S0218488598000094](https://doi.org/10.1142/S0218488598000094).
- [162] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.
- [163] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1–9, doi: [10.1109/CVPR.2015.7298594](https://doi.org/10.1109/CVPR.2015.7298594).
- [164] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2818–2826.
- [165] G. Huang, Z. Liu, and L. van der Maaten, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4700–4708.
- [166] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, vol. 37, Jul. 2015, pp. 448–456. [Online]. Available: <http://proceedings.mlr.press/v37/loffe15.html>
- [167] J. Lei Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*. [Online]. Available: <http://arxiv.org/abs/1607.06450>
- [168] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*. [Online]. Available: <http://arxiv.org/abs/1607.08022>
- [169] Y. Wu and K. He, "Group normalization," in *Proc. ECCV*, Sep. 2018, pp. 3–19.
- [170] Z. Qiu, T. Yao, and T. Mei, "Learning spatio-temporal representation with pseudo-3D residual networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5533–5541.
- [171] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6450–6459.
- [172] D. Ghadiyaram, D. Tran, and D. Mahajan, "Large-scale weakly-supervised pre-training for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 12046–12055.
- [173] L. Wang, W. Li, W. Li, and L. Van Gool, "Appearance-and-relation networks for video classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 1430–1439.
- [174] Y. Chen, Y. Kalantidis, J. Li, S. Yan, and J. Feng, "Multi-fiber networks for video recognition," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 352–367.
- [175] C. Luo and A. Yuille, "Grouped spatial-temporal aggregation for efficient action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5512–5521.
- [176] D. Tran, H. Wang, M. Feiszli, and L. Torresani, "Video classification with channel-separated convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5552–5561.
- [177] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, Jun. 2018.
- [178] N. Hussein, E. Gavves, and A. W. M. Smeulders, "Timeception for complex action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 254–263.
- [179] J. Li, X. Liu, M. Zhang, and D. Wang, "Spatio-temporal deformable 3D ConvNets with attention for action recognition," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107037. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0031320319303383>
- [180] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2019, pp. 7083–7093.
- [181] Y. Zhao, Y. Xiong, and D. Lin, "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. Red Hook, NY, USA: Curran Associates, 2018, pp. 2204–2215. [Online]. Available: <http://papers.nips.cc/paper/7489-trajectory-convolution-for-action-recognition.pdf>
- [182] L. Wang, Y. Qiao, and X. Tang, "Action recognition with trajectory-pooled deep-convolutional descriptors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4305–4314.
- [183] S. Sudhakaran, S. Escalera, and O. Lanz, "Gate-shift networks for video action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1102–1111.
- [184] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: [10.1162/neco.1997.9.8.1735](https://doi.org/10.1162/neco.1997.9.8.1735).
- [185] K. Cho, B. van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Oct. 2014, pp. 1724–1734.
- [186] G. Chen, "A gentle tutorial of recurrent neural network with error backpropagation," 2016, *arXiv:1610.02583*. [Online]. Available: <http://arxiv.org/abs/1610.02583>
- [187] M. Li, "A tutorial on backward propagation through time (BPTT) in the gated recurrent unit (GRU) RNN," Dept. Comput. Sci., Univ. British Columbia, Vancouver, BC, Canada, Tech. Rep., May 2016, doi: [10.13140/RG.2.2.32858.98247](https://doi.org/10.13140/RG.2.2.32858.98247).
- [188] R. C. Staudemeyer and E. Rothstein Morris, "Understanding LSTM—A tutorial into long short-term memory recurrent neural networks," 2019, *arXiv:1909.09586*. [Online]. Available: <http://arxiv.org/abs/1909.09586>
- [189] A. Sherstinsky, "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network," *Phys. D, Nonlinear Phenomena*, vol. 404, Mar. 2020, Art. no. 132306. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0167278919305974>
- [190] Y. Li, J. Zhang, Y. Cheng, K. Huang, and T. Tan, "Semantics-guided multi-level RGB-D feature fusion for indoor semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 1262–1266, doi: [10.1109/ICIP.2017.8296484](https://doi.org/10.1109/ICIP.2017.8296484).
- [191] O. Sagi and L. Rokach, "Ensemble learning: A survey," *WIREs Data Mining Knowl. Discovery*, vol. 8, no. 4, p. e1249, Jul. 2018, doi: [10.1002/widm.1249](https://doi.org/10.1002/widm.1249).
- [192] D. He, Z. Zhou, C. Gan, F. Li, X. Liu, Y. Li, L. Wang, and S. Wen, "StNet: Local and global spatial-temporal modeling for action recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8401–8408.
- [193] B. Jiang, M. Wang, W. Gan, W. Wu, and J. Yan, "STM: Spatiotemporal and motion encoding for action recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2000–2009.
- [194] G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler, "Convolutional learning of spatio-temporal features," in *Computer Vision—ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Germany: Springer, 2010, pp. 140–153.
- [195] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.
- [196] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 4489–4497.

- [197] K. Hara, H. Kataoka, and Y. Satoh, "Can spatiotemporal 3D CNNs retrace the history of 2D CNNs and imagenet?" in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6546–6555.
- [198] Y. Zhu, Z. Lan, S. Newsam, and A. Hauptmann, "Hidden two-stream convolutional networks for action recognition," in *Computer Vision—ACCV 2018*, C. V. Jawahar, H. Li, G. Mori, and K. Schindler, Eds. Cham, Switzerland: Springer, 2019, pp. 363–378.
- [199] N. Crasto, P. Weinzaepfel, K. Alahari, and C. Schmid, "MARS: Motion-augmented RGB stream for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7882–7891.
- [200] J. C. Stroud, D. A. Ross, C. Sun, J. Deng, and R. Sukthankar, "D3D: Distilled 3D networks for video action recognition," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2020, pp. 625–634.
- [201] C. Feichtenhofer, A. Pinz, and R. P. Wildes, "Spatiotemporal multiplier networks for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4768–4777.
- [202] L. Wang, L. Ge, R. Li, and Y. Fang, "Three-stream CNNs for action recognition," *Pattern Recognit. Lett.*, vol. 92, pp. 33–40, Jun. 2017. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0167865517301071>
- [203] L. Wang, J. Zang, Q. Zhang, Z. Niu, G. Hua, and N. Zheng, "Action recognition by an attention-aware temporal weighted convolutional neural network," *Sensors*, vol. 18, no. 7, p. 1979, Jun. 2018, doi: [10.3390/s18071979](https://doi.org/10.3390/s18071979).
- [204] K. Liu, W. Liu, C. Gan, M. Tan, and H. Ma, "T-C3D: Temporal convolutional 3D network for real-time action recognition," in *Proc. Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 7138–7145.
- [205] B. Zhou, A. Andonian, A. Oliva, and A. Torralba, "Temporal relational reasoning in videos," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 803–818.
- [206] M. Zolfaghari, K. Singh, and T. Brox, "ECO: Efficient convolutional network for online video understanding," in *Proc. Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 695–712.
- [207] C. Yang, Y. Xu, J. Shi, B. Dai, and B. Zhou, "Temporal pyramid network for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 591–600.
- [208] J. Donahue, L. A. Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2625–2634.
- [209] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proc. CVPR*, Oct. 2015, pp. 4694–4702.
- [210] Z. Wu, X. Wang, Y.-G. Jiang, H. Ye, and X. Xue, "Modeling spatial-temporal clues in a hybrid deep learning framework for video classification," in *Proc. 23rd ACM Int. Conf. Multimedia*. New York, NY, USA: ACM, Oct. 2015, pp. 461–470, doi: [10.1145/2733373.2806222](https://doi.org/10.1145/2733373.2806222).
- [211] X. Wang, Z. Miao, R. Zhang, and S. Hao, "I3D-LSTM: A new model for human action recognition," in *Proc. IOP Conf. Mater. Sci. Eng.*, vol. 569, Aug. 2019, Art. no. 032035, doi: [10.1088/2F1757-899x/2F569/2F3/2F032035](https://doi.org/10.1088/2F1757-899x/2F569/2F3/2F032035).
- [212] Y. Y. Joefferie and M. Aono, "Action recognition by composite deep learning architecture I3D-DenseLSTM," in *Proc. Int. Conf. Adv. Inform., Concepts, Theory Appl. (ICAICTA)*, Sep. 2019, pp. 1–6, doi: [10.1109/ICAICTA.2019.8904245](https://doi.org/10.1109/ICAICTA.2019.8904245).
- [213] C.-Y. Ma, M.-H. Chen, Z. Kira, and G. AlRegib, "TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition," *Signal Process., Image Commun.*, vol. 71, pp. 76–87, Feb. 2019. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0923596518304922>
- [214] Y. Li, B. Ji, X. Shi, J. Zhang, B. Kang, and L. Wang, "TEA: Temporal excitation and aggregation for action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 909–918.
- [215] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.
- [216] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" 2021, *arXiv:2102.05095*. [Online]. Available: <http://arxiv.org/abs/2102.05095>
- [217] D. Neimark, O. Bar, M. Zohar, and D. Asselmann, "Video transformer network," 2021, *arXiv:2102.00719*. [Online]. Available: <http://arxiv.org/abs/2102.00719>
- [218] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A video vision transformer," 2021, *arXiv:2103.15691*. [Online]. Available: <http://arxiv.org/abs/2103.15691>
- [219] H. Fan, B. Xiong, K. Mangalam, Y. Li, Z. Yan, J. Malik, and C. Feichtenhofer, "Multiscale vision transformers," 2021, *arXiv:2104.11227*. [Online]. Available: <http://arxiv.org/abs/2104.11227>
- [220] J. Gao, Z. Yang, and R. Nevatia, "RED: Reinforced encoder-decoder networks for action anticipation," 2017, *arXiv:1707.04818*. [Online]. Available: <http://arxiv.org/abs/1707.04818>
- [221] K.-H. Zeng, W. B. Shen, D.-A. Huang, M. Sun, and J. C. Niebles, "Visual forecasting by imitating dynamics in natural sequences," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2999–3008.
- [222] P. Gujjar and R. Vaughan, "Classifying pedestrian actions in advance using predicted video of urban driving scenes," in *Proc. Int. Conf. Robot. Autom. (ICRA)*, May 2019, pp. 2097–2103.
- [223] H. Zhao and R. Wildes, "Spatiotemporal feature residual propagation for action prediction," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7003–7012.
- [224] A. Furnari and G. M. Farinella, "Egocentric action anticipation by disentangling encoding and inference," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 3357–3361.
- [225] A. Jain, A. Singh, H. S. Koppula, S. Soh, and A. Saxena, "Recurrent neural networks for driver activity anticipation via sensory-fusion architecture," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2016, pp. 3118–3125.
- [226] M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "Encouraging LSTMs to anticipate actions very early," in *Proc. ICCV*, Oct. 2017, pp. 280–289.
- [227] M. S. Aliakbarian, F. S. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "VIENA2: A driving anticipation dataset," in *Computer Vision—ACCV 2018*, C. V. Jawahar, H. Li, G. Mori, and K. Schindler, Eds. Cham, Switzerland: Springer, 2019, pp. 449–466.
- [228] R. De Geest and T. Tuytelaars, "Modeling temporal structure with LSTM for online action detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Mar. 2018, pp. 1549–1557.
- [229] P.-J. Wang, S.-F. Lien, and M.-S. Lee, "A learning-based prediction model for baby accidents," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 629–633.
- [230] V. Escorcia, F. C. Heilbron, J. C. Niebles, and B. Ghanem, "Daps: Deep action proposals for action understanding," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 768–784.
- [231] S. Buch, V. Escorcia, C. Shen, B. Ghanem, and J. C. Niebles, "SST: Single-stream temporal action proposals," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 2911–2920.
- [232] J. Gao, Z. Yang, K. Chen, C. Sun, and R. Nevatia, "TURN TAP: Temporal unit regression network for temporal action proposals," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3628–3636.
- [233] T. Lin, X. Liu, X. Li, E. Ding, and S. Wen, "BMN: Boundary-matching network for temporal action proposal generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3889–3898.
- [234] T. Wang, Y. Chen, Z. Lin, A. Zhu, Y. Li, H. Snoussi, and H. Wang, "RecapNet: Action proposal generation mimicking human cognitive process," *IEEE Trans. Cybern.*, early access, Jan. 28, 2020, doi: [10.1109/TCYB.2020.2965196](https://doi.org/10.1109/TCYB.2020.2965196).
- [235] Z. Shou, J. Chan, A. Zareian, K. Miyazawa, and S.-F. Chang, "CDC: Convolutional-de-convolutional networks for precise temporal action localization in untrimmed videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5734–5743.
- [236] L. Li, T. Kong, F. Sun, and H. Liu, "Deep point-wise prediction for action temporal proposal," in *Neural Information Processing*, T. Gedeon, K. W. Wong, and M. Lee, Eds. Cham, Switzerland: Springer, 2019, pp. 475–487.
- [237] T. Lin, X. Zhao, and Z. Shou, "Single shot temporal action detection," in *Proc. 25th ACM Int. Conf. Multimedia*. New York, NY, USA: ACM, Oct. 2017, pp. 988–996, doi: [10.1145/3123266.3123343](https://doi.org/10.1145/3123266.3123343).
- [238] B. G. S. Buch, V. Escorcia, and J. C. Niebles, "End-to-end, single-stream temporal action detection in untrimmed videos," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, G. B. T.-K. Kim, S. Zafeiriou, and K. Mikolajczyk, Eds. Newcastle, U.K.: BMVA Press, Sep. 2017, pp. 93.1–93.12, doi: [10.5244/C.31.93](https://doi.org/10.5244/C.31.93).
- [239] Y. Huang, Q. Dai, and Y. Lu, "Decoupling localization and classification in single shot temporal action detection," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2019, pp. 1288–1293.
- [240] F. Long, T. Yao, Z. Qiu, X. Tian, J. Luo, and T. Mei, "Gaussian temporal awareness networks for action localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 344–353.

- [241] M. A. Rahman and R. Laganière, "Single-stage end-to-end temporal activity detection in untrimmed videos," in *Proc. 17th Conf. Comput. Robot Vis. (CRV)*, May 2020, pp. 206–213.
- [242] C. Jin, T. Zhang, W. Kong, T. Li, and G. Li, "Regression before classification for temporal action detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 1–5, doi: [10.1109/ICASSP40776.2020.9053319](https://doi.org/10.1109/ICASSP40776.2020.9053319).
- [243] Y. Huang, Y. Sugano, and Y. Sato, "Improving action segmentation via graph-based temporal reasoning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14024–14034.
- [244] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 580–587.
- [245] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, Sep. 2015, pp. 1440–1448.
- [246] S. Ren, K. He, R. Girshick, and J. Sun, "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. Red Hook, NY, USA: Curran Associates, 2015, pp. 91–99.
- [247] Y. Ye, X. Yang, and Y. Tian, "Discovering spatio-temporal action tubes," *J. Vis. Commun. Image Represent.*, vol. 58, pp. 515–524, Jan. 2019. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S1047320318303468>
- [248] X. Peng and C. Schmid, "Multi-region two-stream R-CNN for action detection," in *Computer Vision—ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham, Switzerland: Springer, 2016, pp. 744–759.
- [249] Z. Yang, J. Gao, and R. Nevatia, "Spatio-temporal action detection with cascade proposal and location anticipation," 2017, *arXiv:1708.00042*. [Online]. Available: <http://arxiv.org/abs/1708.00042>
- [250] G. Singh, S. Saha, M. Sapienza, P. H. S. Torr, and F. Cuzzolin, "Online real-time multiple spatiotemporal action localisation and prediction," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3637–3646.
- [251] D. Li, Z. Qiu, Q. Dai, T. Yao, and T. Mei, "Recurrent tubelet proposal and recognition networks for action detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 303–318.
- [252] R. Hou, C. Chen, and M. Shah, "Tube convolutional neural network (T-CNN) for action detection in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 5822–5831.
- [253] V. Kalogeiton, P. Weinzaepfel, V. Ferrari, and C. Schmid, "Action tubelet detector for spatio-temporal action localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4405–4413.
- [254] X. Yang, X. Yang, M.-Y. Liu, F. Xiao, L. S. Davis, and J. Kautz, "STEP: Spatio-temporal progressive learning for video action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 264–272.
- [255] Y. Zhang, P. Tokmakov, M. Hebert, and C. Schmid, "A structured model for action detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 9975–9984.
- [256] R. Girdhar, J. Carreira, C. Doersch, and A. Zisserman, "Video action transformer network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Dec. 2019, pp. 244–253.
- [257] M. Tomei, L. Baraldi, S. Calderara, S. Bronzin, and R. Cucchiara, "Video action detection by learning graph-based spatio-temporal interactions," 2019, *arXiv:1912.04316*. [Online]. Available: <http://arxiv.org/abs/1912.04316>
- [258] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427–437, 2009. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S0306457309000259>
- [259] A. Tharwat, "Classification assessment methods," *Appl. Comput. Inform.*, vol. 17, no. 1, pp. 168–192, 2021. [Online]. Available: <http://www.sciencedirect.com/science/article/pii/S2210832718301546>
- [260] G. Tsoumakas and I. Katakis, "Multi-label classification: An overview," *Int. J. Data Warehousing Mining*, vol. 3, no. 3, pp. 1–13, 2007.
- [261] X.-Z. Wu and Z.-H. Zhou, "A unified view of multi-label performance measures," in *Proc. 34th Int. Conf. Mach. Learn.*, vol. 70, D. Precup and Y. W. Teh, Eds. Sydney, NSW, Australia: International Convention Centre, Aug. 2017, pp. 3780–3788. [Online]. Available: <http://proceedings.mlr.press/v70/wu17a.html>



MATTHEW S. HUTCHINSON received the B.S. and M.Eng. degrees in computer science and engineering from Massachusetts Institute of Technology, Cambridge, Massachusetts, USA, in 2020.

From 2019 to 2020, he was a Graduate Research Assistant with Massachusetts Institute of Technology (MIT) Lincoln Laboratory Supercomputing Center (LLSC). In 2020, he commissioned into U.S. Air Force (USAF) as a Developmental Engineer and has subsequently transferred into the U.S.

Space Force (USSF). His research interests include artificial intelligence, machine learning, and computer vision.



VIJAY N. GADEPALLY (Senior Member, IEEE) received the B.Tech. degree in electrical engineering from the Indian Institute of Technology, Kanpur, India, and the M.Sc. and Ph.D. degrees in electrical and computer engineering from The Ohio State University, Ohio, USA.

He is currently a Senior Member of the Technical Staff at Massachusetts Institute of Technology (MIT), Lincoln Laboratory, where he leads the research activities of the Lincoln Laboratory Supercomputing Center. Additionally, he works closely with the Computer Science and Artificial Intelligence Laboratory (CSAIL), MIT. His research interests include high performance computing, machine learning, artificial intelligence, and high-performance databases.

Dr. Gadepally's awards and honors, including an Outstanding Graduate Student Award at The Ohio State University, in 2011, and the MIT Lincoln Laboratory's Early Career Technical Achievement Award, in 2016. In 2017, he was named to AFCEA's inaugural 40 under 40 list.

...