HUMANE AI NET

Multimodal Perception with Transformers: Research Challenges and Data Sets

James L. Crowley
Grenoble Institut Polytechnique,
Univ Grenoble Alpes

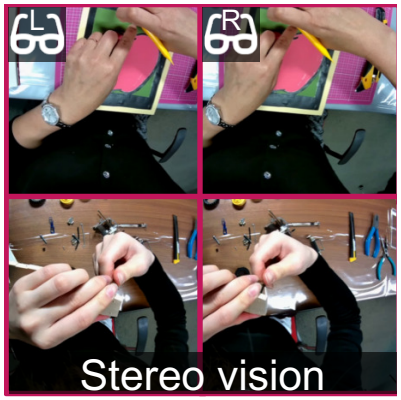
Research Challenges and Data Sets

For research on Multimodal Perception with Transformers

- EGO4D: Ego-Centric 4D Perception
- Ego-Centric Perception: Kitchen activities
  - EPIC-Kitchens 55 (2018)
  - EPIC-Kitchens 100 (2021)
- Visual Question and Answering (VQA)
- Vision and Language Navigation (VLN)
- Social-IQ

Egocentric 4D Perception (Ego4D)

(https://ego4d-data.org/docs/)



- A multimodal egocentric dataset and benchmark suite, with 3,600 hrs of densely narrated video and a wide range of annotations across five new benchmark tasks
- Scenarios of daily life captured in-the-wild by 926 camera wearers from 74 worldwide locations and 9 different countries
- Includes audio, 3D meshes of the environment, eye gaze, stereo, and synchronized videos from multiple egocentric cameras.

Egocentric 4D Perception (Ego4D)

(https://ego4d-data.org/docs/)

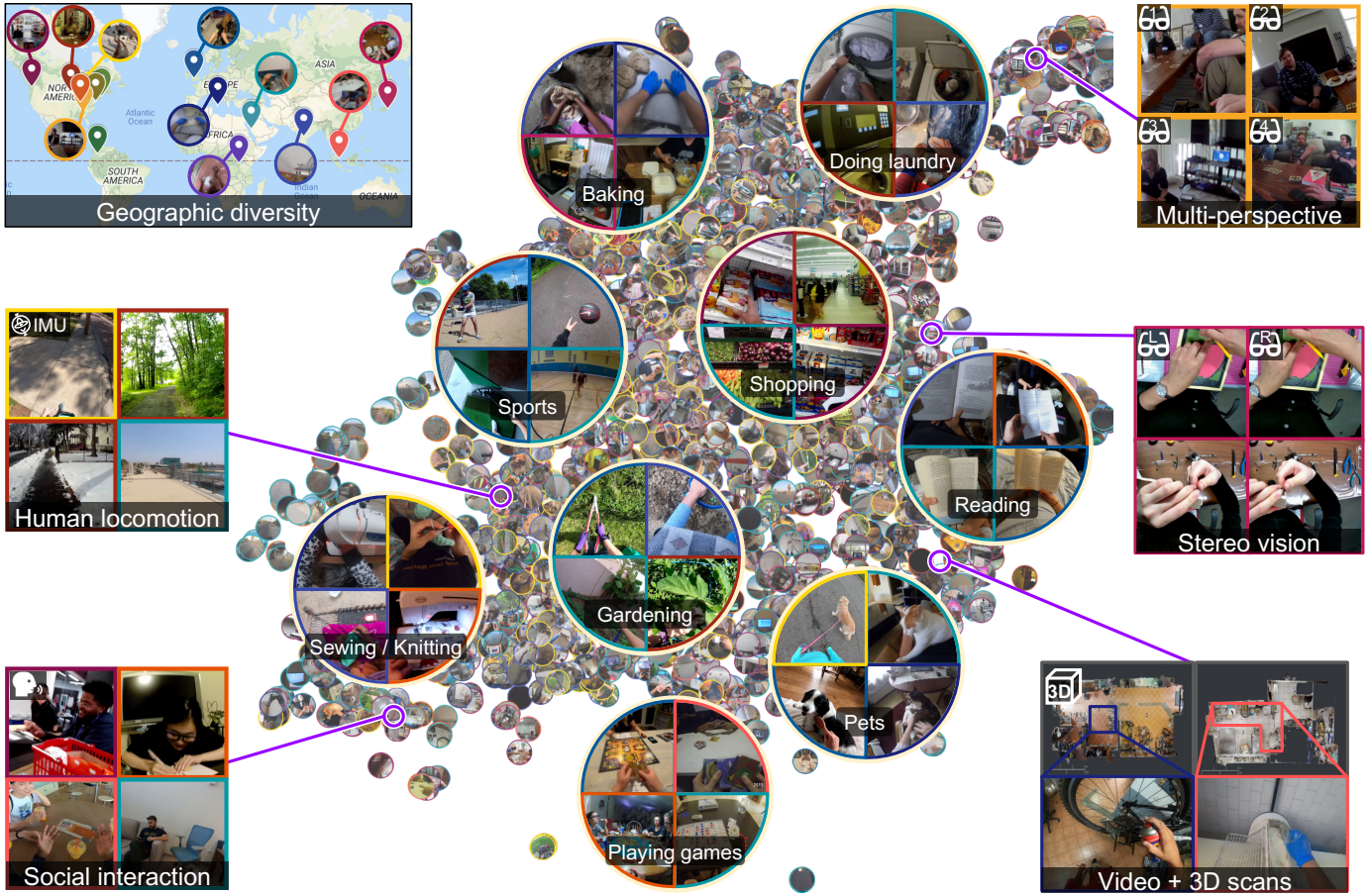


Figure 1. Ego4D is a massive-scale egocentric video dataset of daily life activity spanning 74 locations worldwide. Here we see a snapshot of the dataset (5% of the clips, randomly sampled) highlighting its diversity in geographic location, activities, and modalities. The data includes social videos where participants consented to remain unblurred. See https://ego4d-data.org/fig1.html for interactive figure.

The Ego4D Consortium

EGO4D Consortium

This initiative is led by an international consortium of 13 universities in partnership with Facebook AI, that collaborated to advance egocentric perception.

https://arxiv.org/abs/2110.07058

Computer Science > Computer Vision and Pattern Recognition

[Submitted on 13 Oct 2021 (v1), last revised 11 Mar 2022 (this version, v3)]

Ego4D: Around the World in 3,000 Hours of Egocentric Video

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

We introduce Ego4D, a massive-scale egocentric video dataset and benchmark suite. It offers 3,670 hours of daily-life activity video spanning hundreds of scenarios (household, outdoor, workplace, leisure, etc.) captured by 931 unique camera wearers from 74 worldwide locations and 9 different countries. The approach to collection is designed to uphold rigorous privacy and ethics standards with consenting participants and robust de-identification procedures where relevant. Ego4D dramatically expands the volume of diverse egocentric video footage publicly available to the research community. Portions of the video are accompanied by audio, 3D meshes of the environment, eye gaze, stereo, and/or synchronized videos from multiple egocentric cameras at the same event. Furthermore, we present a host of new benchmark challenges centered around understanding the first-person visual experience in the past (querying an episodic memory), present (analyzing hand-object manipulation, audio-visual conversation, and social interactions), and future (forecasting activities). By publicly sharing this massive annotated dataset and benchmark suite, we aim to push the frontier of first-person perception. Project page: this https URL

Comments: To appear in the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022. This version updates the baseline result numbers for the Hands and Objects benchmark (appendix)

Research Challenges: Episodic Memory

Episodic Memory

Where is my X?

Egocentric video gives a recording of a wearer's daily life, and can be used augment human memory on demand. Such a system might be able to remind a user where they left their keys, if they added salt to a recipe, or recall events they attended.

Querying Memory

There are three different tasks within this benchmark based on the input type used to query the memory: visual query (i.e. find the location given an image of keys), textual query ("how many cups of sugar did I add?"), and a moment query (find all instances of "When did I play with the dog").

Construction of Queries

For the language queries, a set of templates were designed which annotators used to write questions for the task. Examples include "what is the state of object X?" or "where is object X after event Y"? These were then re-written for variety.

Recalling Lives

Given the broad nature of this benchmark, there isn't a subset of activities that were focused on within this task, leading to a realistic and challenging benchmark.

Research Challenges: Episodic Memory

- **Visual queries with 2D localization and VQ 3D localization**: Given an egocentric video clip and an image crop depicting the query object, return the last time the object was seen in the input video, in terms of the tracked bounding box (2D + temporal localization) or the 3D displacement vector from the camera to the object in the environment.

- **Natural language queries**: Given a video clip and a query expressed in natural language, localize the temporal window within all the video history where the answer to the question is evident.

- **Moments queries**: Given an egocentric video and an activity name (e.g., a "moment"), localize all instances of that activity in the past video

Research Challenges: Hand + Object interaction

Hand + Object Interaction

How do objects change during interactions?

Going beyond Action Recognition, this benchmark follows when, where and how an object is changed during its interaction - only possible through a first person Viewpoint.

Changes of State

We capture annotations of objects, as they transform, temporally, spatially and semantically - an onion might be minced. These are represented by three different tasks in the benchmark: Point-of-no-return Temporal Localisation, Active Object Detection and State-Change Classification.

Pre/Post Conditions

Each annotation has been labelled with prior states (i.e. the prior condition) and posterior states as well as the point of no return (PNR) in which the state change is triggered.

World of Interactions

The data for this challenge has been selected from activities with a high level of hand-object interactions such as knitting, carpentry, and baking.

Research Challenges: Hand + Object interaction

- **Temporal localization and classification**: Given an egocentric video clip, localize temporally the key frames that indicate an object state change and identify what kind of state change it is.

- **State change object detection**: Given an egocentric video clip, identify the objects whose states are changing and outline them with bounding boxes.

Research Challenges: Audio-Visual Diarization

Audio-Visual Diarization

Who said what, and when?
Conversations are egocentric in nature, and a human-in-the-loop AI requires skills such as localizing a speaker and transcribing speech content

Looking for Conversation

This benchmark contains 2 different tasks focused on visual data: localizing and tracking of the speakers in the visual field of view. Note that identities are anonymized to match consortium guidelines.

Hearing the Words

The benchmark also includes 2 tasks for the audio modality: diaraization/temporal extent of the sentences spoken and the transcription of the conversation.

Much Ado About Talking

With this task focused on conversations, scenarios were chosen which included multiple participants interacting together, such as eating, playing games or setting up tents.

Research Challenges: Social Interaction

Social Interactions

Who is attending to whom?
An egocentric video provides a unique lens for studying social interactions because it captures utterances and nonverbal cues from each participant's unique view and enables embodied approaches to social understanding.

More than Conversation

Social extends the Audio-Visual Diaraization benchmark towards understanding the conversations of a social group over a longer period of time for specific tasks.

Talking and Listening

This benchmark includes two different tasks focused on when a person is *Looking at Me* and when a person is *Talking to Me*.

Unique Interactions

The data within the Social Interaction task was collected specifically for this task in mind with multi-user scenarios such as social deduction games, eating/drinking and playing basketball.

Audio-Visual Diarization and Social Interaction

- **Audio-visual localization**: Given an egocentric video clip, localize the speakers in the visual field of view.
- **Audio-visual speaker diarization**: Given an egocentric video clip, identify which person spoke and when they spoke.
- **Audio-only Diarization Challenges**
- **Speech transcription**: Given an egocentric video clip, transcribe the speech of each person.
- **Talking to me**: Given an egocentric video clip, identify whether someone in the scene is talking to the camera wearer.
- **Looking at me**: Given an egocentric video clip, identify whether someone in the scene is looking at the camera wearer.

Research Challenges: Forecasting

Forecasting

Predicting the future is a critical skill for AI systems to provide timely assistance for users. With a myriad of long-form, unscripted videos, Ego4D provides an interesting challenge for different forecasting tasks.

Where Will I Move?

Two tasks consider the future motion of the user with hands and feet. Models should predict where the camera wearer will go within the scene and the future location of wearer's hands.

What Will Happen Next?

Two tasks consider short and long term future anticipation. Algorithms should be able to predict the next object interaction that will take place and a countdown towards it taking place as well as the long term - what are the next possible sequence of actions?

Data for Prophets

The data for this challenge has been selected from a diverse set of activities containing many human-object interactions and movements such as brick making, cooking or carpentry.

Research Challenges: Forecasting

- **Locomotion forecasting**: Given a video frame and the past trajectory, predict the future ego positions of the camera wearer (in the form of a 3D trajectory).
- **Hand forecasting**: Given a short preceding video clip, predict where the hand will be visible in the future, in terms of a bounding box center in keyframes.
- **Short-term hand object prediction**: Given a video clip, predict the next active objects, the next action, and the time to contact.
- **Long-term activity prediction**: Given a video clip, the goal is to predict what sequence of activities will happen in the future? For example, after kneading dough, what will the baker do next?

Egocentric Perception of Non-scripted Daily activity

Egocentric Perception of Non-scripted Daily activity
Data Sets: Epic Kitchens https://epic-kitchens.github.io/2021

Key References

Damen, D., et al. (2018). Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 720-736). (Also appeared in PAMI 2020.

Damen, D., et al ., EPIC-KITCHENS-55 - 2020 Challenges Report, CVPR 2019.

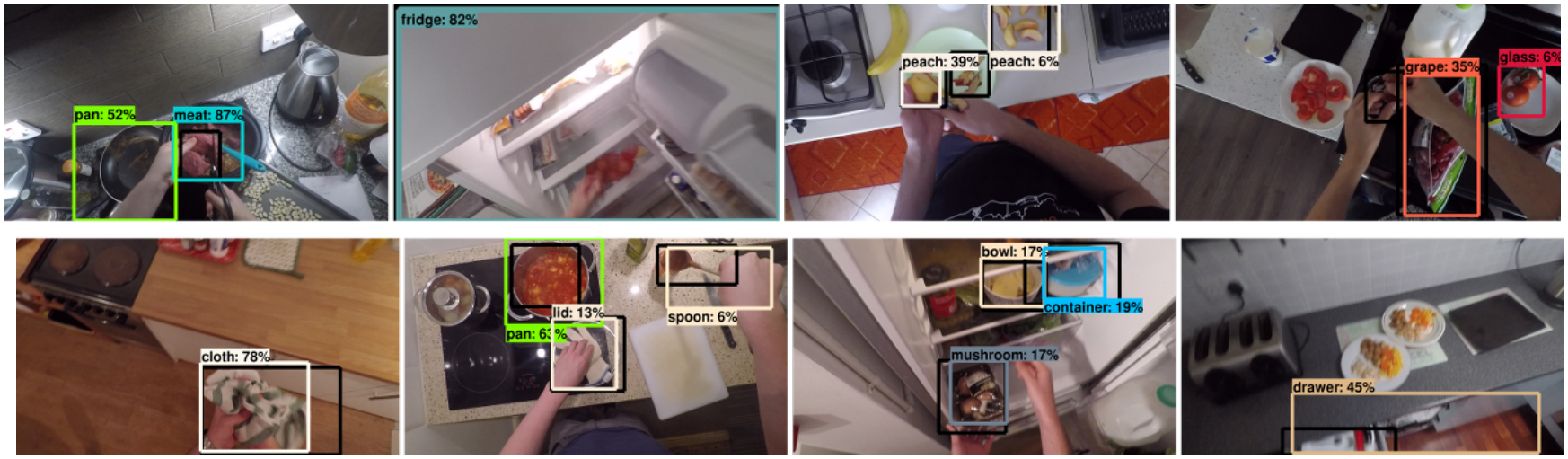
Damen, D., et al ., EPIC-KITCHENS-200 - Rescaling Egocentric Vision, 2021

EPIC: Egocentric Perception of Non-scripted Daily activity

EPIC Kitchens-55: a large-scale egocentric video benchmark recorded by 32 participants in their native kitchen environments. Videos depict **nonscripted** daily activities accompanied by Audio Narration. 55 hours of video (11.5M frames). Ground truth labeling for 39.6K action segments and 454.2K object bounding boxes. Narrations (speech and text) added post-recording by participants

Damen, D., et al. (2018). Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 720-736)

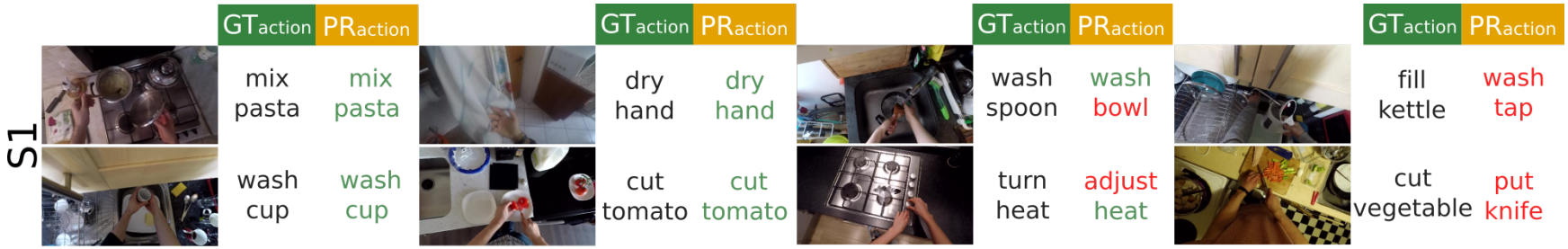
EPIC-55 Research Challenges: Object Detection Challenge



Object Detection: 125 Visual object classes and 331 Noun classes, grouped into grouped into 19 super categories

Evaluation Metrics: mean average precision (mAP) metric from PASCAL VOC, using IoU thresholds of 0.05, 0.5 and 0.75 similar to MS-COCO

EPIC-55 Research Challenges: Action Recognition Challenge



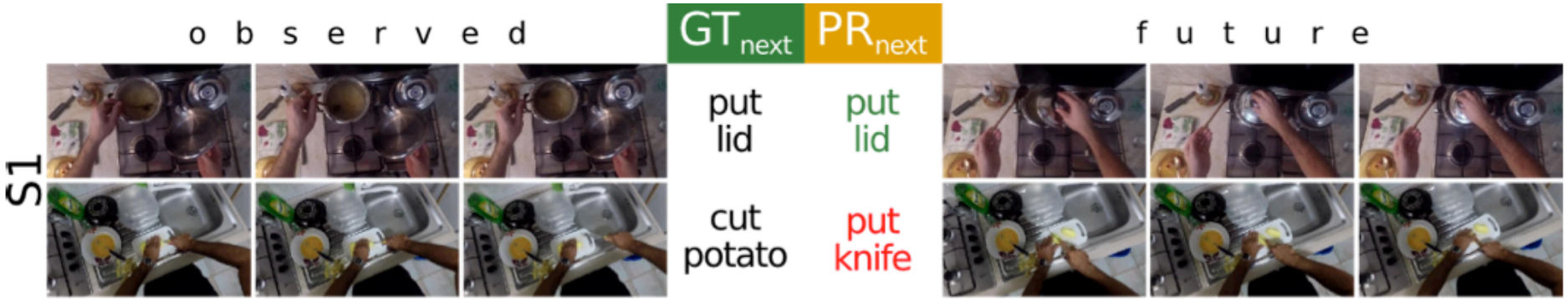
Action Recognition Challenge: Given an action segment, classify the segment into its action class, where classes are defined (verb, noun), with 26 verbs and 70 noun classes.

Evaluation Metrics:

(1) Aggregate metrics: top- 1 and top-5 accuracy for cv, cn and (cv,cn) – we refer to these as 'verb', 'noun' and 'action'.

(2) Per-class metric: precision and recall for classes with more than 100 samples in training

EPIC-55 Research Challenges: Action Anticipation Challenge



Action Anticipation Challenge: Given an action segment, predict the action class by observing the video segment *preceding* the action.

Evaluation Metrics:

(1) Aggregate metrics: top- 1 and top-5 accuracy for cv, cn and (cv,cn) – we refer to these as 'verb', 'noun' and 'action'.

(2) Per-class metric: precision and recall for classes with more than 100 samples in training

EPIC-55 Results: CVPR June 2019

D. Damen, E. Kazakos, W. Price, J. Ma, H. Doughty, A. Furnari, G. M. Farinella,
EPIC-KITCHENS-55- 2020 Challenges Report, at CVPR 2019, Los Angeles, June 2019

Object Detection Challenge:

	Rank	Team	Submissions		Few Shot Classes (%)			Many Shot Classes (%)			All Classes (%)		
			Entries	Date	IoU >0.05	IoU >0.5	IoU >0.75	IoU >0.05	IoU >0.5	IoU >0.75	IoU >0.05	IoU >0.5 ▲	IoU >0.75
S1	1	**hutom**	51	05/30/20	**47.44**	**35.75**	**14.32**	**60.77**	**46.50**	**15.60**	**58.27**	**44.48**	**15.36**
S1	2	**DHARI**	27	05/29/20	**54.98**	**32.40**	**14.55**	**68.74**	**43.88**	**15.38**	**66.15**	**41.72**	**15.23**
S1	3	**FB AI**	69	04/01/20	**26.55**	**19.01**	**8.22**	**58.44**	**46.22**	**15.61**	**52.44**	**41.10**	**14.22**
S1	4	CVG Lab Uni Bonn	23	05/12/20	39.36	26.66	7.89	53.50	41.28	12.46	50.84	38.53	11.60
S1	5	VCL	61	05/18/20	33.23	23.16	5.00	50.78	37.91	9.79	47.48	35.13	8.89
S1	6	[2] (baseline)	-	09/03/18	30.63	20.28	2.75	49.55	37.39	9.82	45.99	34.18	8.49

Action Recognition Challenge:

	Rank	Team	Submissions		Top-1 Accuracy			Top-5 Accuracy			Avg Class Precision			Avg Class Recall		
			Entries	Date	VERB	NOUN	ACTION ▲	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION
	1	**UTS-Baidu**	14	05/28/20	**70.41**	**52.85**	**42.57**	**90.78**	**76.62**	**63.55**	**60.44**	**47.11**	**24.94**	**45.82**	**50.02**	**26.93**
	2	**NUS-CVML**	18	05/29/20	**63.23**	**46.45**	**41.59**	**87.50**	**70.49**	**64.11**	**51.54**	**42.09**	**25.37**	**40.99**	**42.69**	**26.98**
		UTS-Baidu	16	05/30/19	**69.80**	**52.27**	**41.37**	**90.95**	**76.71**	**63.59**	**63.55**	**46.86**	**25.13**	**46.94**	**49.17**	**26.39**
	3	**SAIC-Cambridge**	34	05/27/20	**69.43**	**49.71**	**40.00**	**91.23**	**73.18**	**60.53**	**60.01**	**45.74**	**24.95**	**47.40**	**46.78**	**25.27**
S1	3	**FBK-HuPBA**	50	05/29/20	**68.68**	**49.35**	**40.00**	**90.97**	**72.45**	**60.23**	**60.63**	**45.45**	**21.82**	**47.19**	**45.84**	**24.34**
	4	GT-WISC-MPI	12	01/30/20	68.51	49.96	38.75	89.33	72.30	58.99	51.04	44.00	23.70	43.70	47.32	23.92
	5	G-Blend	14	05/28/20	66.67	48.48	37.12	88.90	71.36	56.21	51.86	41.26	20.97	44.33	44.92	21.48

Action Anticipation Challenge

	Rank	Team	Submissions		Top-1 Accuracy			Top-5 Accuracy			Avg Class Precision			Avg Class Recall		
			Entries	Date	VERB	NOUN	ACTION ▲	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION
S1	1	**NUS_CVML**	18	05/29/20	**37.87**	**24.10**	**16.64**	**79.74**	**53.98**	**36.06**	**36.41**	**25.20**	**9.64**	**15.67**	**22.01**	**10.05**
S1	2	**VI-I2R**	28	05/23/20	**36.72**	**24.61**	**16.02**	**80.39**	**54.90**	**37.11**	**31.03**	**26.02**	**8.68**	**15.28**	**22.03**	**8.70**
S1	3	**Ego-OMG**	16	05/26/20	**32.20**	**24.90**	**16.02**	**77.42**	**50.24**	**34.53**	**14.92**	**23.25**	**4.03**	**15.48**	**19.16**	**5.36**
S1	4	UNIPD-UNICT	16	05/26/20	36.73	24.26	15.67	79.87	53.76	36.31	35.86	25.16	7.42	14.12	21.30	7.62
S1	5	GT-WISC-MPI	20	11/12/19	36.25	23.83	15.42	79.15	51.98	34.29	24.90	24.03	6.93	15.31	21.91	7.88

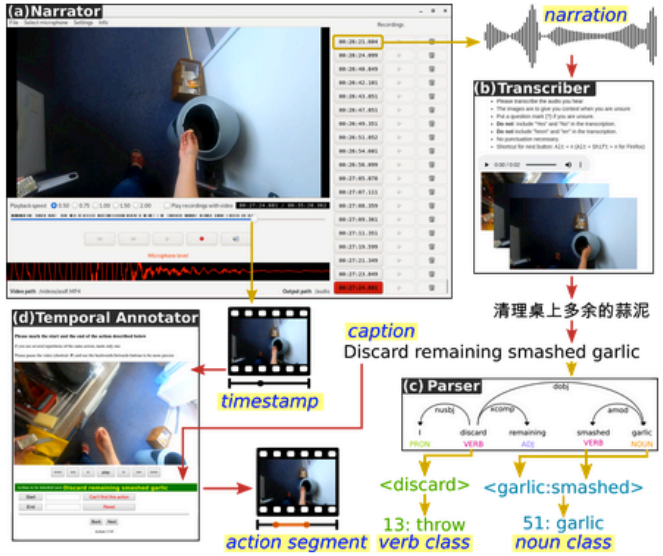
EPIC Kitchens-100

EPIC Kitchens-100: 100 hours, 20M frames, 90K actions in 700 variable-length videos, capturing long-term unscripted activities in 45 environments, using head-mounted cameras. Annotated with denser and more complete annotations of fine-grained actions (54% more actions per minute, +128% more action segments)

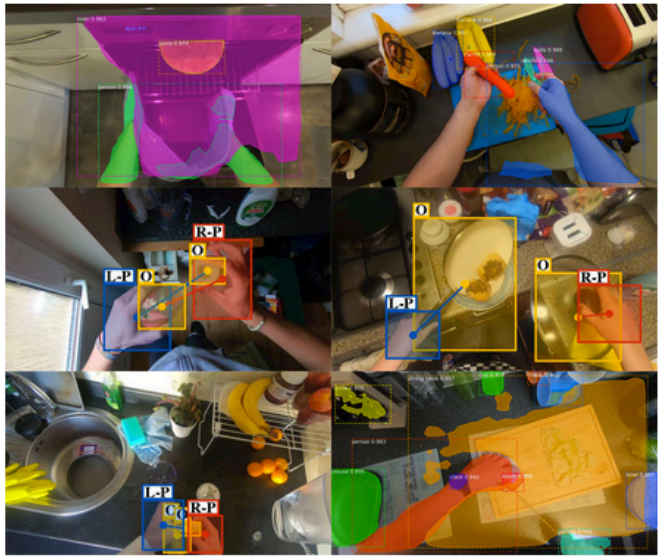
Ground truth labeling for 39.6K action segments and 454.2K object bounding boxes. Narrations (speech and text) added post-recording by participants

Damen, D., et al. (2021). ReScaling egocentric vision: The epic-kitchens dataset. IJCV 2021

EPIC Kitchens-100 Data Collection



Annotation Pipeline



Automatic Annotations

45 participants in 4 cities collected video over 2 to 4 days using GoPro Hero7 black.

Videos are narrated off-line in native language using “Pause and talk” to provide synchronized audio-visual recording

Narratives are translated English with Amazon Mechanical Turk, spell checked and transformed to verbs/nouns

https://epic-kitchens.github.io/2021

EPIC-Kitchens-100: Five research challenges

Five research challenges

1) Action Recognition
2) Action Detection
3) Action Anticipation
4) Cross-modal retrieval
5) Domain adaptation

EPIC-100 Research Challenges: Action Recognition Challenge

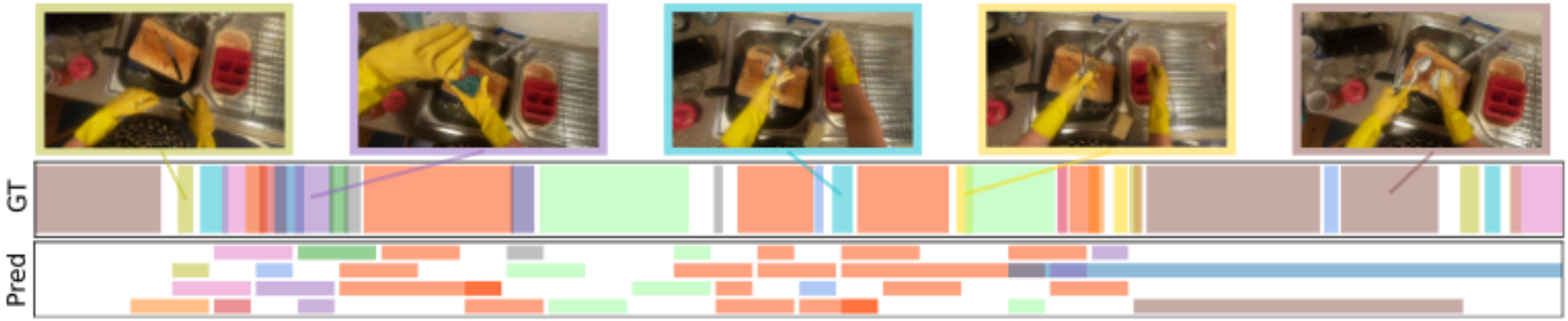
Action Recognition Challenge: Given an action segment, classify the segment into its action class. Data contains 53 action classes with 128 instances

Evaluation Metrics:

(1) Aggregate metrics: top- 1 and top-5 accuracy for cv, cn and (cv,cn) – we refer to these as 'verb', 'noun' and 'action'.

(2) Per-class metric: precision and recall for classes with more than 100 samples in training

EPIC-100: Action Detection Challenge



Action Detection Challenge

Action Detection: Given a video, detect Action instances with Start Time, Stop time, verb, noun and action class.

Data: 100 hours of audio-video recording, 4053 action classes, 89977 action instances, average 128.5 actions/video and 53.2 classes/video, 28% overlap

Evaluation Metrics: mean average precision (mAP) metric. Temporal segments are matched with Intersection over Union from 0.1 to 0.5

EPIC-100: Action Anticipation Challenge

Action Anticipation Challenge: Given an action segment, predict the (Verb, Noun, Action) classes by observing a segment preceding the action segment by 1 second.

Evaluation Metrics:

(1) Aggregate metrics: top- 1 and top-5 accuracy for (Verb, Noun, Action) classes

(2) Per-class metric: precision and recall for (Verb, Noun, Action) classes

EPIC-100: Cross-Modal Action Retrieval Challenge

Cross-Modal Action Retrieval Challenge: Given an query segment, rank segments in a gallery set that are semantically relevant

Text to video: Query is text caption, gallery contains videos

Video to text: Query is video: gallery contains text captions.

Evaluation Metrics:

(1) Normalized Discounted Cumulative Gain (nDCG). Given query x_r, and a gallery C_r

$$nDCG(x_i, C_r) = \frac{DCG(x_i, C_r)}{IDCG(x_i, C_r)}$$

Where:

$$DCG(x_i, C_r) = \sum_{j=1}^{|C_r|} \frac{\mathcal{R}(x_i, c_j)}{log(j+1)}$$

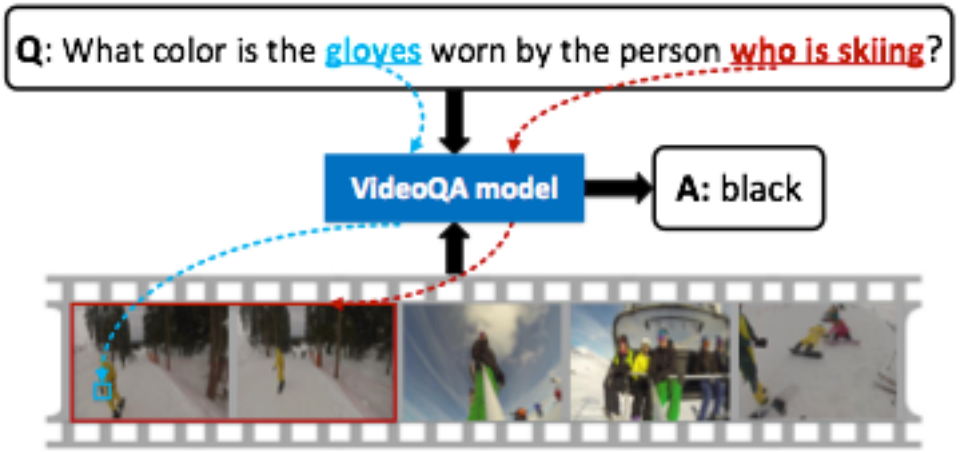
$$IDCG(x_i, C_r) = DCG(x_i, \hat{C}_r))$$

EPIC-100: Domain Adaptation Challenge

Unsupervised Domain Adaptation Challenge: Given a labeled source domain (kitchen) from 2018 learn to adapt to an unlabeled target domain from 2020. Source and Targets are from the 16 participants who provided recordings from both 2018 and 2020.

Evaluation Metrics: Same as with action recognition - Given an action segment, classify the segment into its action class, where classes are defined (verb, noun), with 26 verbs and 70 noun classes.

Visual Question and Answering



VisualQA Problem: Generate natural language answer to a question about a video

Image from Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., and Tao, D. (2019, July). Activitynet-QA: A dataset for understanding complex web videos via question answering. AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 9127-9134).

VQA Datasets:

Datasets	Video source	QA pairs generation	QA tasks	# Videos	# QA pairs	Average video length
MSVD-QA (Xu et al. 2017)	MSVD	Automatic	OE	1,970	50,505	10s
MSRVTT-QA (Xu et al. 2017)	MSRVTT	Automatic	OE	10,000	243,680	15s
TGIF-QA (Jang et al. 2017)	TGIF	Automatic & Human	OE & MC	56,720	103,919	3s
MovieQA (Tapaswi et al. 2016)	Movies	Human	MC	6,771	6,462	200s
Video-QA (Zeng et al. 2017)	Jukinmedia	Automatic	OE	18,100	174,775	45s
ActivityNet-QA (Ours)	ActivityNet	Human	OE	5,800	58,000	180s

VisualQA Problem: Generate natural language answer to a question about a video
As the videos are collected, Question-Answer Pairs are generated for each video.

Most data sets exploit narrative descriptions or captions provided with the video.
Activity net uses crowdsourcing to generate QA pairs.

Table from Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., and Tao, D. (2019, July). Activitynet-QA: A dataset for understanding complex web videos via question answering. AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 9127-9134).

HowTo100M: 100 Million Narrated Video Clips

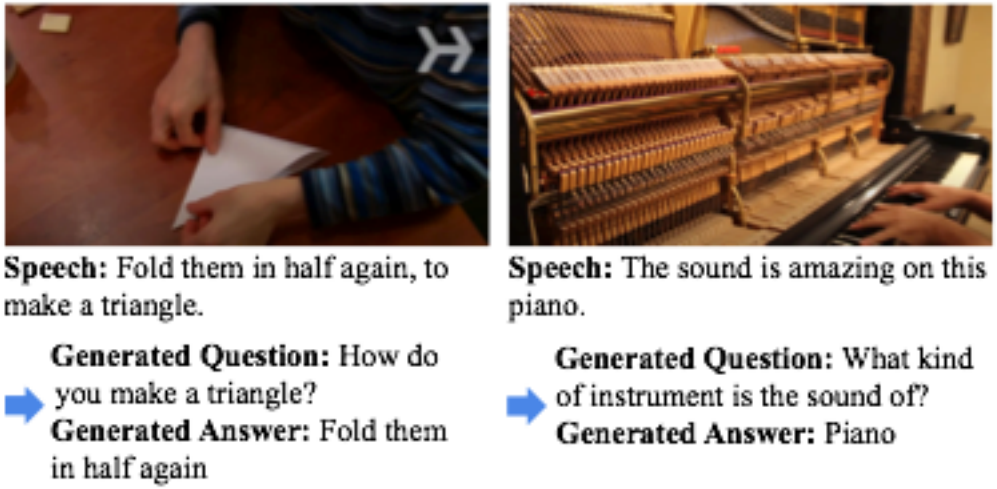
Dataset of narrated instructional videos where content creators teach complex tasks with an explicit intention of explaining the visual content on screen.

Includes 136M video clips with captions sourced from 1.2M Youtube videos (15 years of video) showing **23k activities** from domains such as cooking, hand crafting, personal care, gardening or fitness. Each video is associated with a narration available as subtitles automatically downloaded from Youtube.

Challenges: text based action localization and text-to-video retrieval

Miech, A., Zhukov, D., Alayrac, J. B., Tapaswi, M., Laptev, I., and Sivic, J. (2019). Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. *IEEE International Conference on Computer Vision, CVPR 2019,* pp. 2630-2640.

HowToVQA69M: Question-answer triplets for HowTo100M



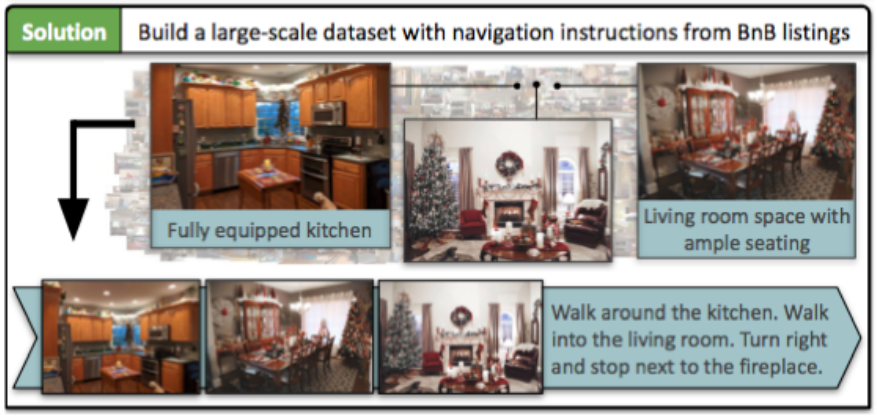
A large dataset with 69 Million video-question-answer triplets generated using transformers to automatically generate questions for videos in HowTo100M.

Approach: Use transformers trained on a question-answering text to generate a non-scripted questions and corresponding open-vocabulary answers from text using the HowTo100M data set.

Challenge: Given a video and a question, Generate a natural language answer.

Yang, A., Miech, A., Sivic, J., Laptev, I. and Schmid, C., 2021. Just ask: Learning to answer questions from millions of narrated videos. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1686-1697).

Vision and Language Navigation



Task: Enable a Robot to navigate in realistic environments using natural language instructions.

Dataset: BnB: image-caption (IC) pairs from listings from online rental marketplace, with 1.4M indoor images and 0.7M captions. Static image-caption pairsare transformed into visual paths and navigation-like instructions

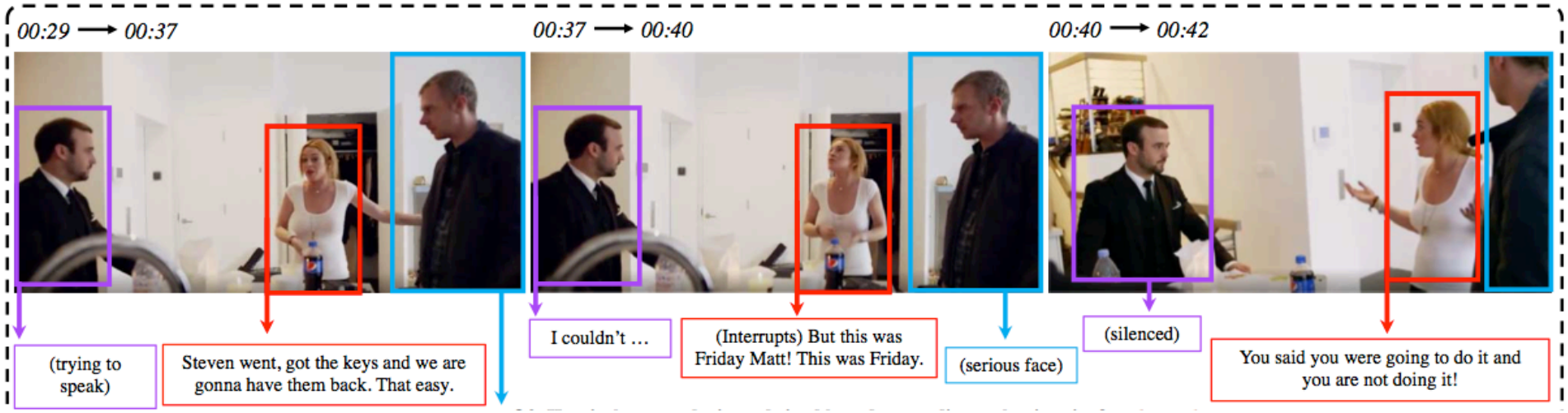
Challenges:

Path Discrimination. Choose the base path from a set of candidates

Path Generation: sequentially predict actions

Guhur, P.L., Tapaswi, M., Chen, S., Laptev, I. and Schmid, C., 2021. Airbert: In-domain Pretraining for Vision-and-Language Navigation. In IEEEF International Conference on Computer Vision, ICCV2021, pp. 1634-1643, Oct 2021

Social-IQ



DataSet: 1,250 natural in-the-wild Annotated videos, with 7, 500 questions and 52, 500 correct and incorrect answers, in 3 classes: (easy, intermediate, advanced)

Challenge: generate answer for question from video

Example:

Q1: How is the discussion between the woman and the man in the white shirt ?

A3: They are having a romantic conversation. <easy>

Zadeh, A., Chan, M., Liang, P.P., Tong, E. and Morency, L.P., Social-IQ: A question answering benchmark for artificial social intelligence. *IEEE Conference on Computer Vision and Pattern Recognition, CVPR2019,* pp. 8807-8817, June 2019

https://github.com/A2Zadeh/Social-IQ