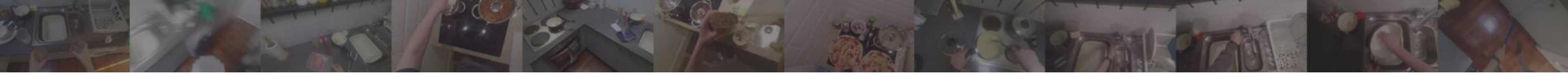
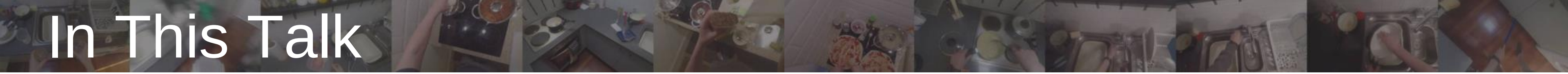


Egocentric Video Understanding

Toby Perrett

In This Talk



- A (very) brief history of video understanding
- Why is egocentric vision important?
- EPIC Kitchens
- EPIC Kitchens VISOR
- EPIC Kitchens FIELDS
- Why we need JADE

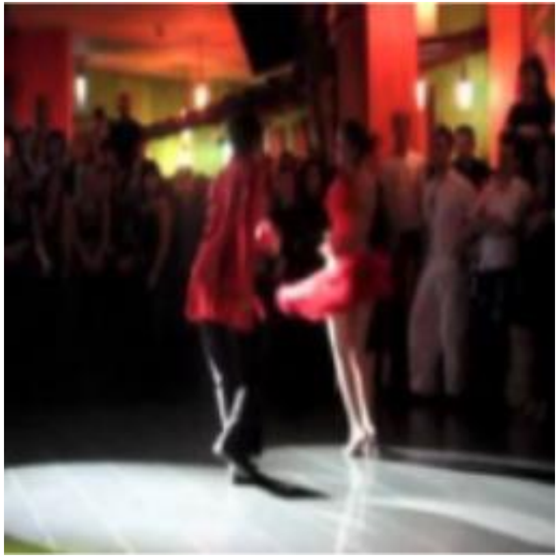
A (very) brief history of video understanding

Video datasets

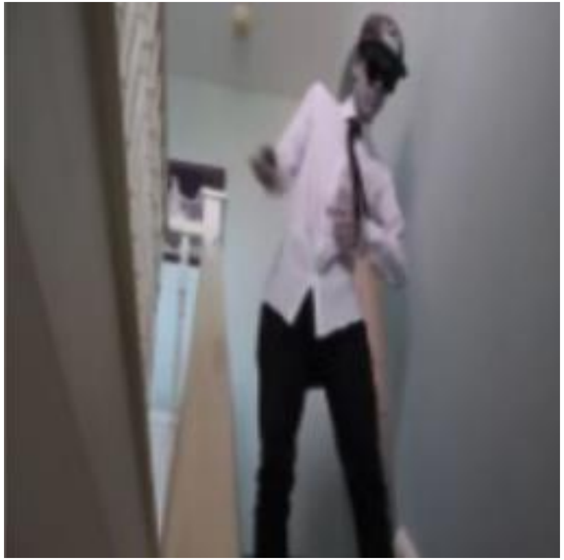
- Models haven’t improved that much
- Capabilities driven by data
- Traditionally 3rd person
- Scraped from Youtube

Can you guess?

- Robot dancing
- Dancing ballet
- Salsa dancing
- Mosh pit dancing
- Tap dancing
- Breakdancing



?



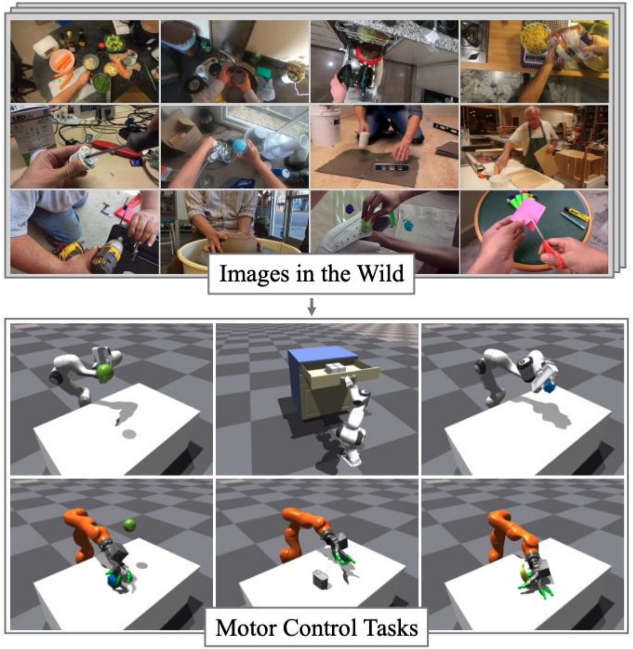
?

Why Egocentric?

How animals see and learn from the world

Assistive applications

Robotics



Solve egocentric first -> everything else is easy!

Why Egocentric?

Q: How do we encourage progress in egocentric video understanding?

A: Provide data, models and challenges

EPIC Kitchens

2017: “We’re going to create the largest egocentric dataset”

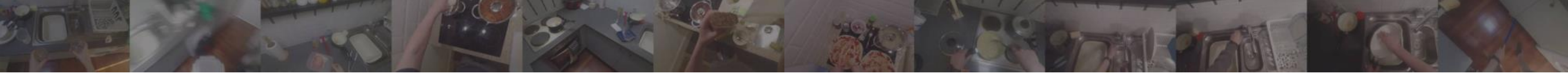
Why Kitchens?

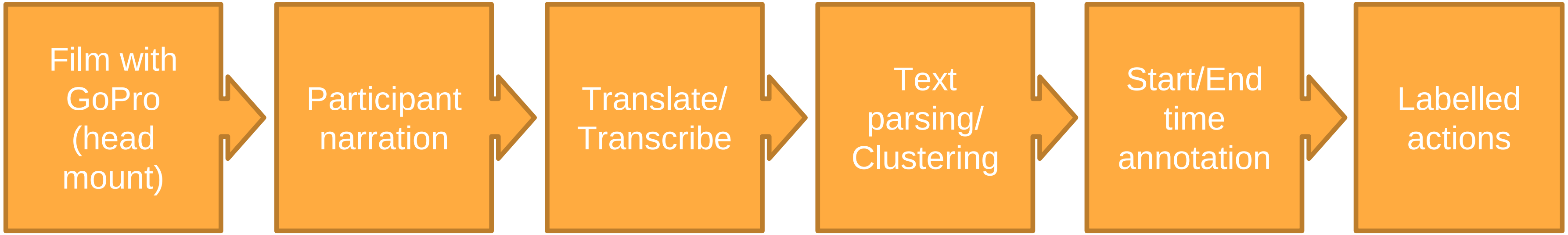
- Lots of actions to understand
- Cultural variation
- Familiar environment
- Messy

D Damen et al. Scaling Egocentric Vision: The EPIC-Kitchens Dataset, ECCV 2018, TPAMI 2020, IJCV 2021

EPIC Kitchens

D Damen et al. Scaling Egocentric Vision: The EPIC-Kitchens Dataset, ECCV 2018, TPAMI 2020, IJCV 2021





Film with GoPro (head mount)

Participant narration

Translate/ Transcribe

Text parsing/ Clustering

Start/End time annotation

Labelled actions

amazon mechanical turk

Put the pot on the counter

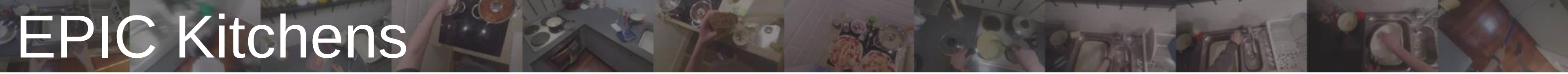
amazon mechanical turk

Verb: put
Noun: pot
Start: 10.5
End: 12.9

D Damen et al. Scaling Egocentric Vision: The EPIC-Kitchens Dataset, ECCV 2018, TPAMI 2020, IJCV 2021

EPIC Kitchens

D Damen et al. Scaling Egocentric Vision: The EPIC-Kitchens Dataset, ECCV 2018, TPAMI 2020, IJCV 2021



EPIC Kitchens

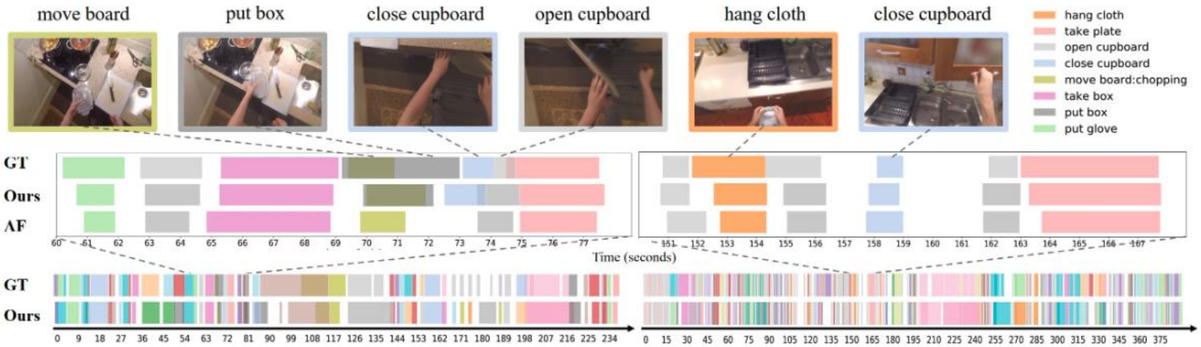
- 45 kitchens
- 100 hours
- 90k action instances
- 97 verb classes
- 300 noun classes
- Standard benchmark for video understanding (1000’s of citations)

D Damen et al. Scaling Egocentric Vision: The EPIC-Kitchens Dataset, ECCV 2018, TPAMI 2020, IJCV 2021

EPIC Kitchens

- 6 challenges
- Run twice a year

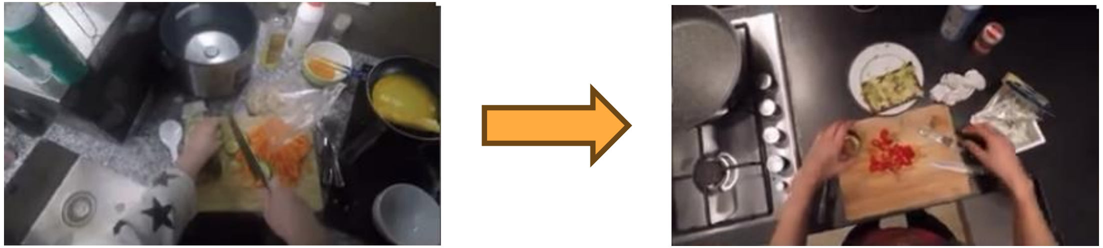
Detection



Retrieval

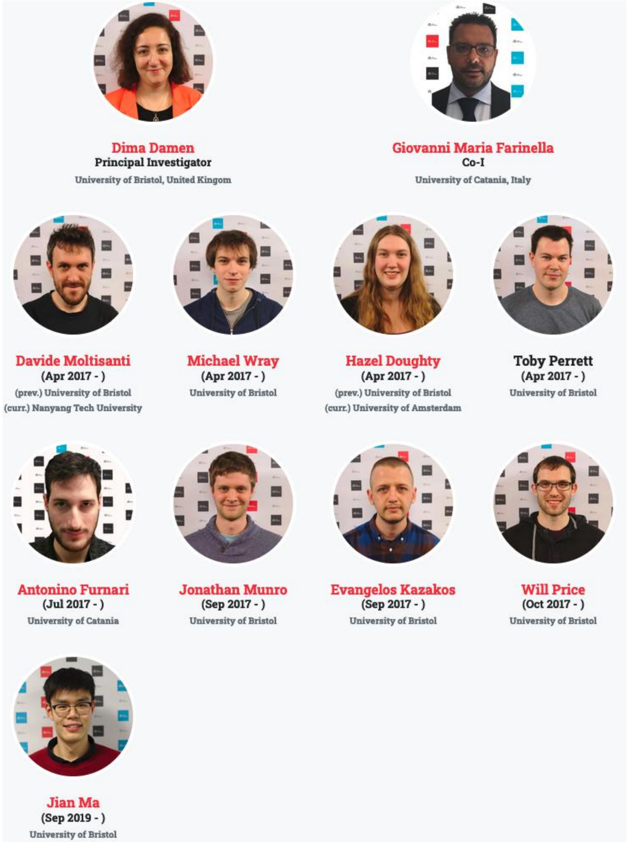
Wash

Adaptation



D Damen et al. Scaling Egocentric Vision: The EPIC-Kitchens Dataset, ECCV 2018, TPAMI 2020, IJCV 2021

EPIC Kitchens



Dima Damen
Principal Investigator
University of Bristol, United Kingom

Giovanni Maria Farinella
Co-I
University of Catania, Italy

Davide Moltisanti
(Apr 2017 -)
(prev.) University of Bristol
(curr.) Nanyang Tech University

Michael Wray
(Apr 2017 -)
University of Bristol

Hazel Doughty
(Apr 2017 -)
(prev.) University of Bristol
(curr.) University of Amsterdam

Toby Perrett
(Apr 2017 -)
University of Bristol

Antonino Furnari
(Jul 2017 -)
University of Catania

Jonathan Munro
(Sep 2017 -)
University of Bristol

Evangelos Kazakos
(Sep 2017 -)
University of Bristol

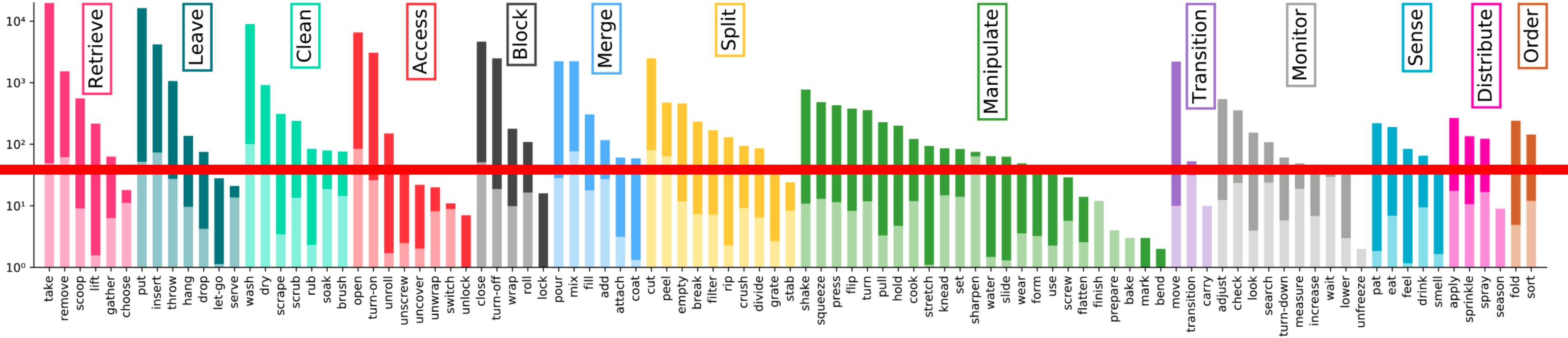
Will Price
(Oct 2017 -)
University of Bristol

Jian Ma
(Sep 2019 -)
University of Bristol

D Damen et al. Scaling Egocentric Vision: The EPIC-Kitchens Dataset, ECCV 2018, TPAMI 2020, IJCV 2021

Long-Tail

- Solve egocentric first

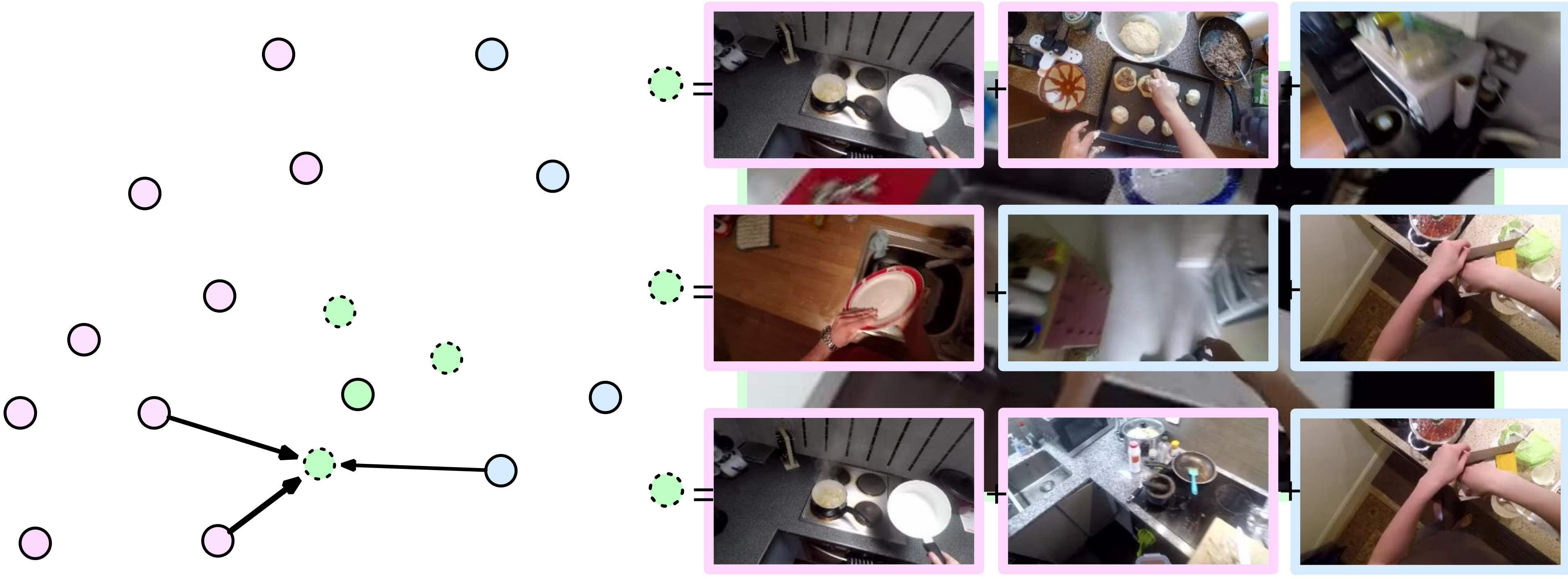


- Models fail on few-shot classes
- Previous video datasets have ignored these cases
- Interest from other fields

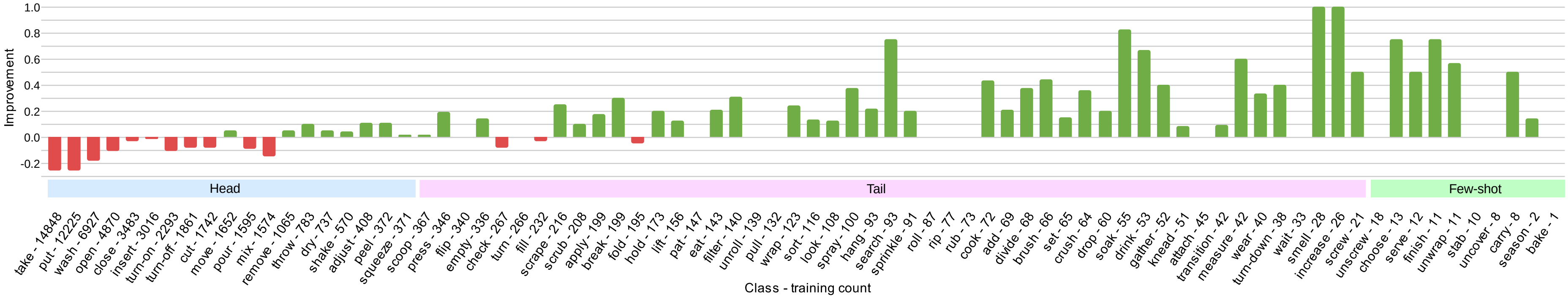
T Perrett et al. Use Your Head: Improving Long-Tail Video Recognition, CVPR 2023

Long-Tail

- We reconstruct few-shot samples from multiple visually similar head samples
- Expands the class boundaries for few-shot classes

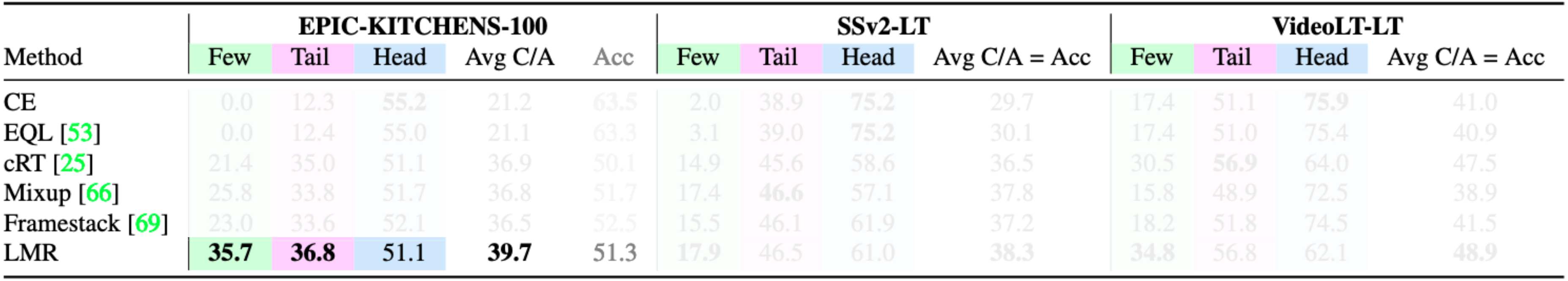


Long-Tail



T Perrett et al. Use Your Head: Improving Long-Tail Video Recognition, CVPR 2023

Long-Tail

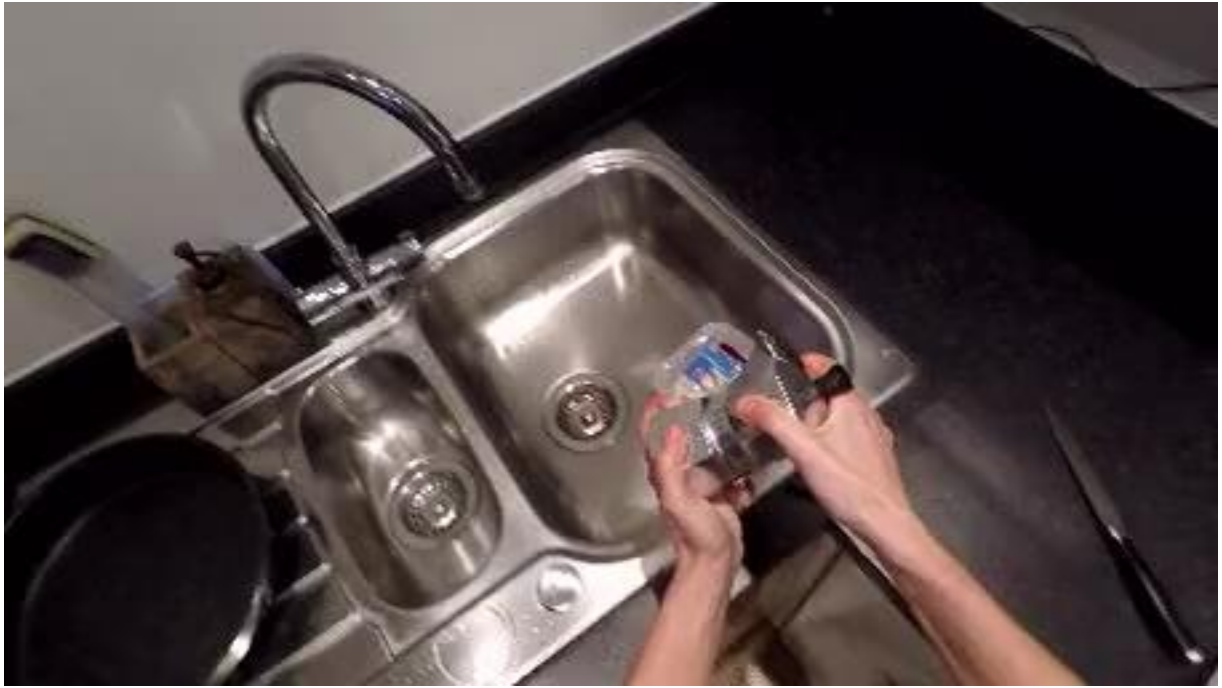


Method	EPIC-KITCHENS-100					SSv2-LT				VideoLT-LT			
	Few	Tail	Head	Avg C/A	Acc	Few	Tail	Head	Avg C/A = Acc	Few	Tail	Head	Avg C/A = Acc
CE	0.0	12.3	**55.2**	21.2	**63.5**	2.0	38.9	**75.2**	29.7	17.4	51.1	**75.9**	41.0
EQL [53]	0.0	12.4	55.0	21.1	63.3	3.1	39.0	**75.2**	30.1	17.4	51.0	75.4	40.9
cRT [25]	21.4	35.0	51.1	36.9	50.1	14.9	45.6	58.6	36.5	30.5	**56.9**	64.0	47.5
Mixup [66]	25.8	33.8	51.7	36.8	51.7	17.4	**46.6**	57.1	37.8	15.8	48.9	72.5	38.9
Framestack [69]	23.0	33.6	52.1	36.5	52.5	15.5	46.1	61.9	37.2	18.2	51.8	74.5	41.5
LMR	**35.7**	**36.8**	51.1	**39.7**	51.3	**17.9**	46.5	61.0	**38.3**	**34.8**	56.8	62.1	**48.9**

- Video architecture: Motionformer, 2021, Facebook/University of Oxford
- 1 run: 1 day on 1 JADE node (8 GPUs)
- Smallest model we could get away with

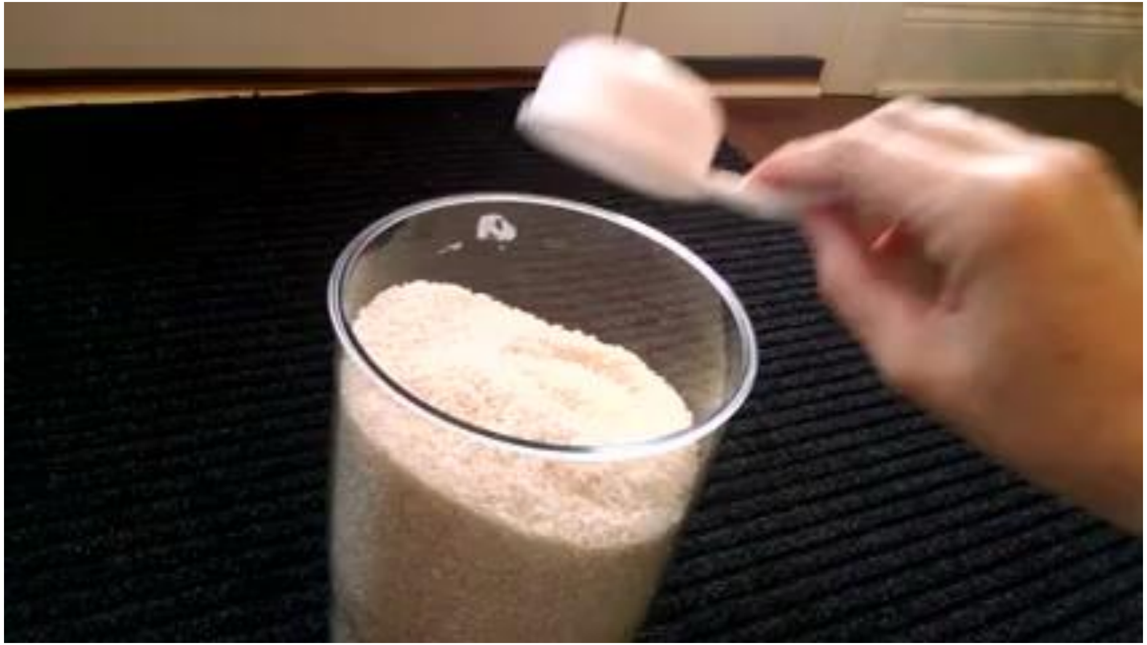
T Perrett et al. Use Your Head: Improving Long-Tail Video Recognition, CVPR 2023

Long-Tail

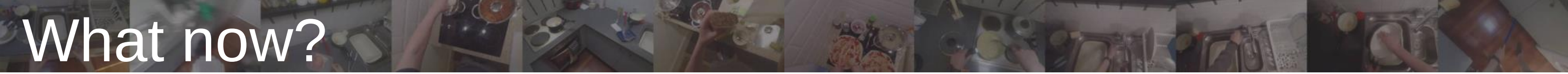


Ground truth:	21	Screw	
Standard training:	6927	Wash	✗
Ours:	21	Screw	✓

Long-Tail



Ground Truth:	8	Pretending to scoop something up with something	
Standard Training:	883	Taking something out of something	✗
Ours:	8	Pretending to scoop something up with something	✓



What now?

Now we have

- Lots of video
- Methods can do video -> verbs/nouns
- Short-term only

What's missing?

- Segmentations – what’s in the video and where is it?
- State changes
- Long term
- How do we interact?
- What about 3D?

Coming up: two EPIC-Kitchens extensions

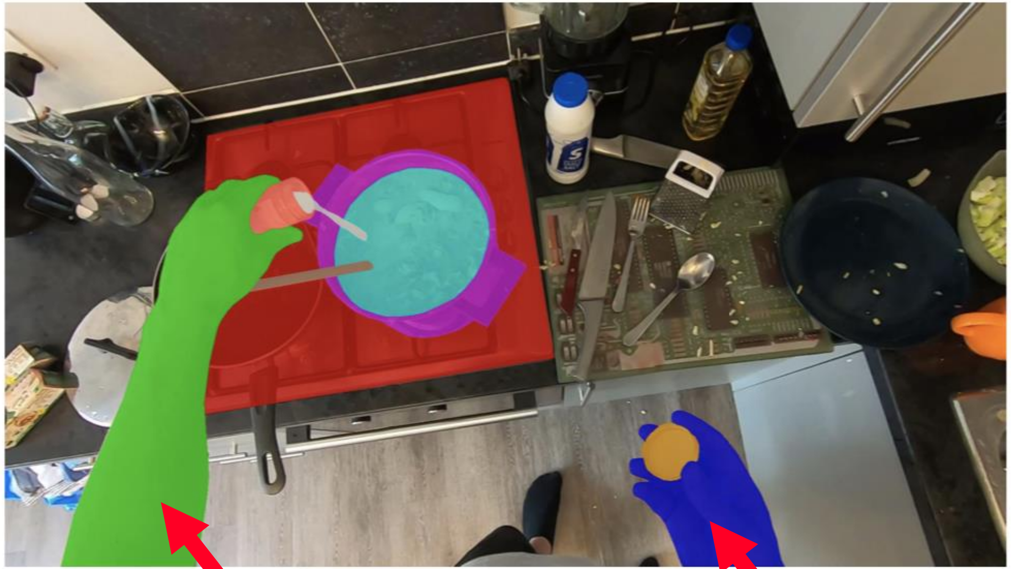
EPIC-KITCHENS VISOR

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

VISOR is….

pour spice

action



🟩 left hand

🟦 right hand

🟥 hob

🟪 saucepan

⬜ spice

🟧 spice container

🟫 spoon

🟦 soup

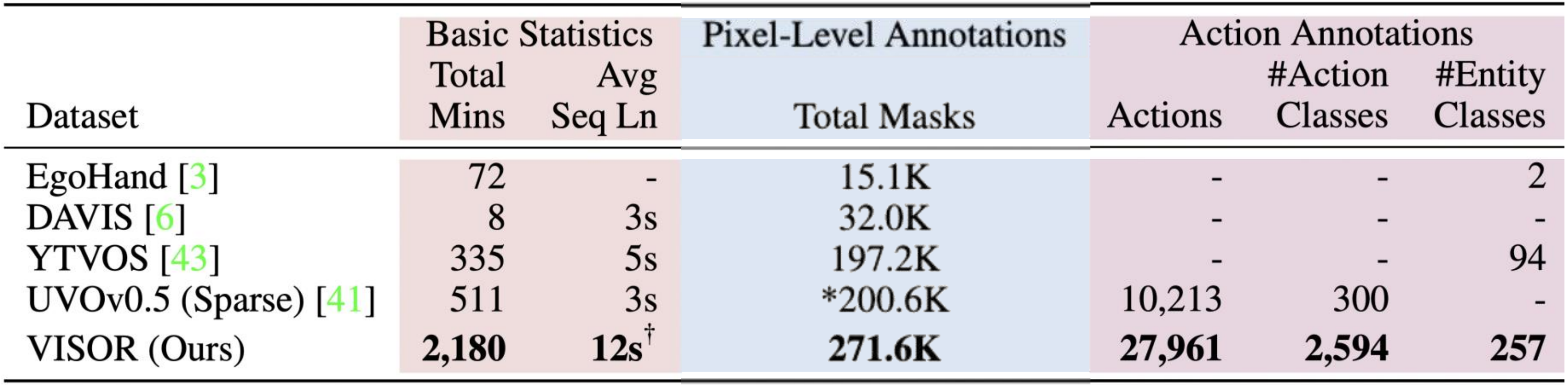
🟨 pepper container lid

in-contact (spice container)

in-contact (container lid)

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

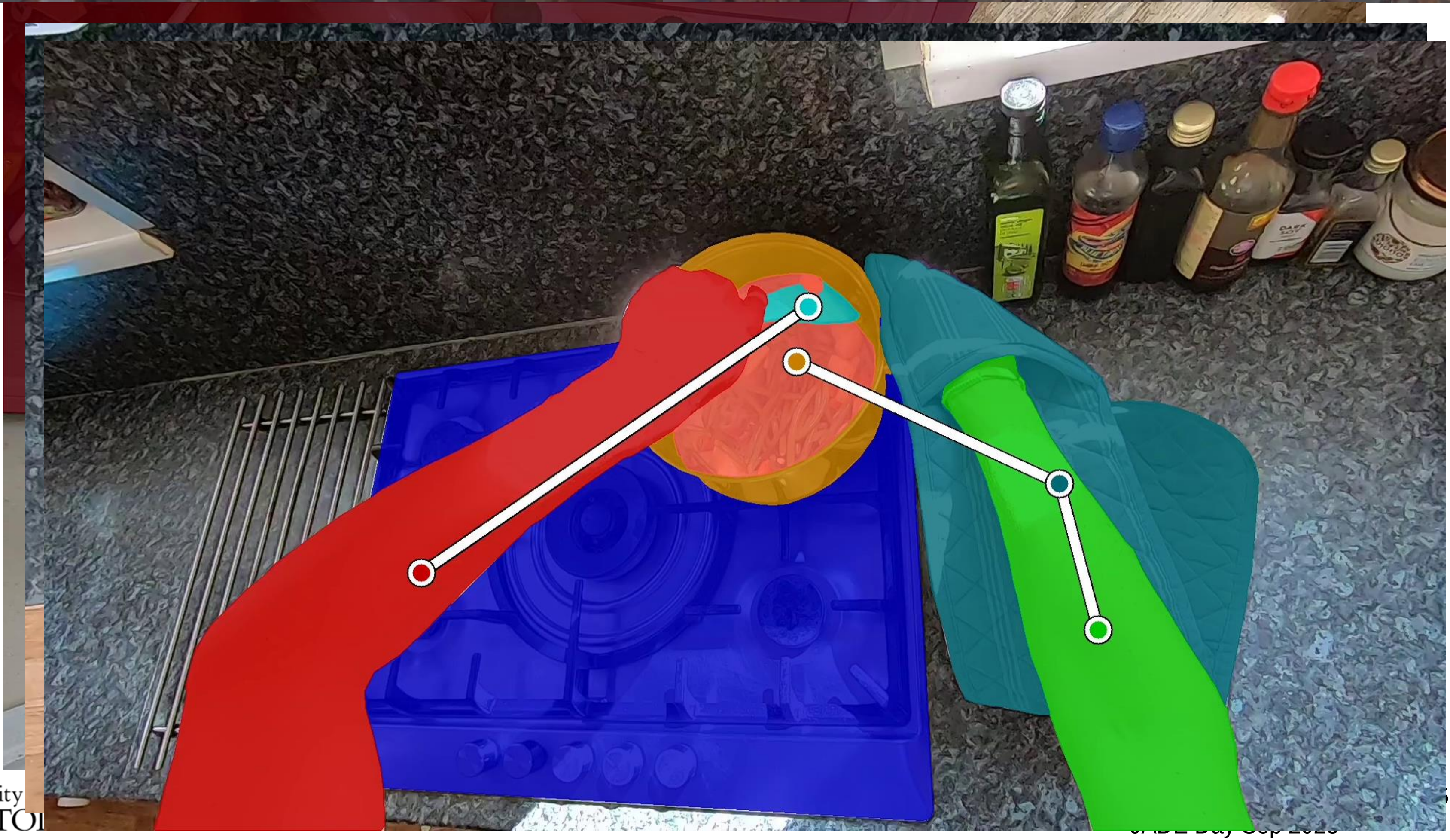
Comparative Stats



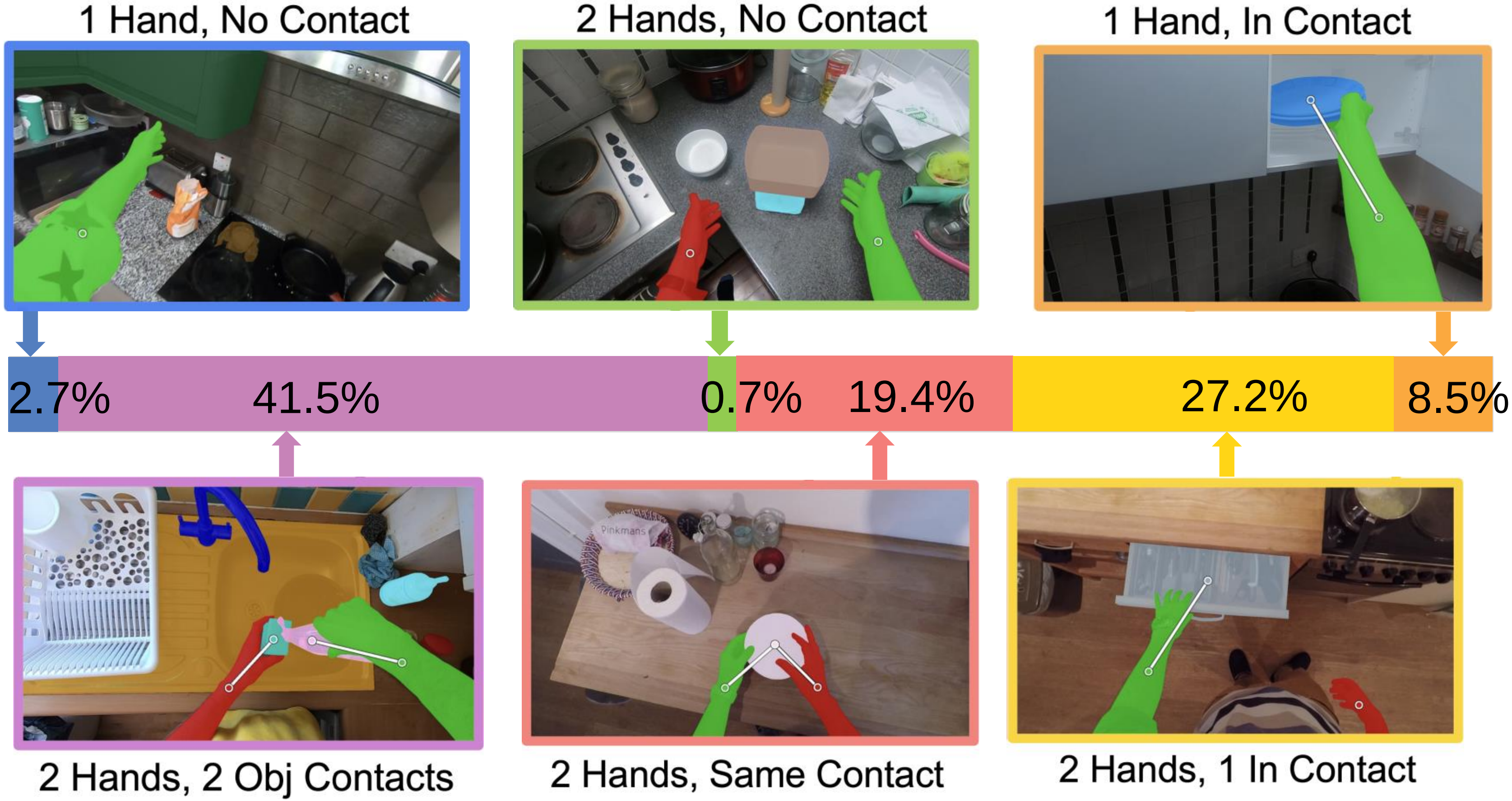
	Basic Statistics		Pixel-Level Annotations	Action Annotations		
	Total	Avg			#Action	#Entity
Dataset	Mins	Seq Ln	Total Masks	Actions	Classes	Classes
EgoHand [3]	72	-	15.1K	-	-	2
DAVIS [6]	8	3s	32.0K	-	-	-
YTVOS [43]	335	5s	197.2K	-	-	94
UVOv0.5 (Sparse) [41]	511	3s	*200.6K	10,213	300	-
VISOR (Ours)	**2,180**	**12s**†	**271.6K**	**27,961**	**2,594**	**257**

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

VISOR Relations



Object relation stats



1 Hand, No Contact

2 Hands, No Contact

1 Hand, In Contact

2.7% 41.5% 0.7% 19.4% 27.2% 8.5%

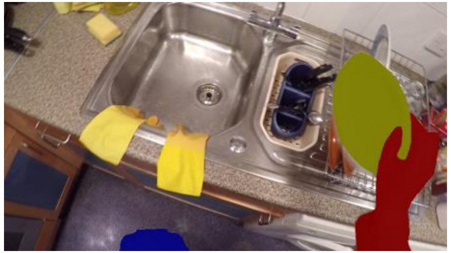
2 Hands, 2 Obj Contacts

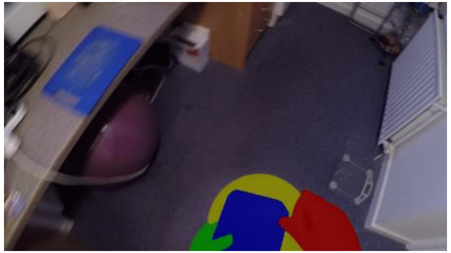
2 Hands, Same Contact

2 Hands, 1 In Contact

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

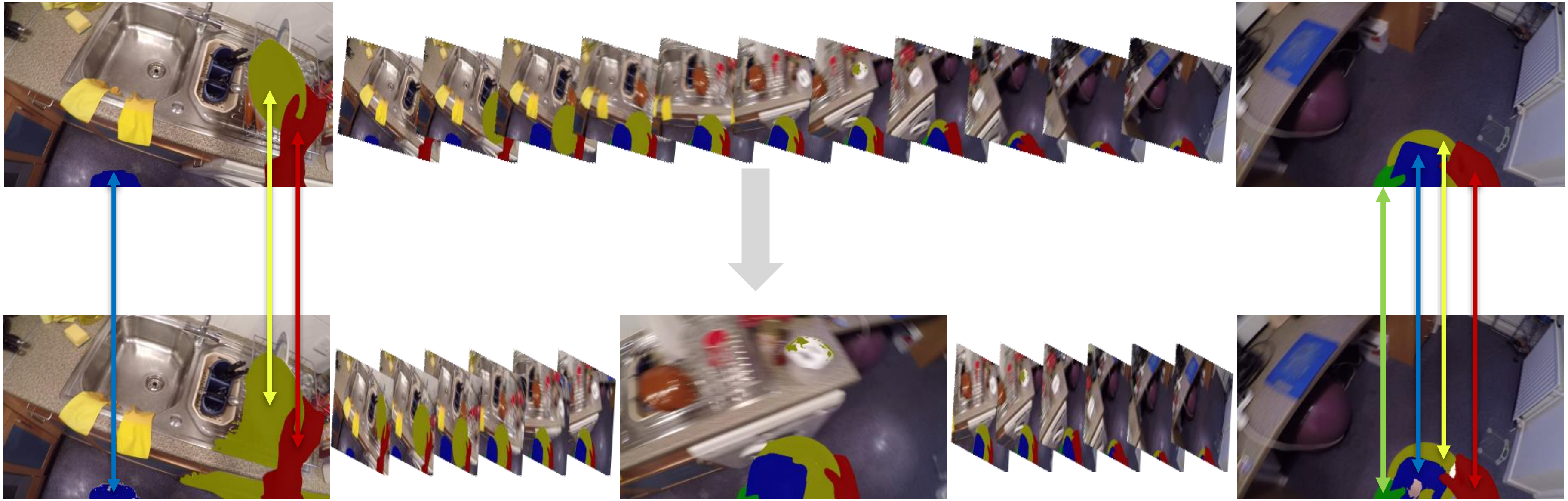
Dense Annotations





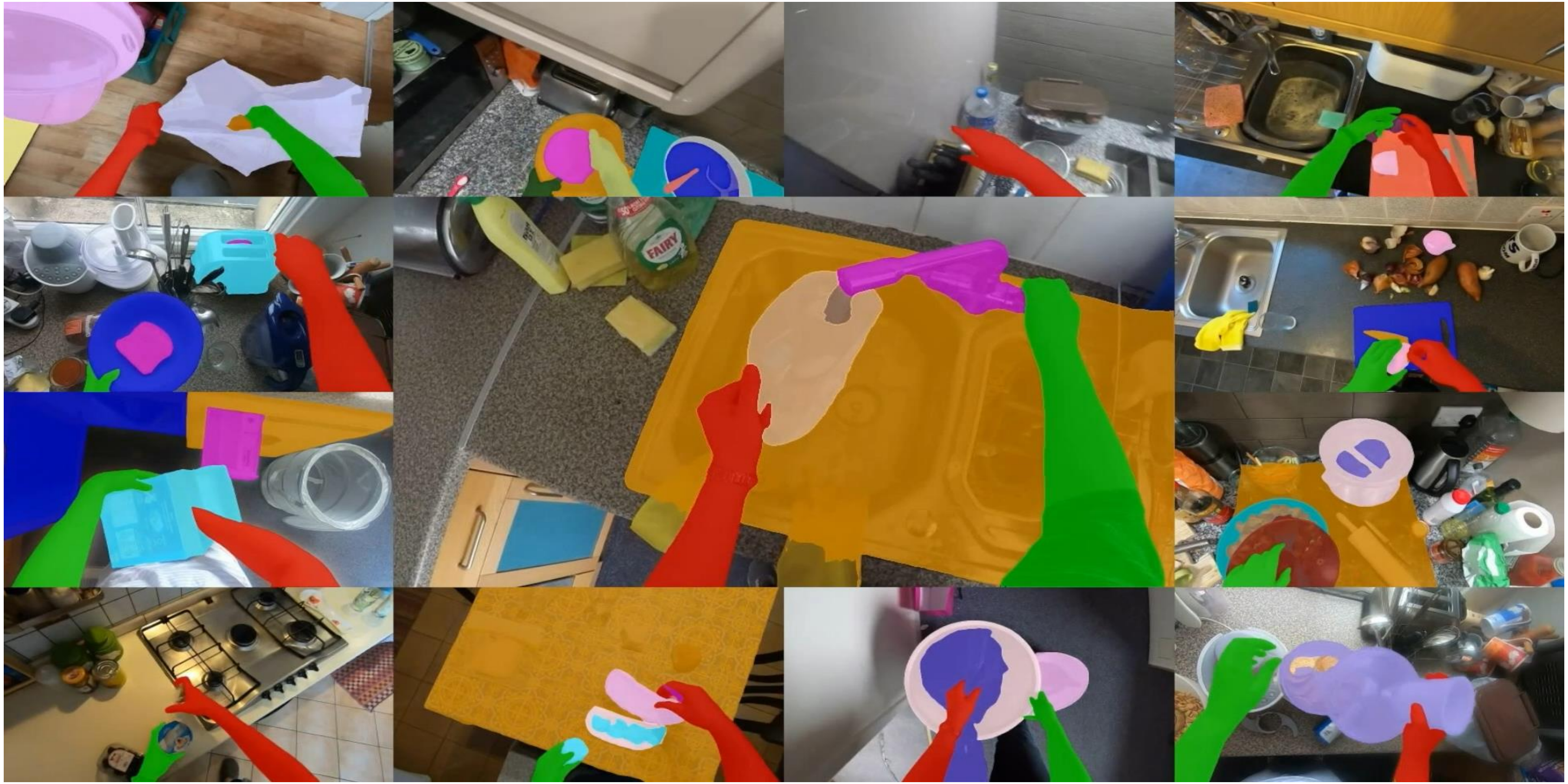
A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

Dense Annotations



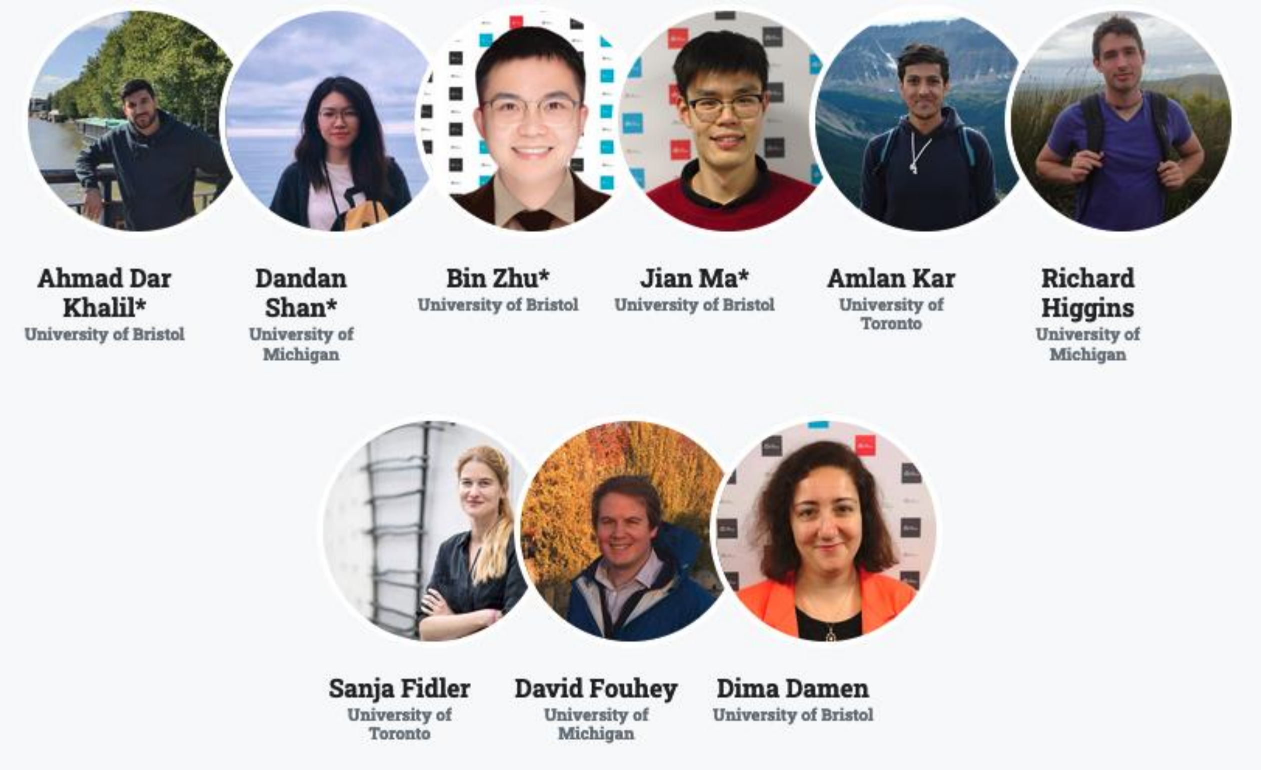
A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

EPIC-KITCHENS VISOR



A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

EPIC-KITCHENS VISOR



Ahmad Dar Khalil*
University of Bristol

Dandan Shan*
University of Michigan

Bin Zhu*
University of Bristol

Jian Ma*
University of Bristol

Amlan Kar
University of Toronto

Richard Higgins
University of Michigan

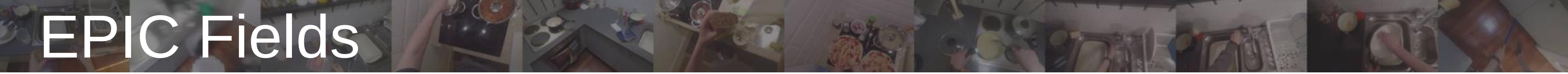
Sanja Fidler
University of Toronto

David Fouhey
University of Michigan

Dima Damen
University of Bristol

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

EPIC-KITCHENS



EPIC Fields

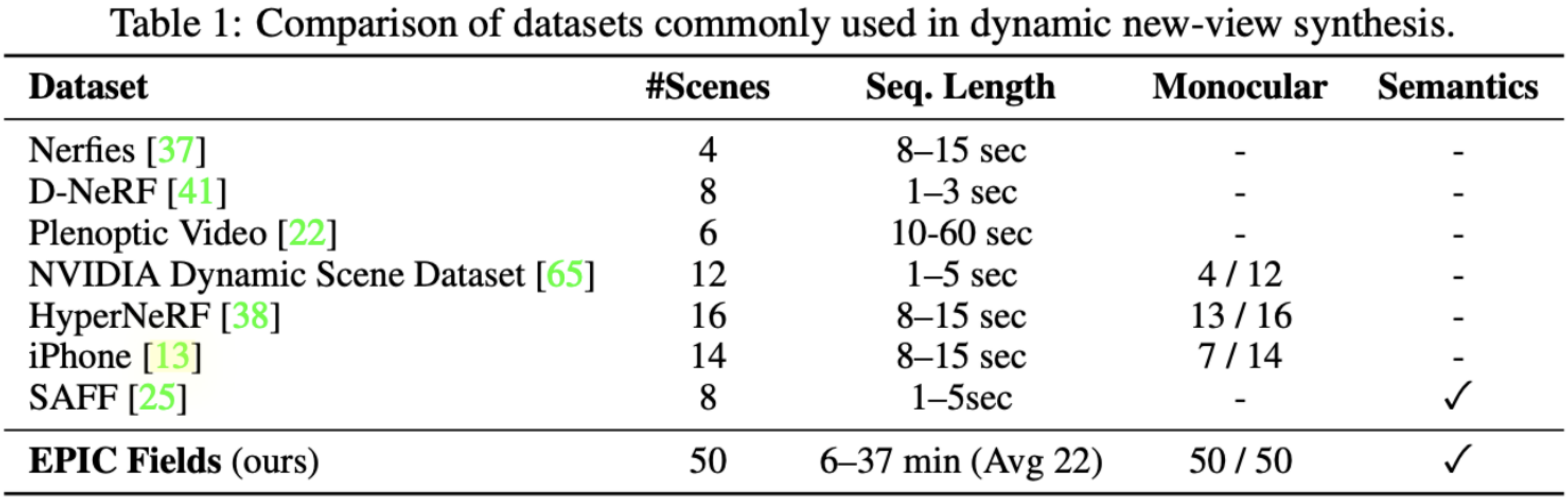
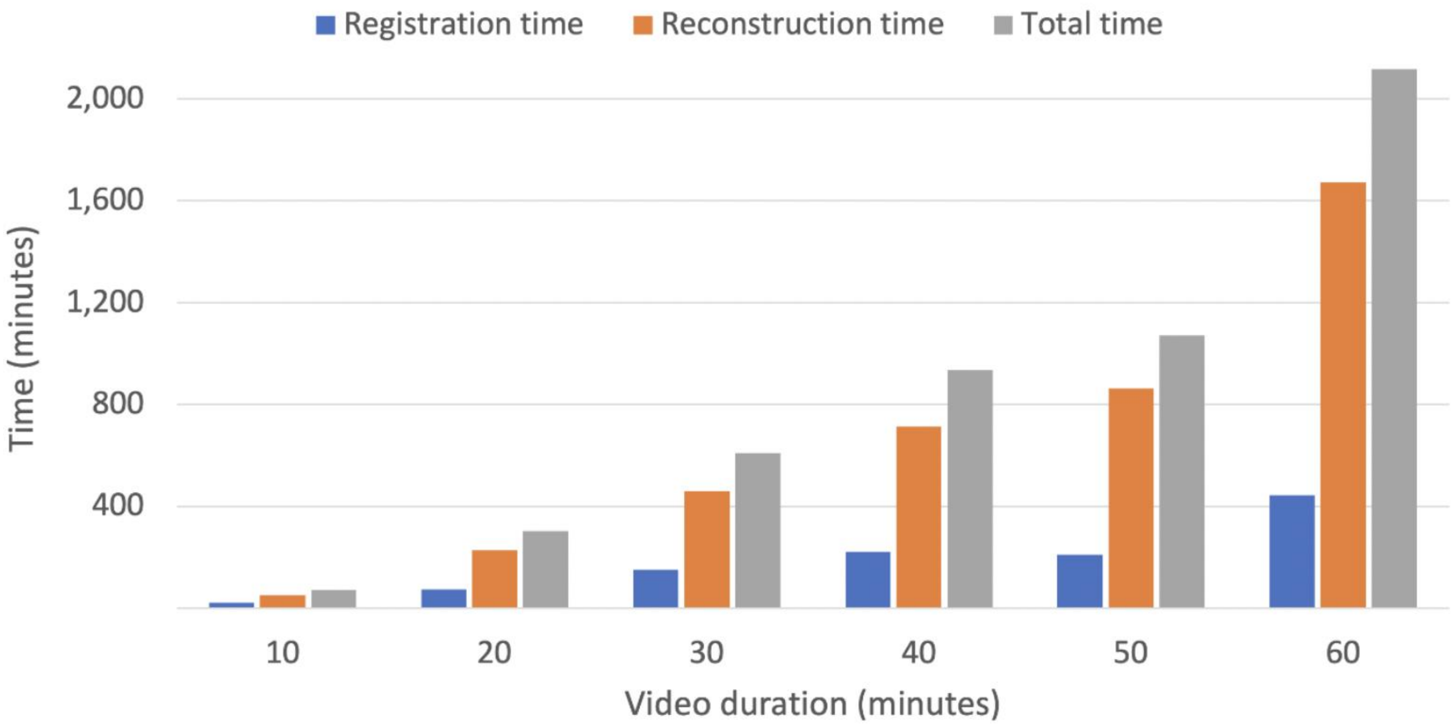


Table 1: Comparison of datasets commonly used in dynamic new-view synthesis.

Dataset	#Scenes	Seq. Length	Monocular	Semantics
Nerfies [37]	4	8–15 sec	-	-
D-NeRF [41]	8	1–3 sec	-	-
Plenoptic Video [22]	6	10-60 sec	-	-
NVIDIA Dynamic Scene Dataset [65]	12	1–5 sec	4 / 12	-
HyperNeRF [38]	16	8–15 sec	13 / 16	-
iPhone [13]	14	8–15 sec	7 / 14	-
SAFF [25]	8	1–5sec	-	✓
EPIC Fields (ours)	50	6–37 min (Avg 22)	50 / 50	✓

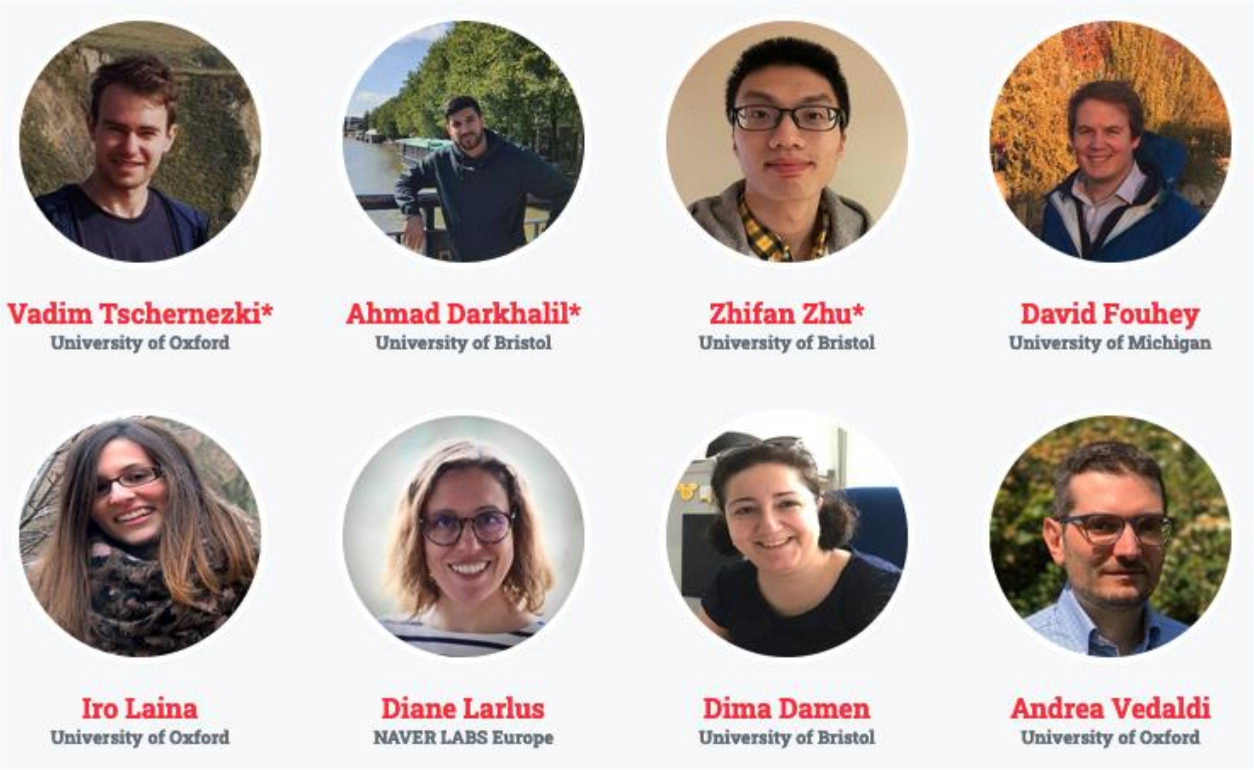
V Tschernezkiet et al. EPIC Fields: Marrying 3D Geometry and Viideo Understanding. *NeurIPS 2023*

EPIC Fields



V Tschernezkiet et al. EPIC Fields: Marrying 3D Geometry and Viideo Understanding. *NeurIPS 2023*

EPIC Fields



Vadim Tschernezki*
University of Oxford

Ahmad Darkhalil*
University of Bristol

Zhifan Zhu*
University of Bristol

David Fouhey
University of Michigan

Iro Laina
University of Oxford

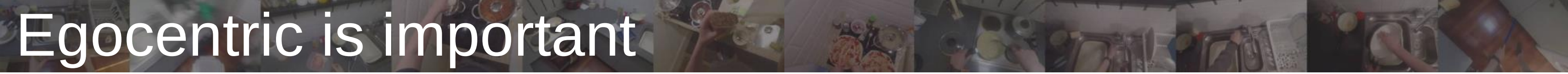
Diane Larlus
NAVER LABS Europe

Dima Damen
University of Bristol

Andrea Vedaldi
University of Oxford

V Tschernezkiet et al. EPIC Fields: Marrying 3D Geometry and Video Understanding. *NeurIPS 2023*

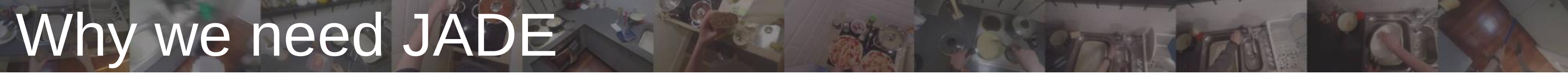
Egocentric is important



- How humans see the world
- Challenge for machine learning
- Lots of applications

What’s next?

- How do we use 3D? Can we calculate on the fly?
- Move to open vocabulary
- Develop models for true long-term understanding
- One model to do all tasks



Why we need JADE

- Large datasets
- Large video models
- Providing pre-trained models
- Move to 3D
- Compete/collaborate with industry

- Code, data and pretrained models publicly available for everything in this talk
- https://epic-kitchens.github.io
- https://tobyperrett.github.io/lmr