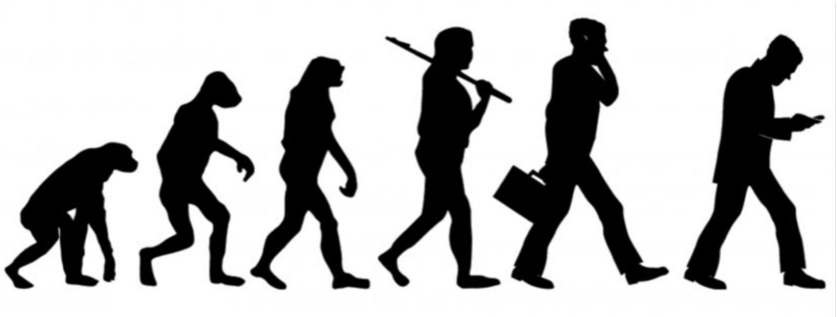
Opportunities in Egocentric Video Understanding

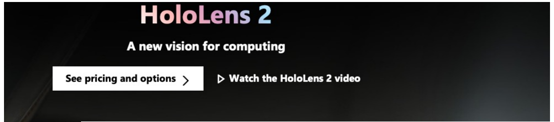
Dima Damen
WACV2024 – Waikoloa, Hawaii

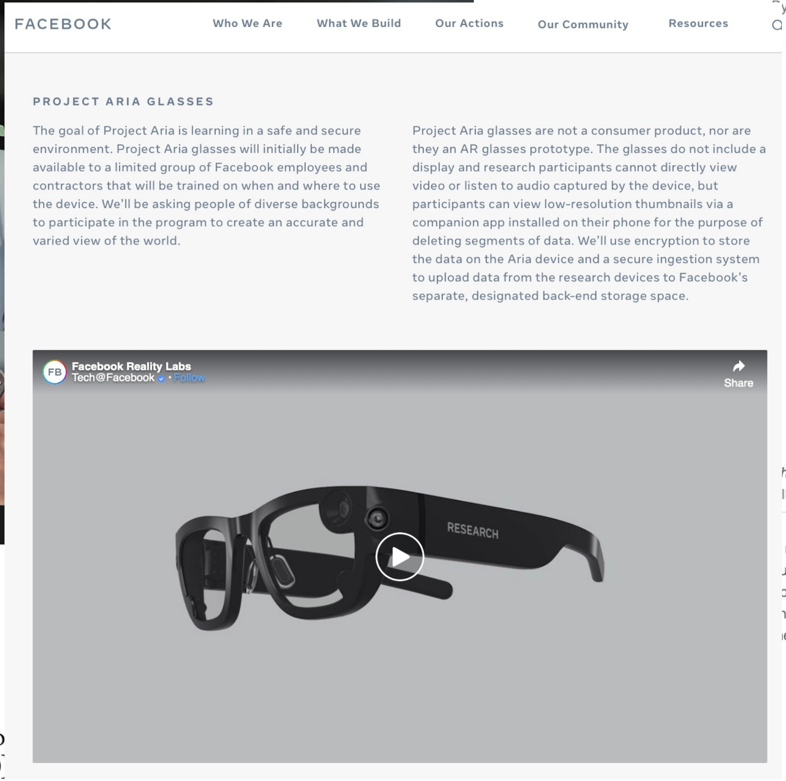
The present…

Photo *Illustration* by Pelle Cass



The future…




FACEBOOK
Who We Are | What We Build | Our Actions | Our Community | Resources

PROJECT ARIA GLASSES

The goal of Project Aria is learning in a safe and secure environment. Project Aria glasses will initially be made available to a limited group of Facebook employees and contractors that will be trained on when and where to use the device. We'll be asking people of diverse backgrounds to participate in the program to create an accurate and varied view of the world.

Project Aria glasses are not a consumer product, nor are they an AR glasses prototype. The glasses do not include a display and research participants cannot directly view video or listen to audio captured by the device, but participants can view low-resolution thumbnails via a companion app installed on their phone for the purpose of deleting segments of data. We'll use encryption to store the data on the Aria device and a secure ingestion system to upload data from the research devices to Facebook's separate, designated back-end storage space.

Facebook Reality Labs
Tech@Facebook · Follow
Share
RESEARCH

Samsung patent application reveals augmented reality headset design

It comes as the Gear VR slowly fades away

The future…

University of BRISTOL

Surveillance vs Sousveillance

Surveillance

Sousveillance

GEORGE FLOYD

Teen with 'cell phone and sheer guts' credited for Derek Chauvin's murder conviction

CNNWire By Holly Yan, CNN

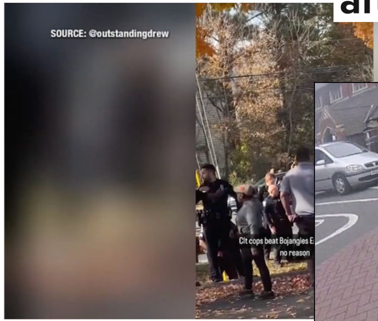
Wednesday, April 21, 2021 6:07PM

Video shows Charlotte officer repeatedly hitting pinned woman during arrest: 'Not easy to watch'

WTVD-AP

Friday, November 17, 2023

'They could've killed him': Jacksonville family wants justice after video of arrest goes viral

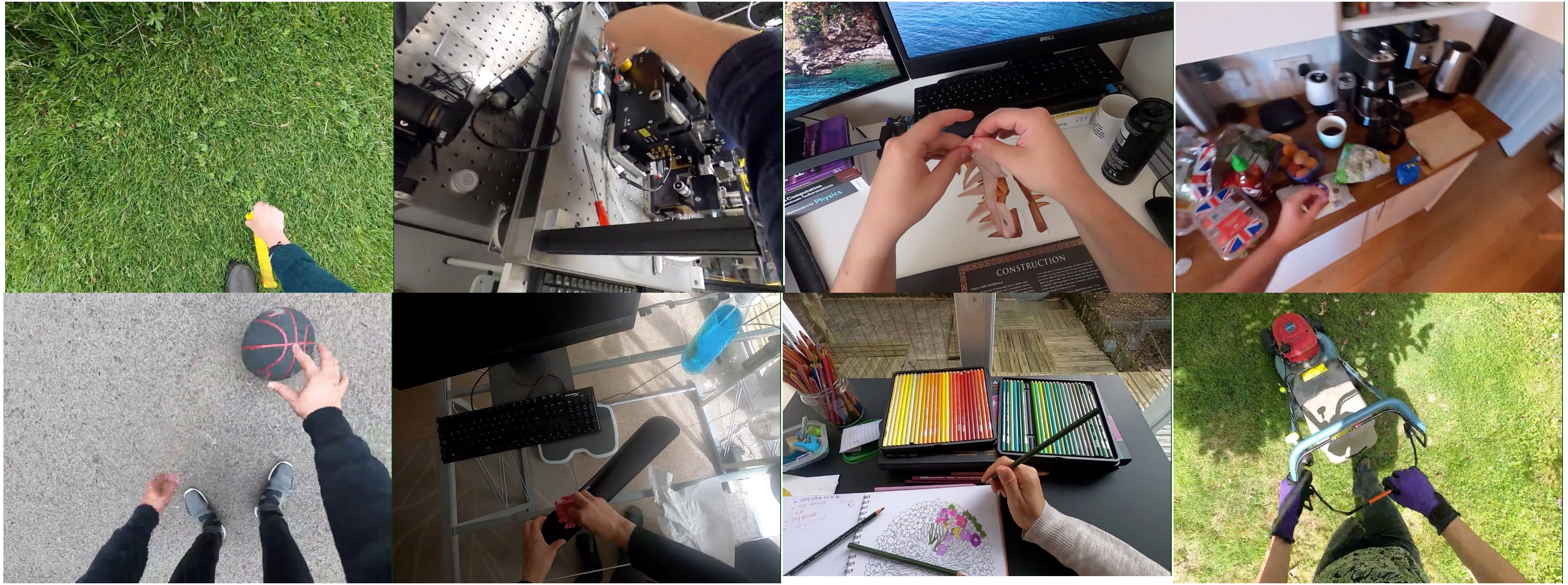


Cyclist's GoPro footage captures

Egocentric cameras are coming

What can we do with such footage?

Egocentric Videos?



Egocentric Videos?

Data Collection Exercises

EPIC KITCHENS

2017 - now

100 hours
45 kitchens
4 countries
Long-term recording
Kitchen-based activities

EGO 4D

2020 - now

6730 hours
923 participants
74 locations
9 countries
Short-term recording
All daily activities

Data Collection Exercises

EGO-EXO4D

2022 - now

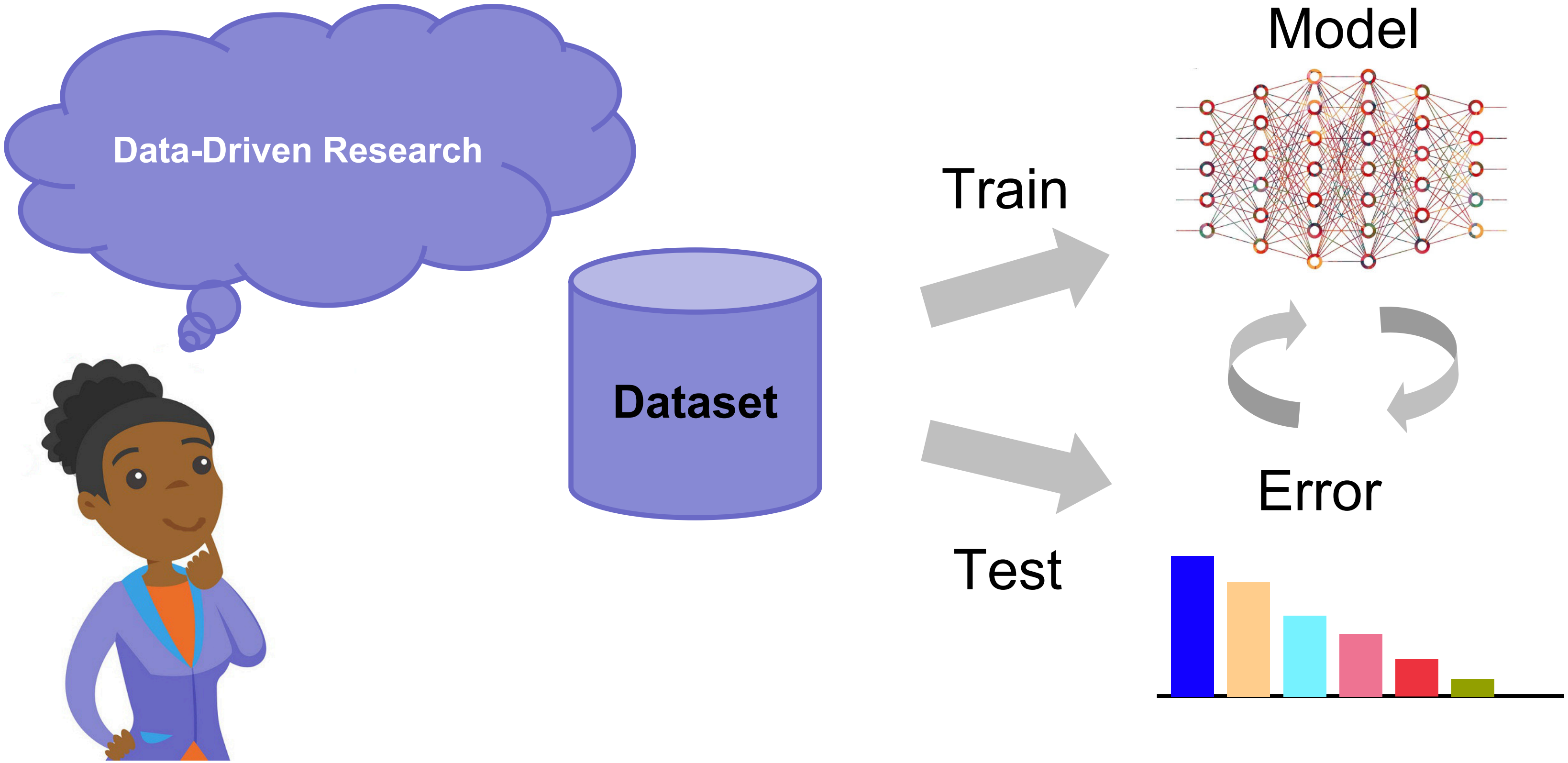
Released Dec 2023
1422 hours
8 skilled activities
839 camera wearers
Ego-Exo recordings

2024 – [coming]

[new recordings]

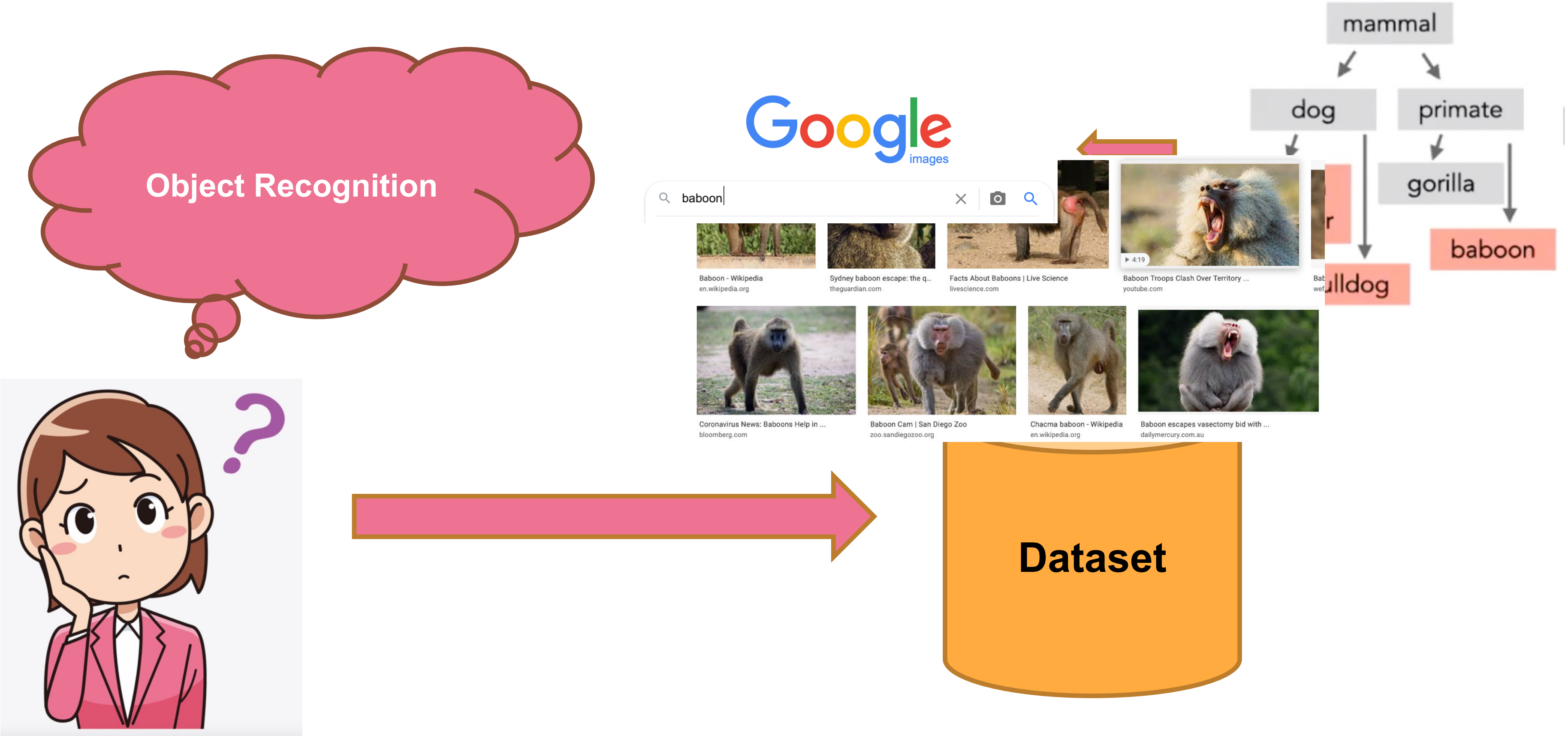
University of BRISTOL

What is … Data…



University of BRISTOL

ImageNet Dataset

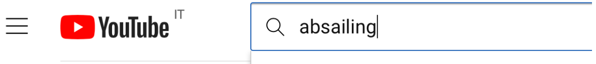


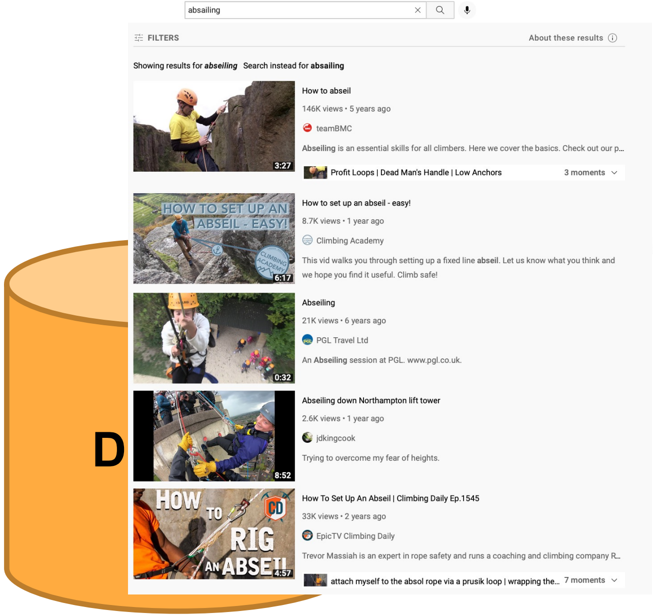
Kinetics Dataset

A. List of Kinetics Human Action Classes

This is the list of classes included in the human action video dataset. The number of clips for each action class is given by the number in brackets following each class name.

1. abseiling (1146)
2. air drumming (1132)
3. answering questions (478)
4. applauding (411)
5. applying cream (478)
6. archery (1147)

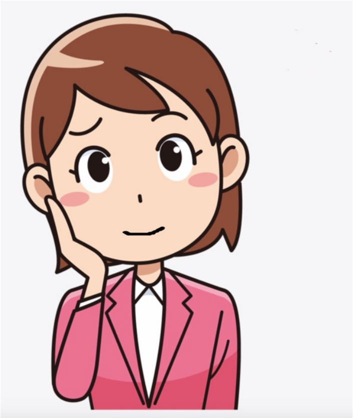




Data First…

~~Object Recognition~~

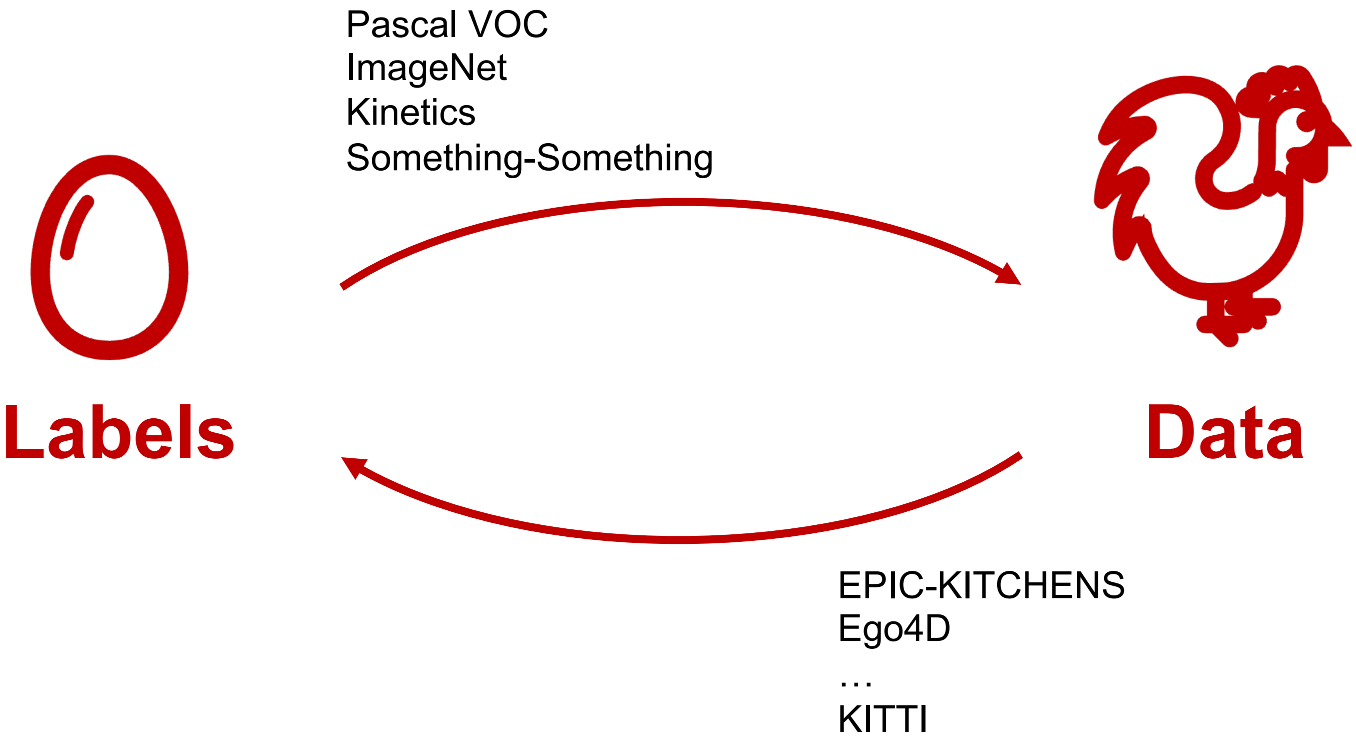
Let's collect Data!



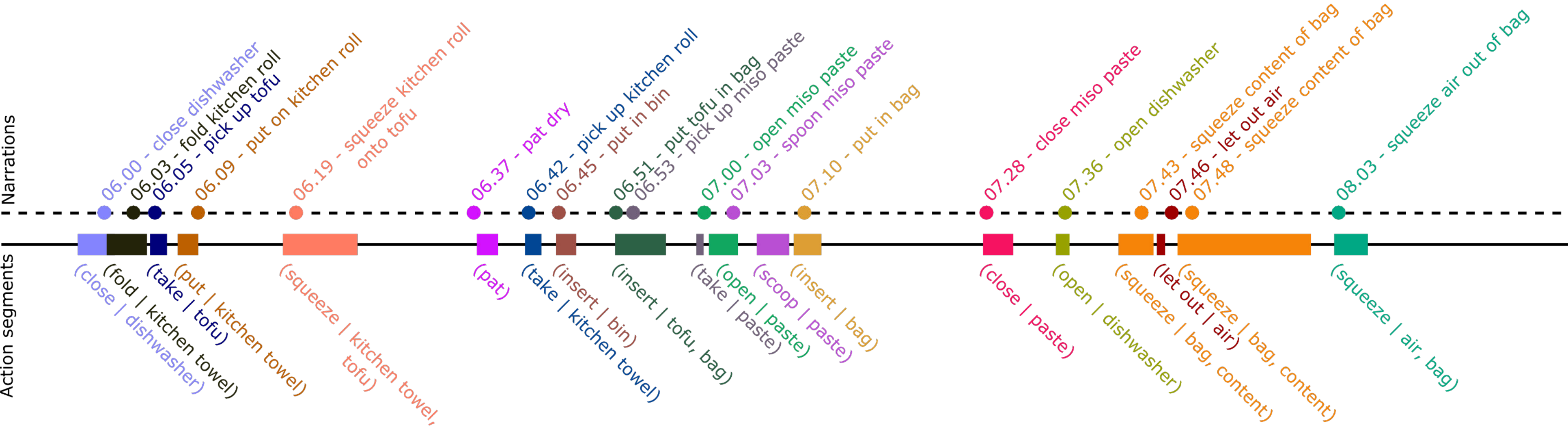
EPIC
KITCHENS-100

EPIC-KITCHENS

Data Collection Exercise



EPIC-KITCHENS



EPIC-KITCHENS

Store
Lay-down **PUT**
Place Rest-on

stove burner
hob
stovetop gas

open vocab	lay-down	stovetop
closed vocab	put	hob
category	leave	appliance

with: Kristen Grauman
+83 authors

Narration

C: camera wearer

#C C scraps off wood filler from one putty knife with the other putty knife

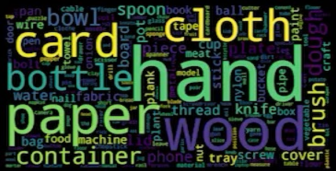
#C C picks up another putty knife from the white board

13.2 sentences/min

3.8 M sentences

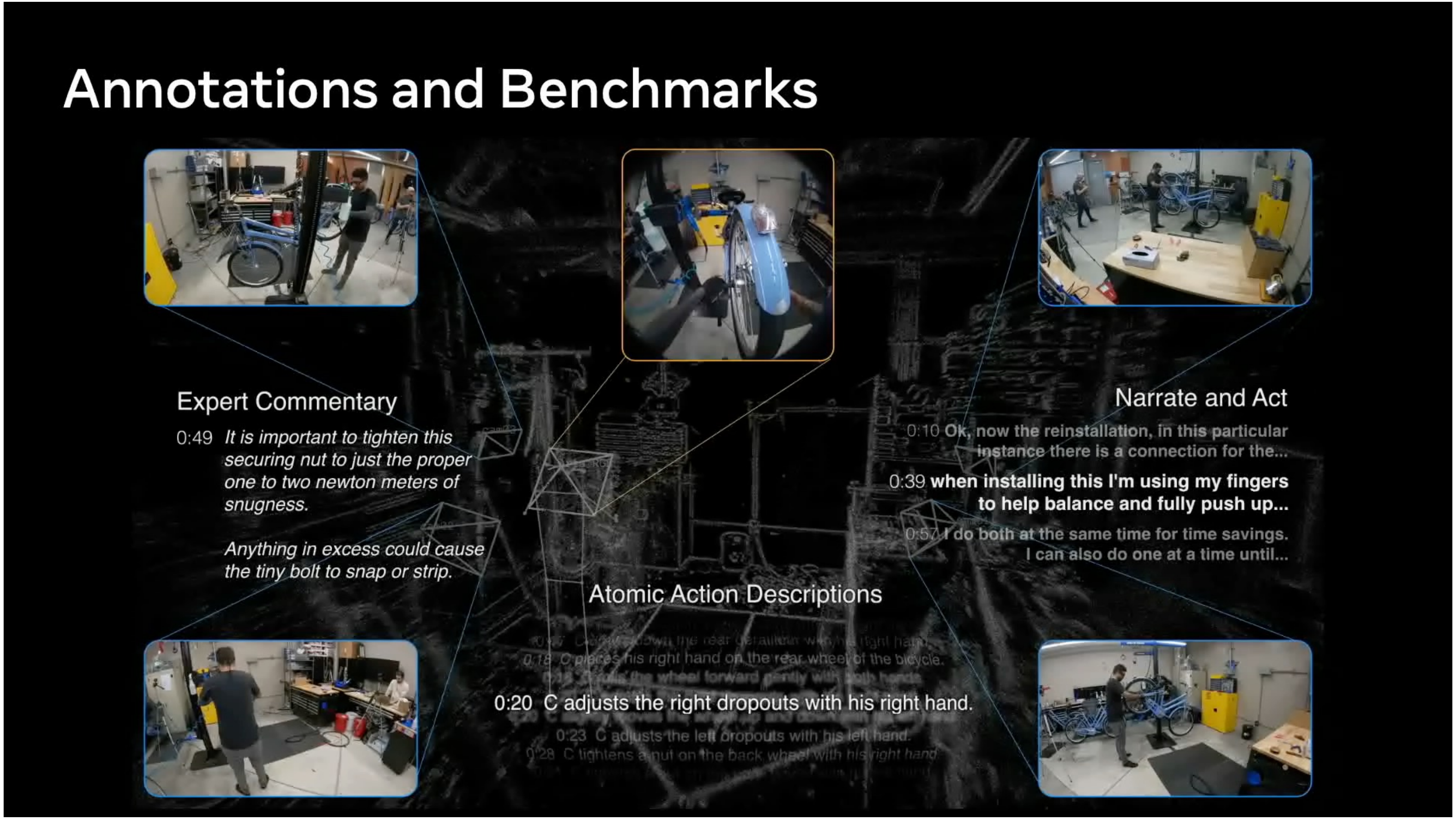
1,772 verbs

4,336 nouns



Ego-Exo4D

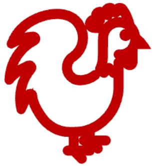
with: Kristen Grauman
+102 authors



University of BRISTOL

The chicken or the egg...

Data



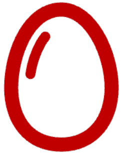
Naturally unbalanced

Harder to label (exposes ambiguity)

Closer to application

Many research opportunities…

Labels



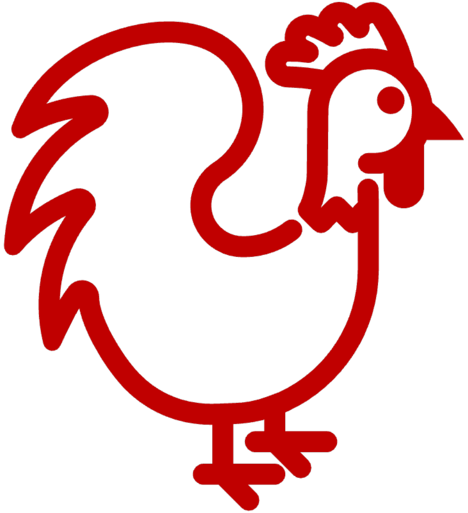
Unnaturally balanced (or nearly)

Easier to label (hides ambiguity)

Can be expanded

Single task

The chicken or the egg...



Data first brings out many opportunities

Opportunities in Egocentric Video Understanding

Opportunities in Egocentric Vision

Tasks are harder

Detection, 3D Mapping, Tracking, VOS, Hand-Object, Generative, …

Solutions prove more rewarding

Weak supervision, Domain Adap/Gen., Audio-Visual, long-term understanding

Opportunities in Egocentric Vision

Tasks are harder

Detection, 3D Mapping, Tracking, VOS, Hand-Object, Generative, …

Solutions prove more rewarding

Weak supervision, Domain Adap/Gen., Audio-Visual, long-term understanding

Opportunities in Egocentric Vision

Tasks are harder

Detection, 3D Mapping, Tracking, VOS, Hand-Object, Generative, …

Solutions prove more rewarding

Weak supervision, Domain Adap/Gen., Audio-Visual, long-term understanding

Action Detection

with: Hanyuan Wang
Majid Mirmehdi
Toby Perrett

Task	Method	0.1	0.2	0.3	0.4	0.5	Avg
Verb	BMN [18,36]	10.8	9.8	8.4	7.1	5.6	8.4
	G-TAD [76]	12.1	11.0	9.4	8.1	6.5	9.4
	Ours	**26.6**	**25.6**	**24.4**	**22.4**	**18.3**	**23.4**
Noun	BMN [18,36]	10.3	8.3	6.2	4.5	3.4	6.5
	G-TAD [76]	11.0	10.0	8.6	7.0	5.4	8.4
	Ours	**25.5**	**24.3**	**22.6**	**20.3**	**16.6**	**21.9**

Table 2. **Results on EPIC-Kitchens 100 validation set.**

Zhang et al (2022). ActionFormer: Localizing Moments of Actions with Transformers. ECCV

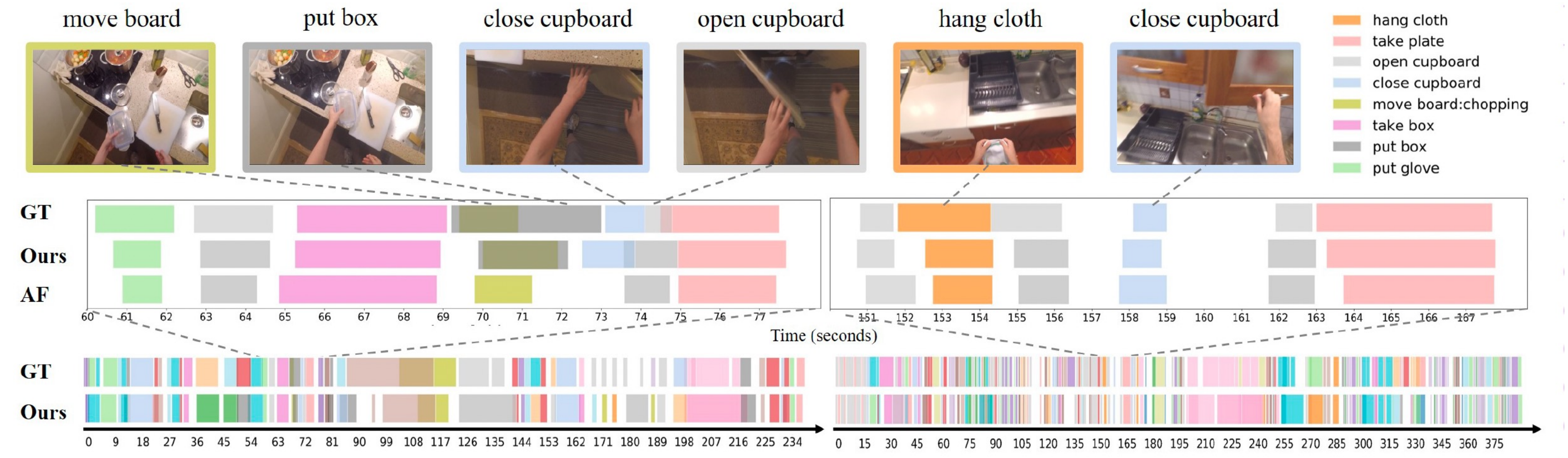


Figure 3. Qualitative results on the EPIC-KITCHENS-100 validation set. Ground truth and predictions are shown with colour-coded class

University of BRISTOL

H Wang et al (2024). Refining Action Boundaries for One-stage Detection. *BMVCW*

Opportunities in Egocentric Vision

Tasks are harder

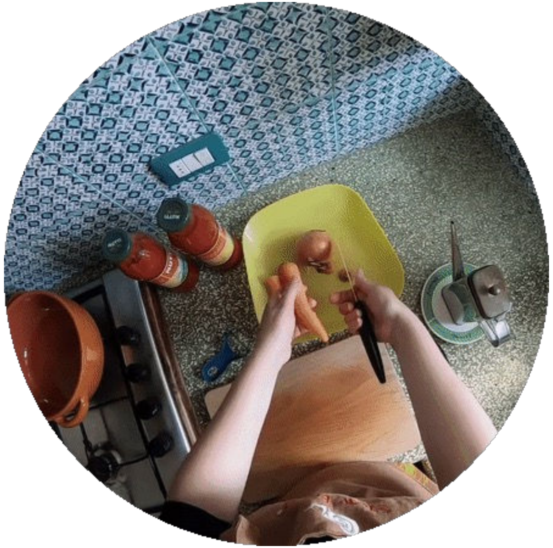
Detection, 3D Mapping, Tracking, VOS, Hand-Object, Generative, …

Solutions prove more rewarding

Weak supervision, Domain Adap/Gen., Audio-Visual, long-term understanding

Generalisation across Scenarios and Locations

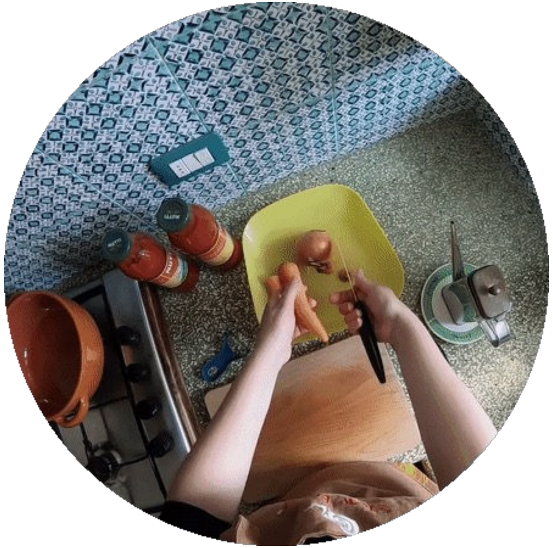
with: Chiara Plizzari
Toby Perrett

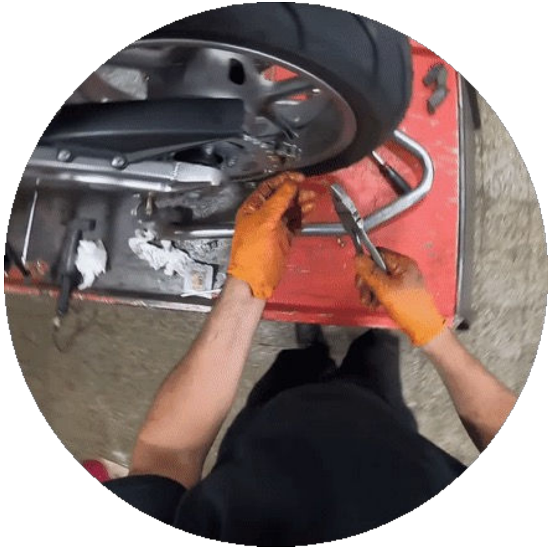


Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

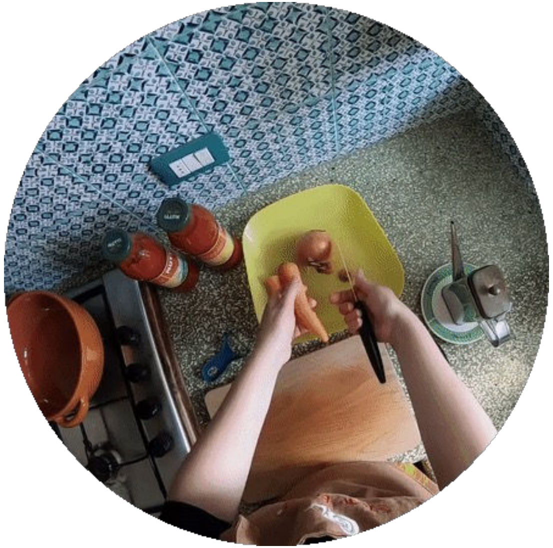


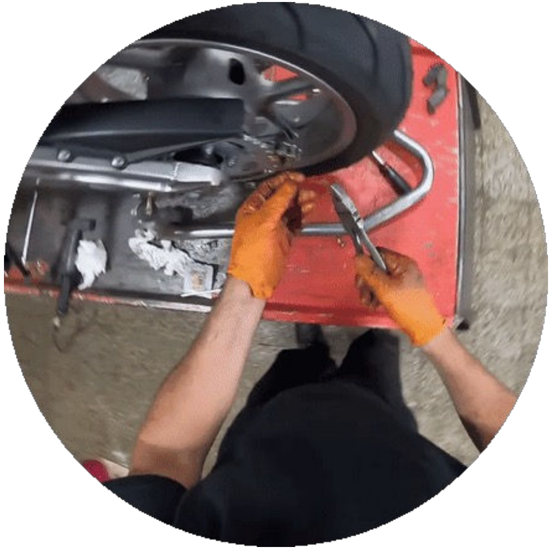


Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett



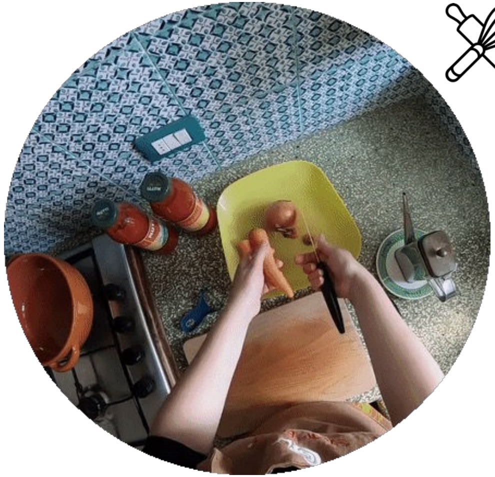


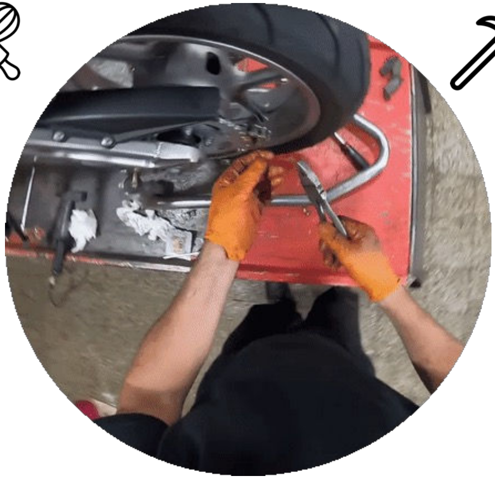
University of BRISTOL

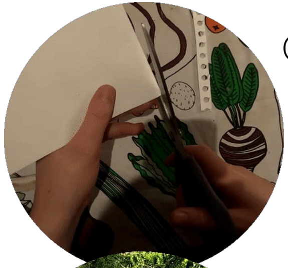
Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett







Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

University of BRISTOL

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Dataset: ARGO1M

with: Chiara Plizzari
Toby Perrett

- We introduce **ARGO1M**, the first dataset to perform **Action Recognition Generalisation** Over Scenarios and Locations

EGO4D

13 locations

10 scenarios

Dataset: ARGO1M

with: Chiara Plizzari
Toby Perrett

- We introduce **ARGO1M**, the first dataset to perform **Action Recognition Generalisation** Over Scenarios and Locations

EGO4D

13 locations

10 scenarios

60 action classes

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Dataset: ARGO1M

with: Chiara Plizzari
Toby Perrett

- We introduce **ARGO1M**, the first dataset to perform **Action Recognition Generalisation** Over Scenarios and Locations

NEW **1.1M samples**

EGO4D

13 locations

10 scenarios

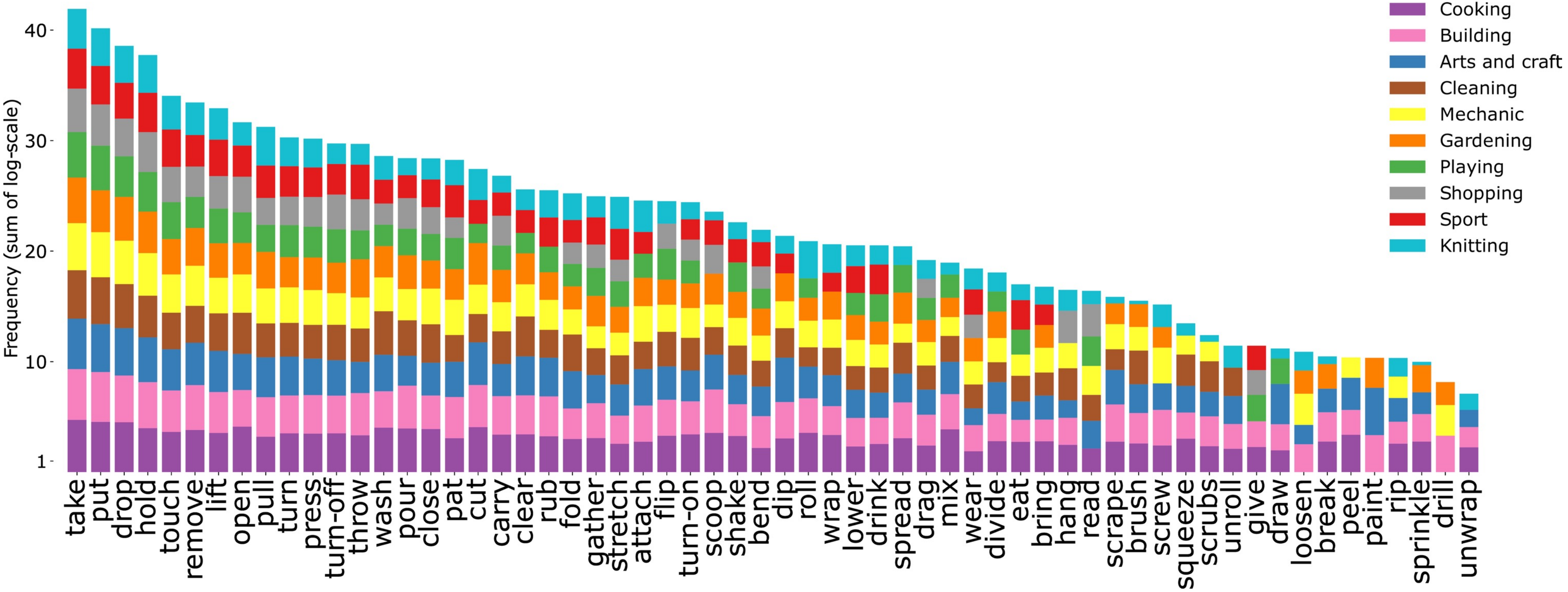
60 action classes

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

with: Chiara Plizzari
Toby Perrett

ARGO1M: 1.05M action clips from 60 action classes recorded in 13 locations within 10 scenarios



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

ARGO1M Splits

with: Chiara Plizzari
Toby Perrett

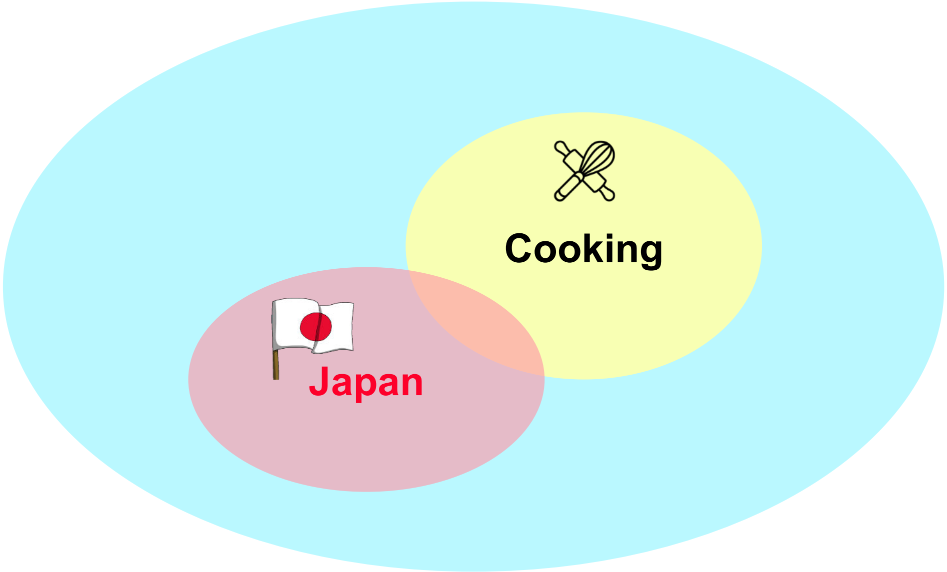
ARGO1M

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

ARGO1M Splits

with: Chiara Plizzari
Toby Perrett

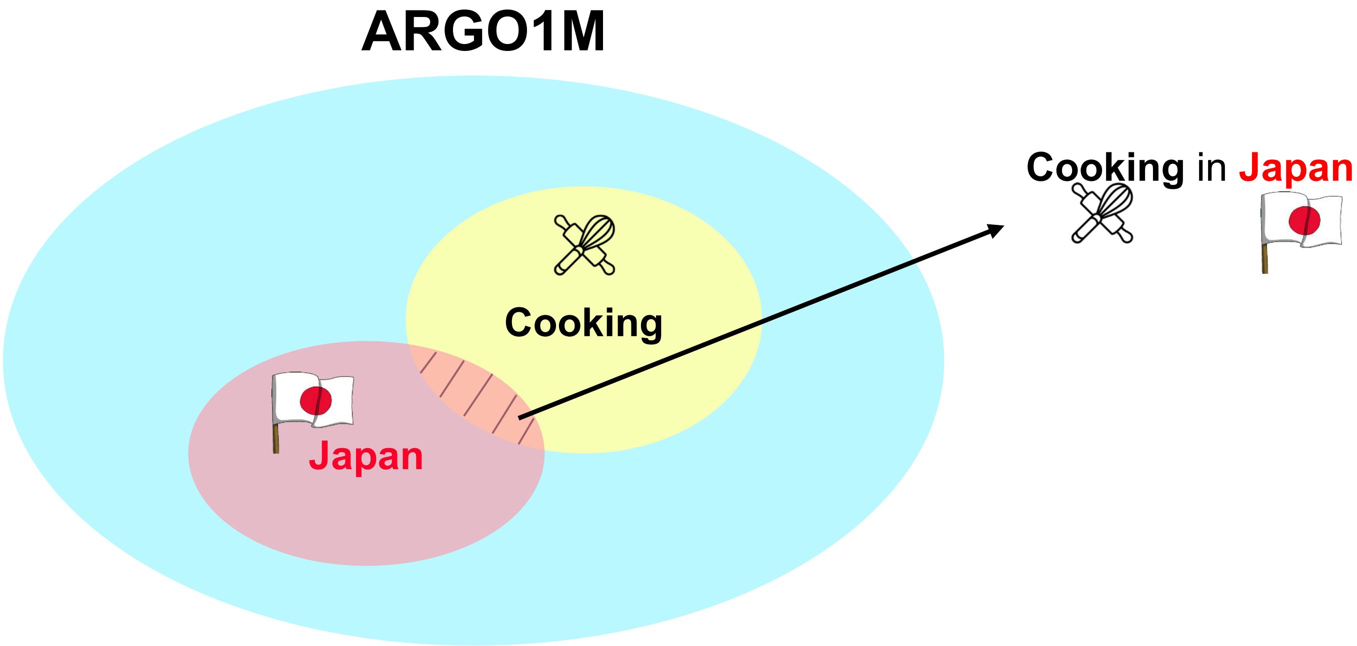
ARGO1M



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

ARGO1M Splits

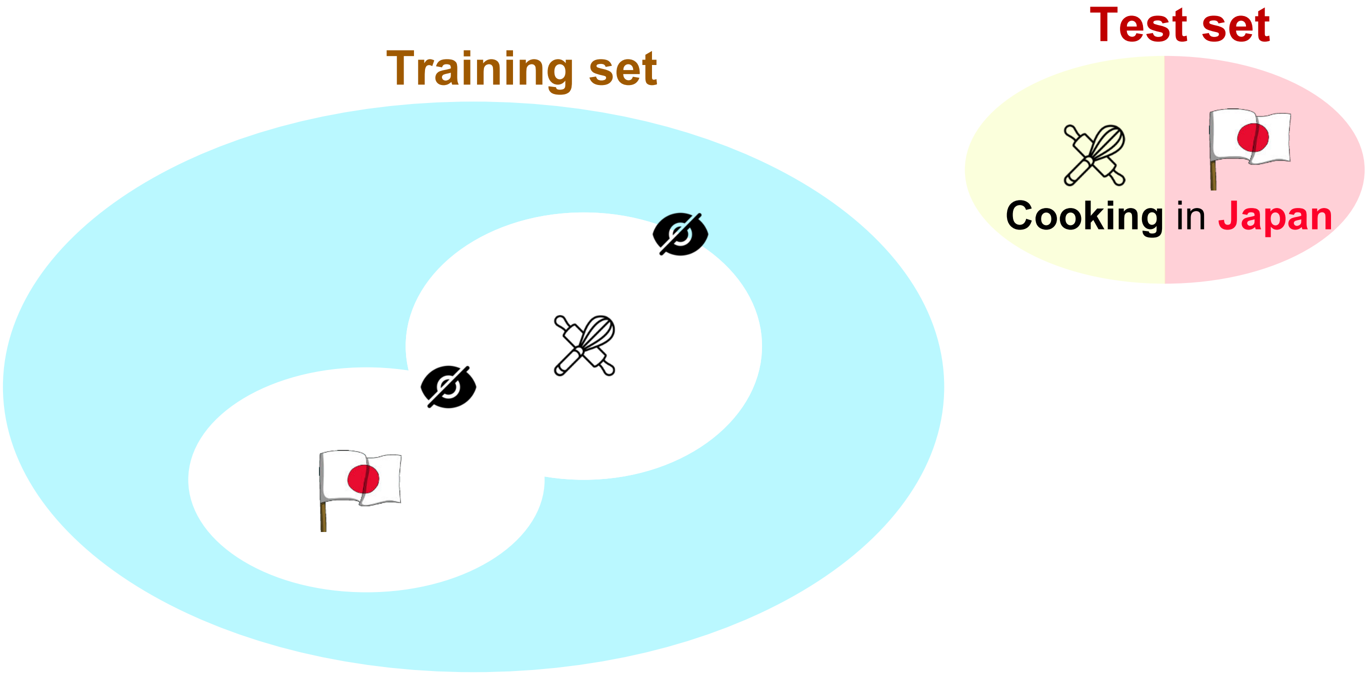
with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

ARGO1M Splits

with: Chiara Plizzari
Toby Perrett

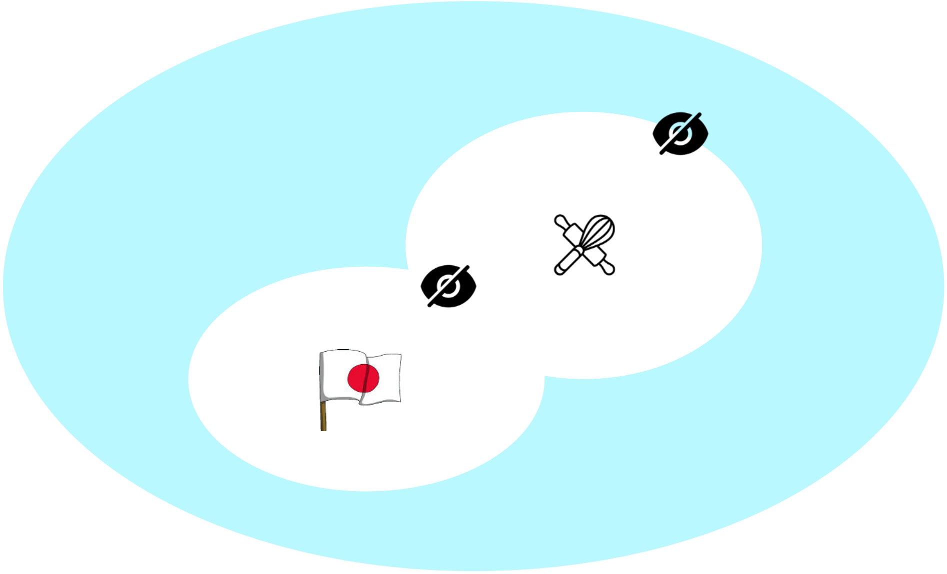


Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

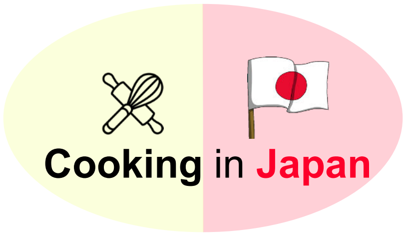
ARGO1M Splits

with: Chiara Plizzari
Toby Perrett

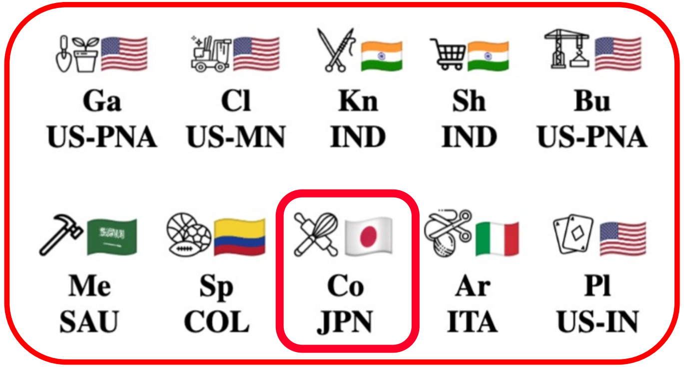
Training set



Test set



10 test sets

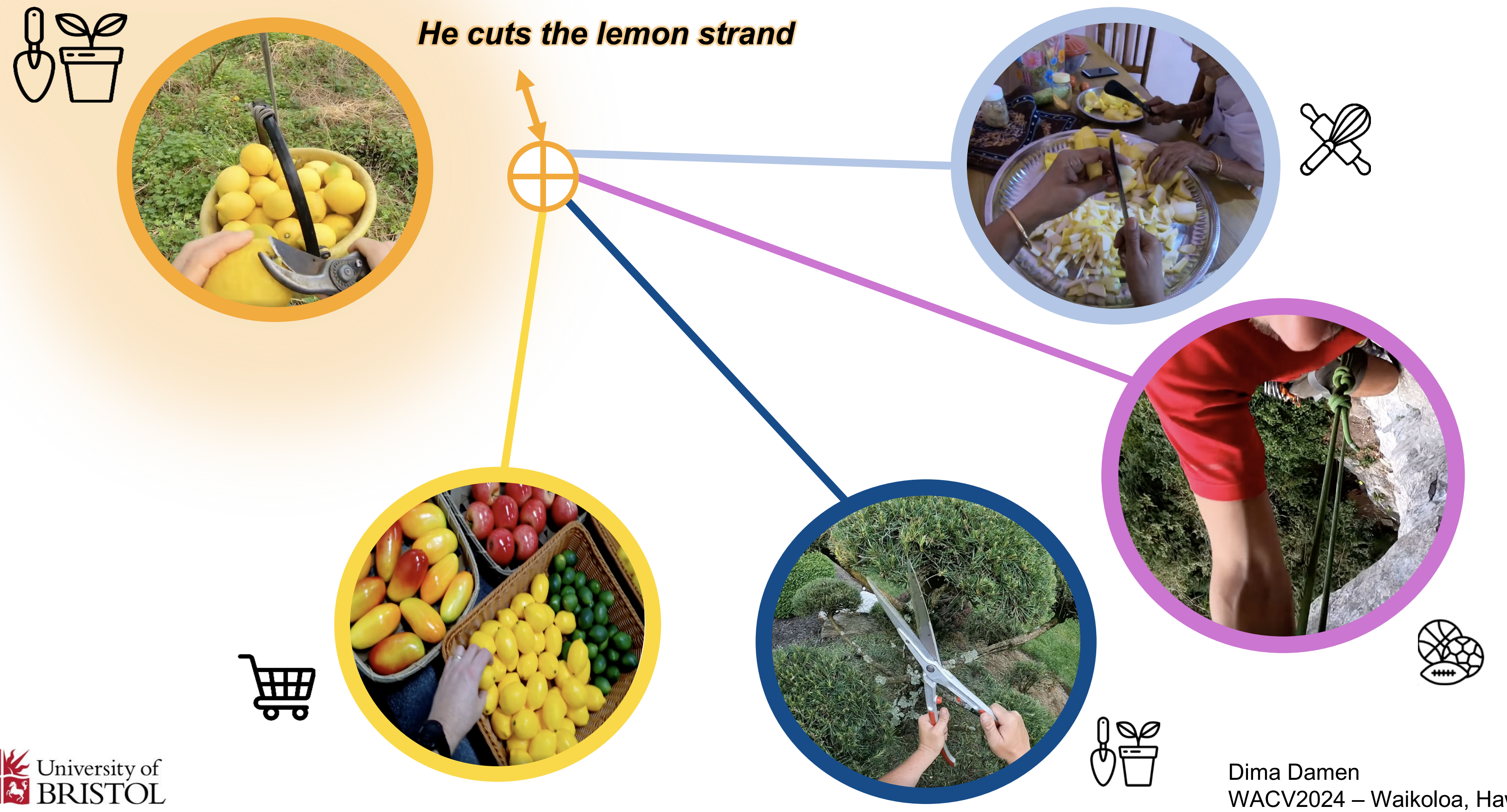


Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Generalisation across Scenarios and Locations

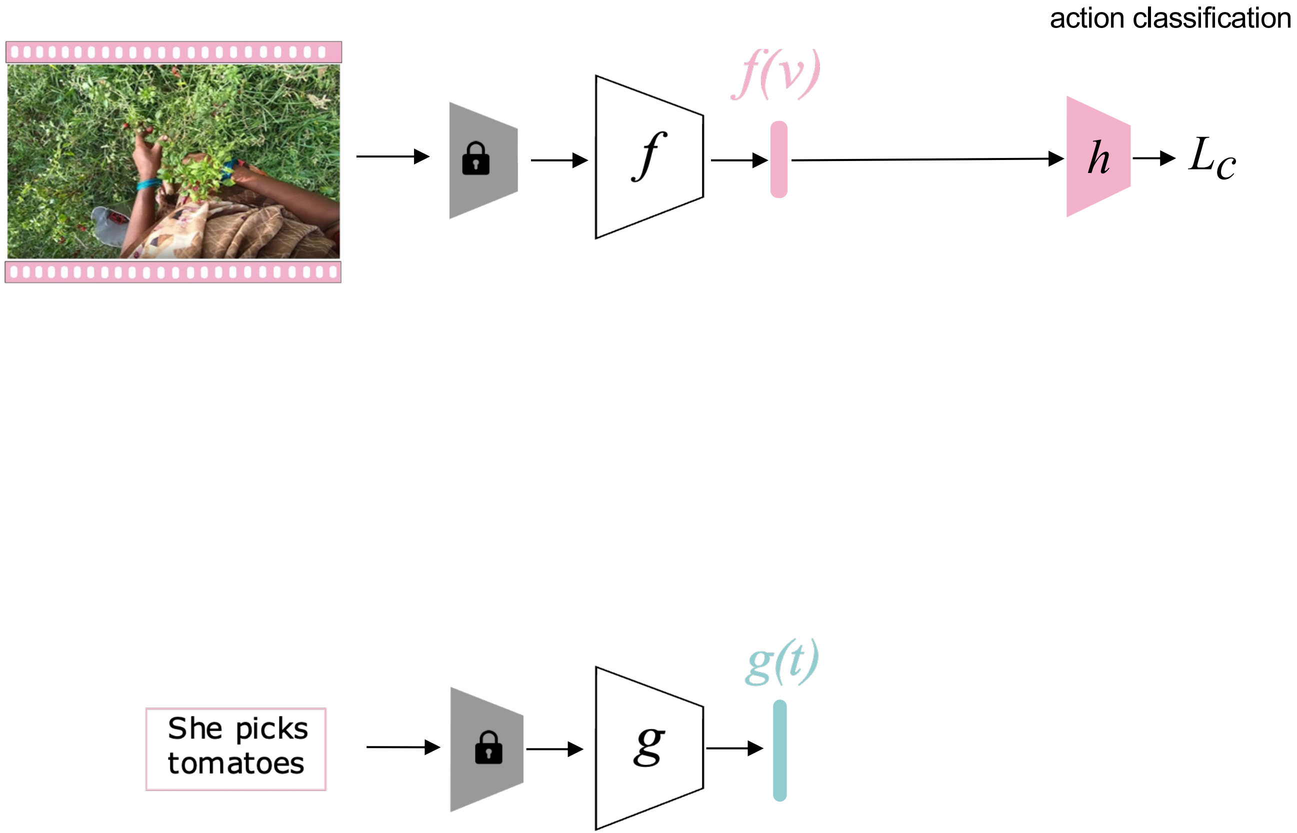
with: Chiara Plizzari
Toby Perrett

He cuts the lemon strand



Proposed method: CIR

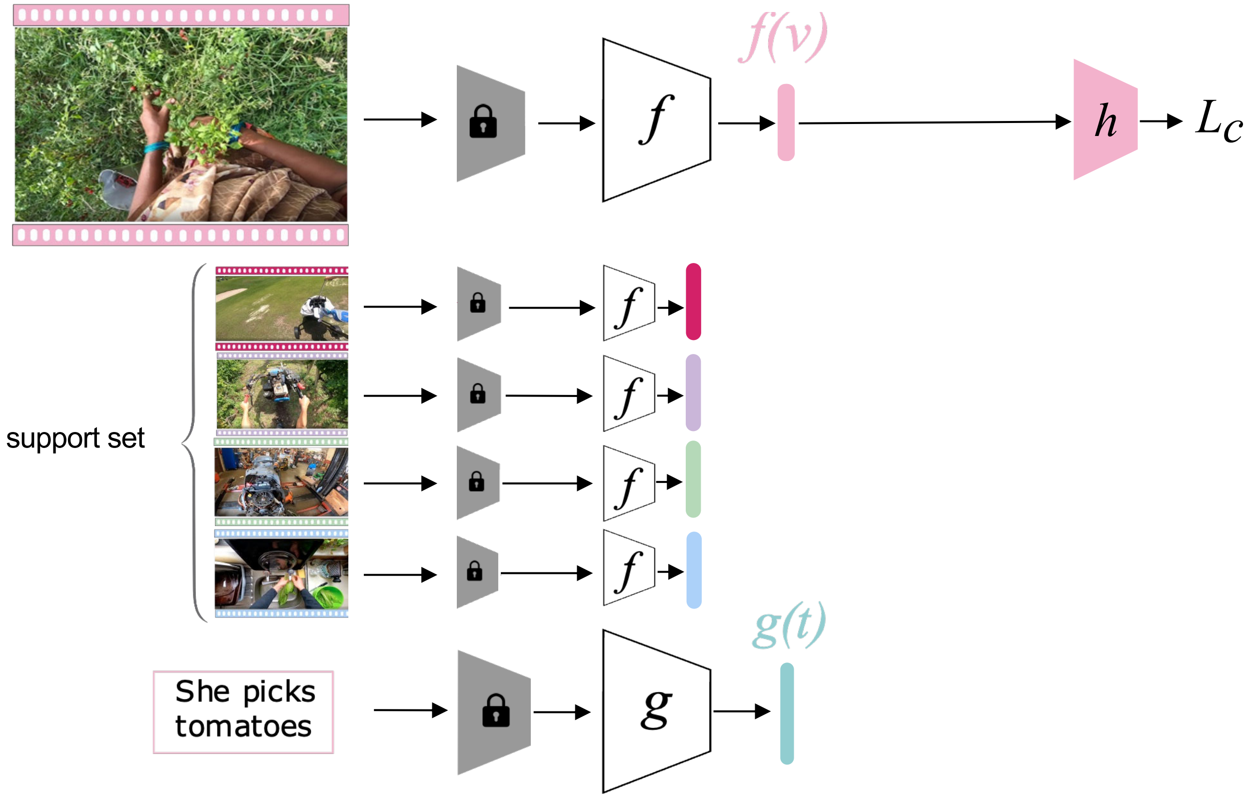
with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Proposed method: CIR

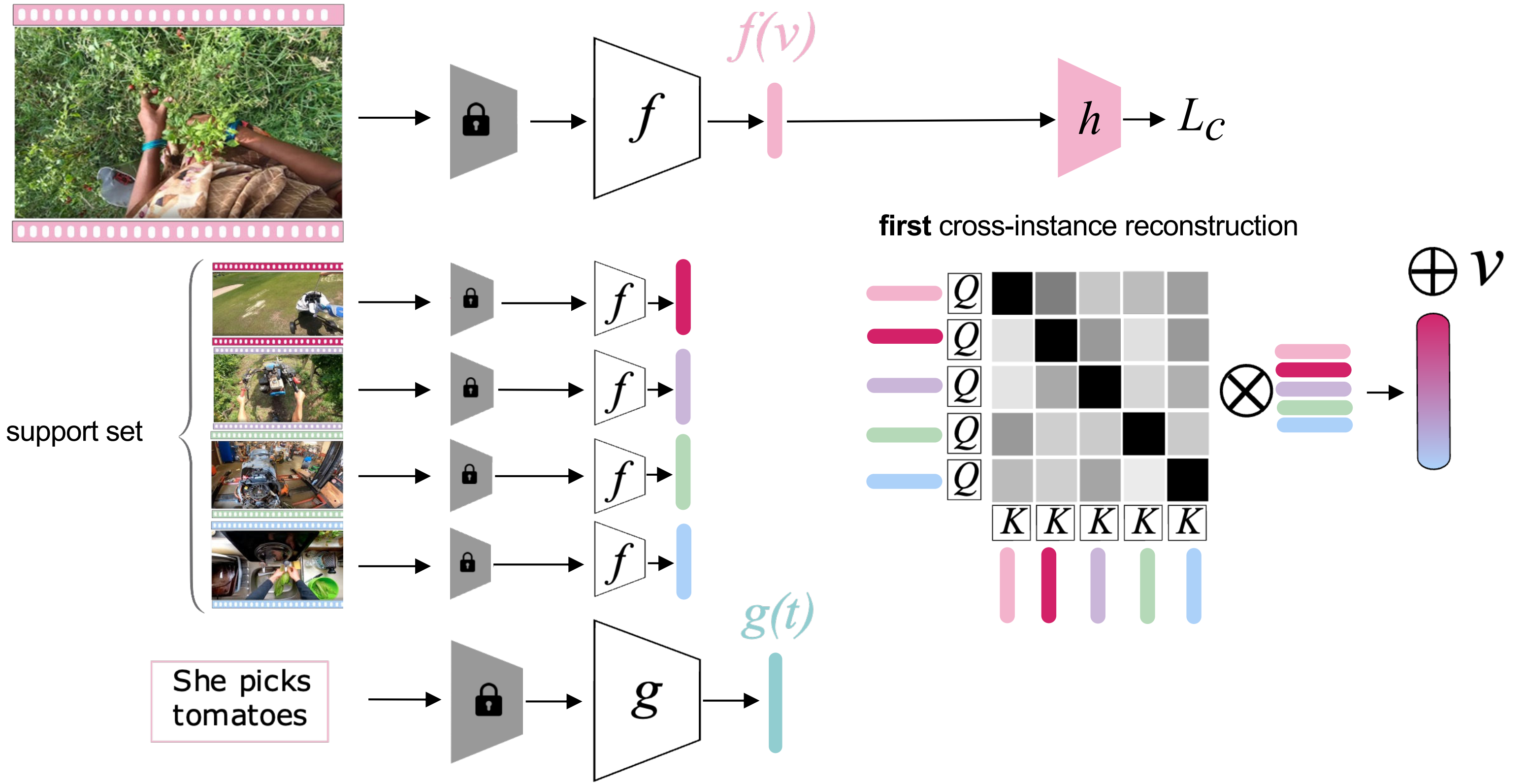
with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Proposed method: CIR

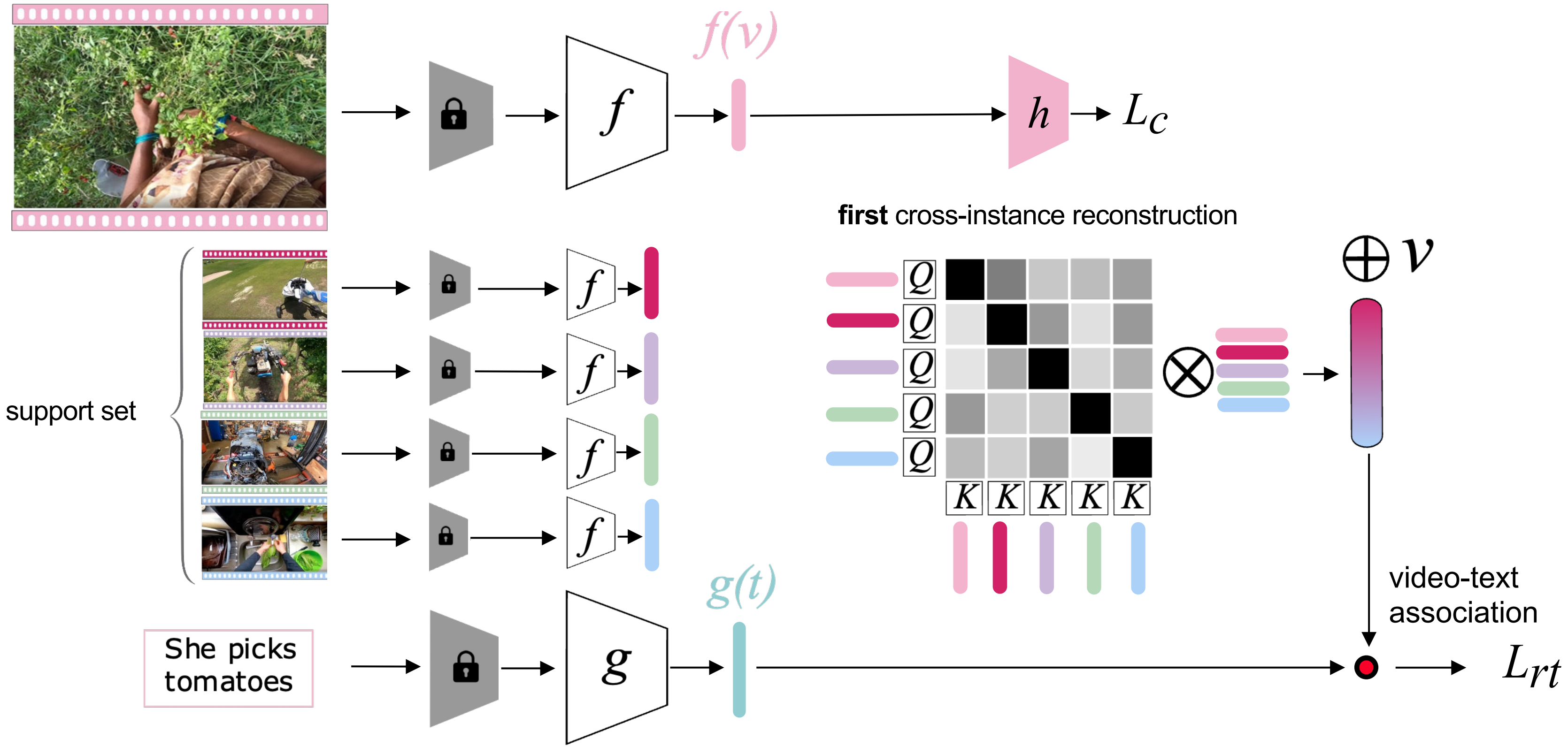
with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Proposed method: CIR

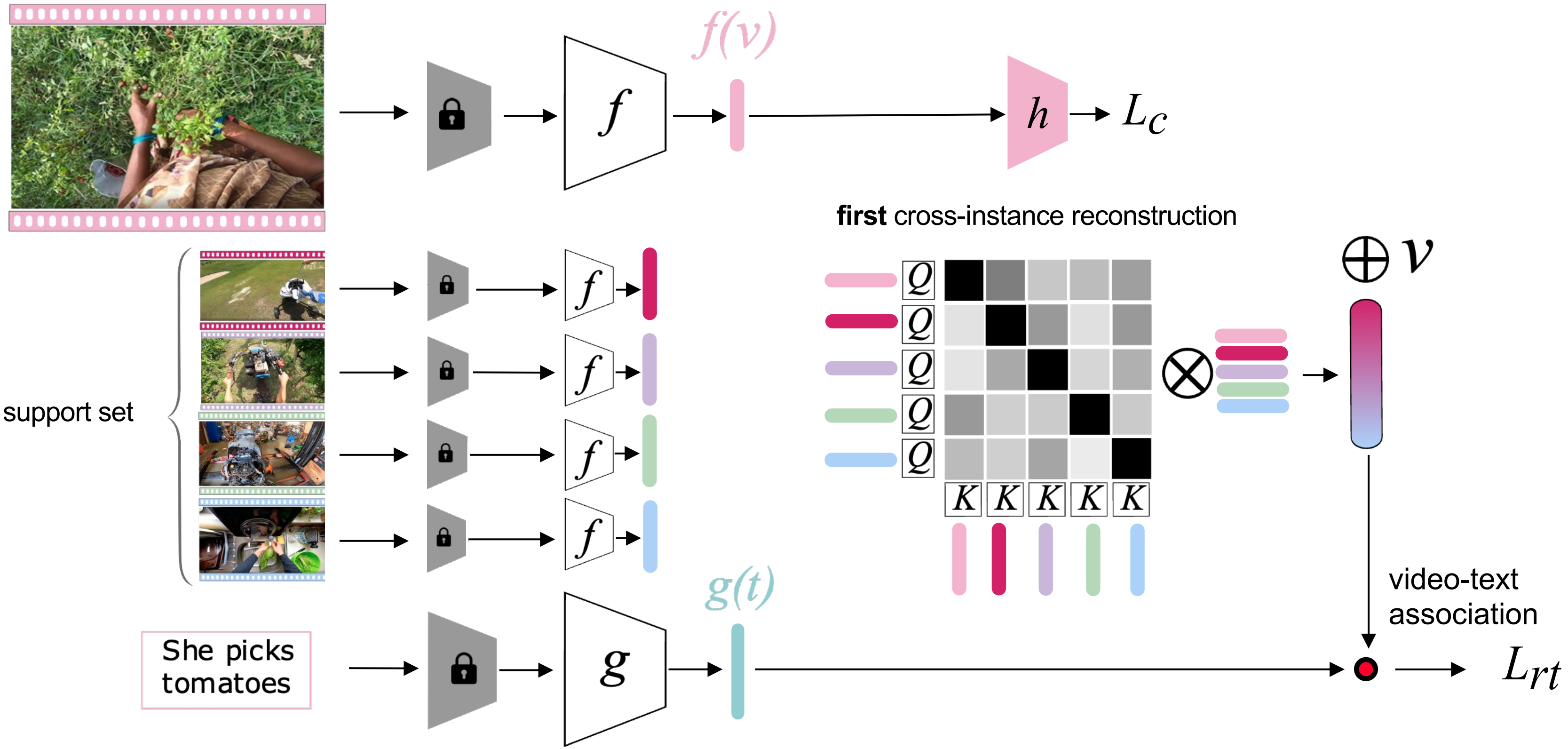
with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Proposed method: CIR

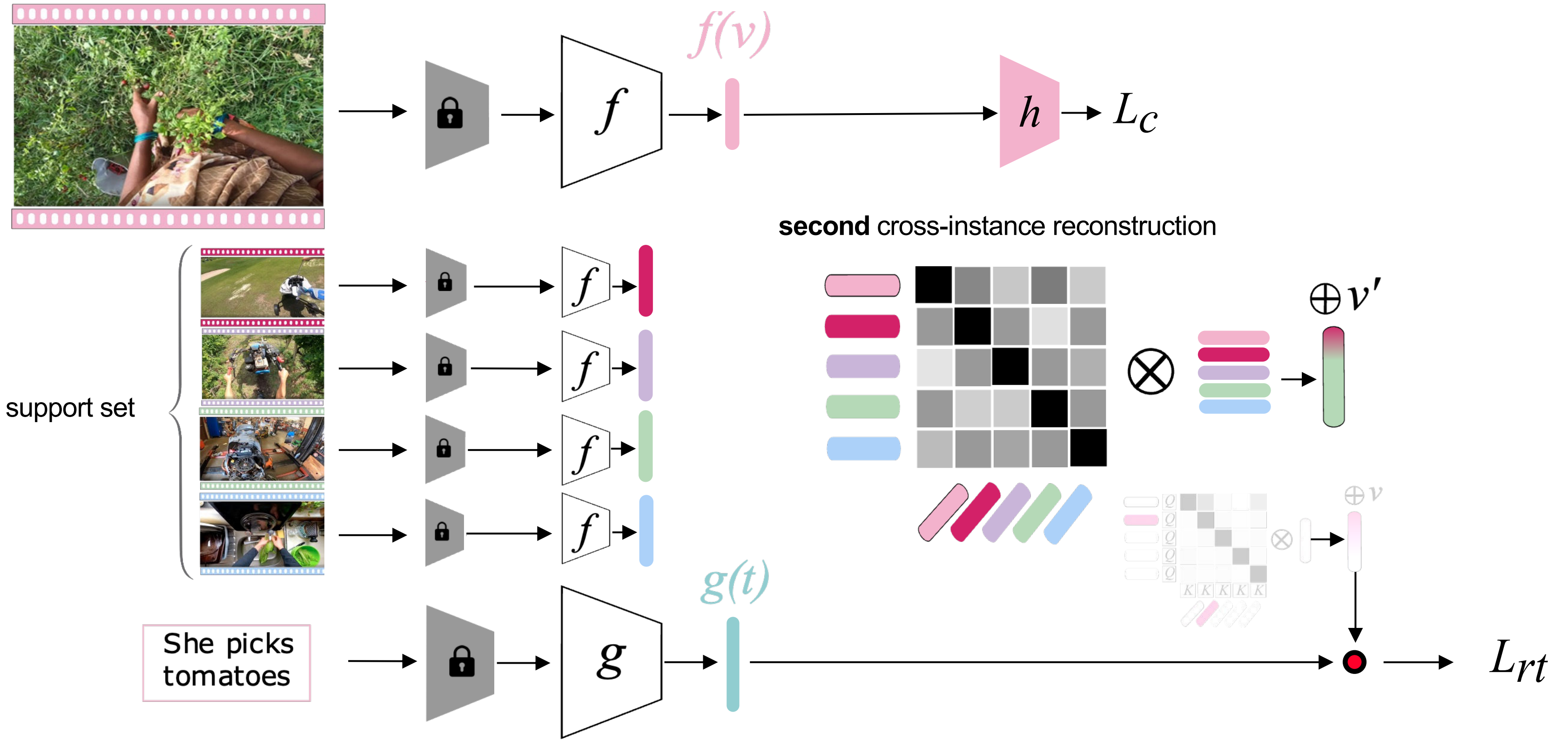
with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Proposed method: CIR

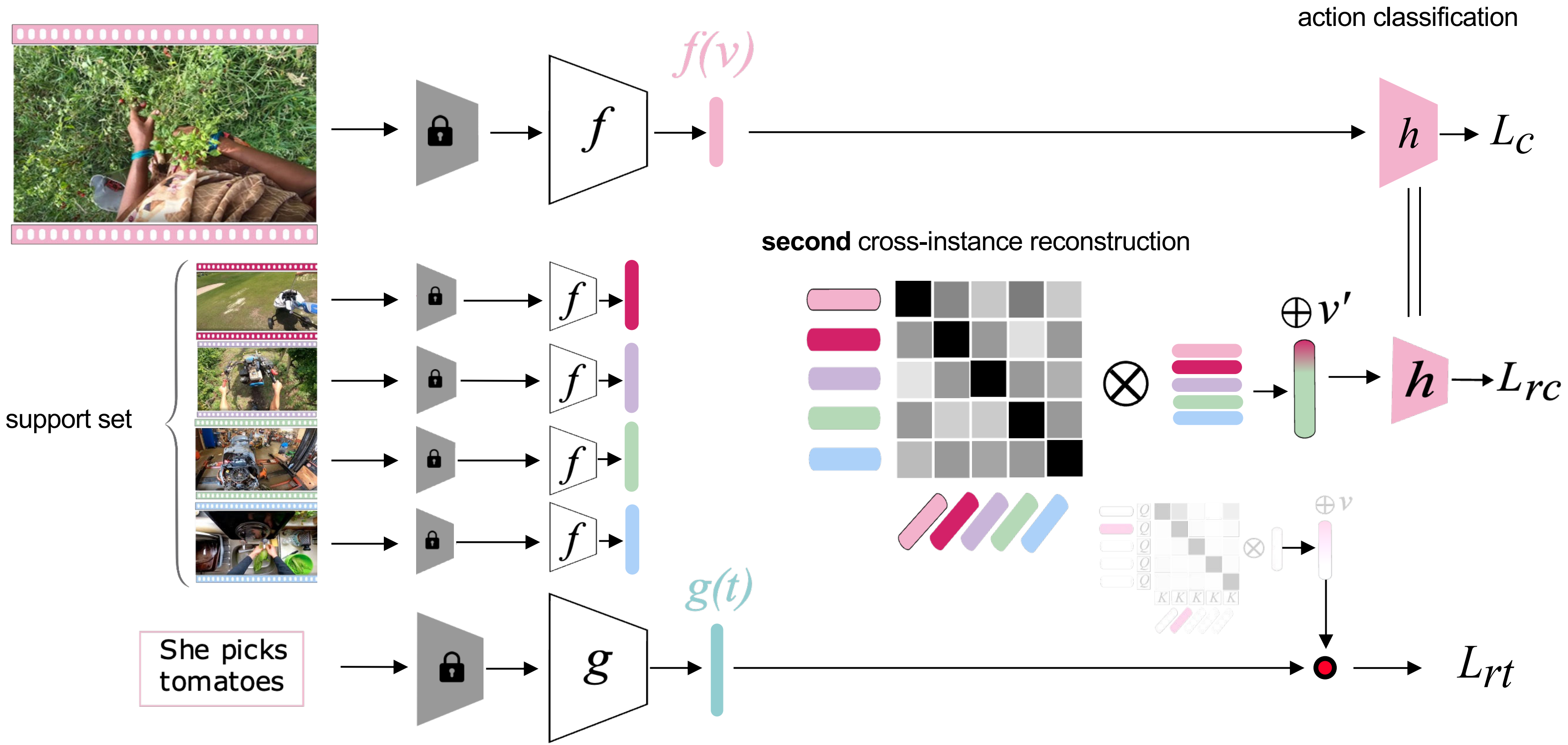
with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Proposed method: CIR

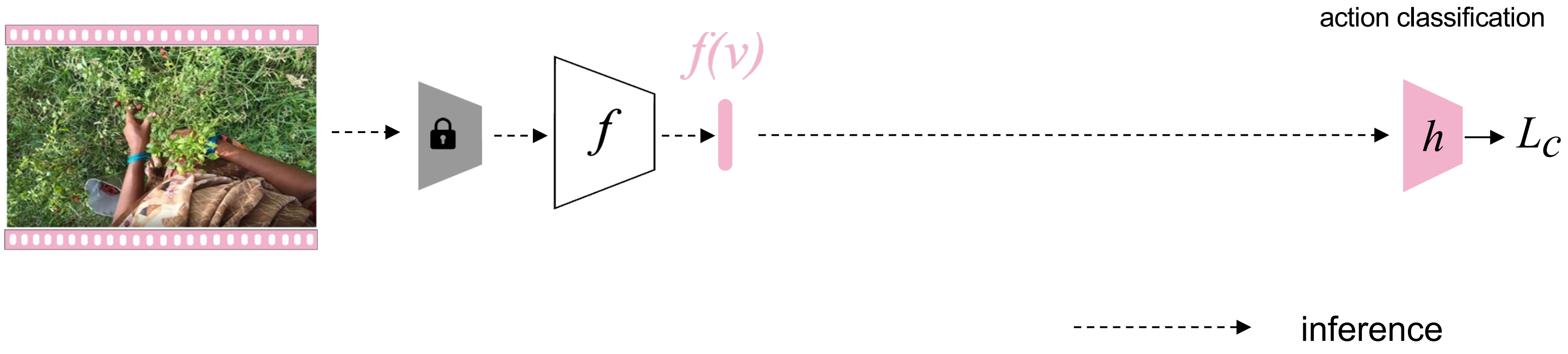
with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Proposed method: CIR

with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Chiara Plizzari
Toby Perrett
Dima Damen

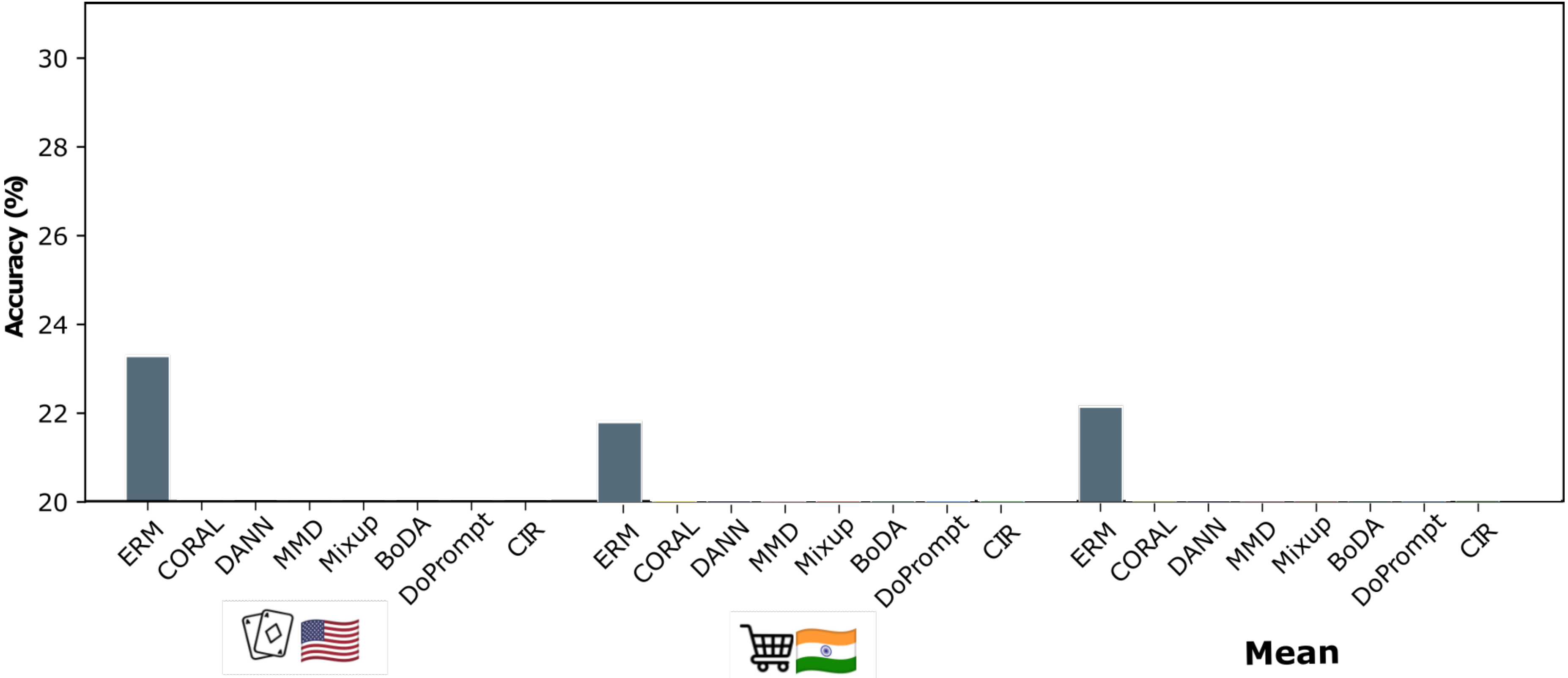
#C C drops the cut vegetables

query

support 1 support 2 support 3 support 4 support 5

Proposed method: CIR

with: Chiara Plizzari
Toby Perrett



Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

What can a cook in Italy teach a mechanic in India?

with: Chiara Plizzari
Toby Perrett

arXiv:2306.08713v1 [cs.CV] 14 Jun 2023

What can a cook in Italy teach a mechanic in India?
Action Recognition Generalisation Over Scenarios and Locations

Chiara Plizzari[♠*] Toby Perrett[♦] Barbara Caputo[♠] Dima Damen[♦]

[♠] Politecnico di Torino, Italy [♦] University of Bristol, United Kingdom

Abstract

We propose and address a new generalisation problem: can a model trained for action recognition successfully classify actions when they are performed within a previously unseen scenario and in a previously unseen location? To answer this question, we introduce the Action Recognition Generalisation Over scenarios and locations dataset (ARGO1M), which contains 1.1M video clips from the large-scale Ego4D dataset, across 10 scenarios and 13 locations. We demonstrate recognition models struggle to generalise over 10 proposed test splits, each of an unseen scenario in an unseen location. We thus propose CIR, a method to represent each video as a Cross-Instance Reconstruction of videos from other domains. Reconstructions are paired with text narrations to guide the learning of a domain generalisable representation. We provide extensive analysis and ablations on ARGO1M that show CIR outperforms prior domain generalisation works on all test splits. Code and data: `https://chiaraplizz.github.io/what-can-a-cook/`.

1. Introduction

A notable distinction between human and machine intelligence is the ability of humans to generalise. We can see an example of the action "cut" performed by a cook in Italy, and recognise the same action performed in a different geographic *location*, *e.g.* India, despite having never visited. We can also recognise actions within new *scenarios*, such as a mechanic cutting metal, even if we are unfamiliar with the tools they use.

This problem is known as domain generalisation [62], where a model trained on a set of labelled data fails to generalise to a different distribution in inference. The gap between distributions is known as *domain shift*. To date, works have focused on generalising over visual domain shifts [25, 46, 31, 10, 39]. In this paper, we introduce the *scenario shift*, where the same action is performed as part

*Work carried during Chiara's research visit to the University of Bristol

Figure 1: Problem statement and samples from the ARGO1M dataset. The same action, e.g. "cut", is performed differently based on the *scenario* and the *location* in which it is carried out. We aim to generalise so as to recognise the same action within a new scenario, *unseen* during training, and in an *unseen location, e.g., Mechanic* ( ) in *India* ( ).

of a different activity, impacting the tools used, objects interacted with, goals and behaviour. We combine this with the location shift, generalising over both simultaneously.

In Fig. 1, the action "cut" is performed using a knife whilst cooking ( ), pliers whilst building ( ) and scissors for arts and crafts ( ). Tools are not specific for a scenario and can vary over locations – e.g. in Fig. 1, seaweed sheets are cut with scissors while cooking in Japan. Generalising would be best achieved by learning the notion of "cutting" as separating an object into two or more pieces, regardless of the tool or background location. Successful generalisation can thus enable recognising metal being "cut" by a mechanic in India using an angle grinder (Fig. 1 Test).

Our investigation is enabled by the recent introduction of the Ego4D [17] dataset of egocentric footage from around the world. We curate a setup specifically for action generalisation, called ARGO1M. It contains 1.1M action clips of 60 classes from 73 unique scenario/location combinations.

To tackle the challenge of ARGO1M, we propose a new method for domain generalisation. We represent each video

ARGO1M-What-can-a-cook Public Watch 2

master 1 branch 0 tags Go to file Add file Code

Chiaraplizz Update download_all.py 5ecb693 4 days ago 10 commits

code	Update download_all.py	4 days ago
data	Initial commit	last week
data_csv	Initial commit	last week
resources	Initial commit	last week
LICENSE.md	Initial commit	last week
README.md	Update README.md	4 days ago
environment.yml	Initial commit	last week

ARGO1M Dataset
CIR Method
Code and Models

RELEASED

Plizzari et al (2023). What can a cook in Italy teach a mechanic in India? Action Recognition Generalisation Over Scenarios and Locations. IEEE/CVF International Conference on Computer Vision (ICCV).

Opportunities in Egocentric Vision

Tasks are harder

Detection, 3D Mapping, Tracking, VOS, Hand-Object, Generative, …

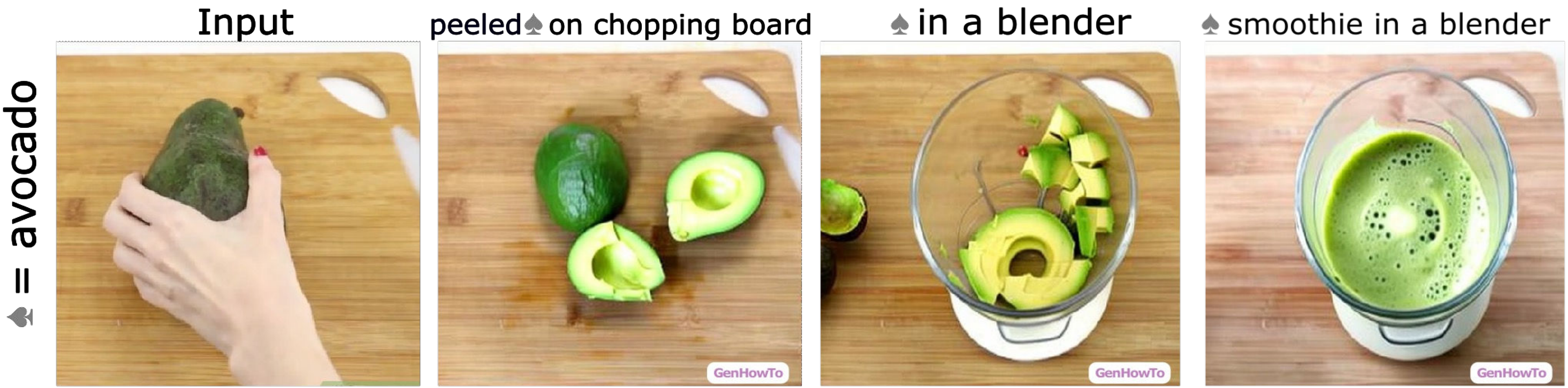
Solutions prove more rewarding

Weak supervision, Domain Adap/Gen., Audio-Visual, long-term understanding

GenHowTo…

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

- Hands transform objects….



T Soucek et al (2023). GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. ArXiv

GenHowTo…

with: Tomas Soucek, Ivan Laptev, Michael Wray, Josef Sivic

Input	GenHowTo	EF-DDPM	InstructPix2Pix

Prompt: a frosted cake with strawberries around the top

Prompt: a person kneading dough on a cutting board

Prompt: a person cutting a fish on a cutting board

GenHowTo…

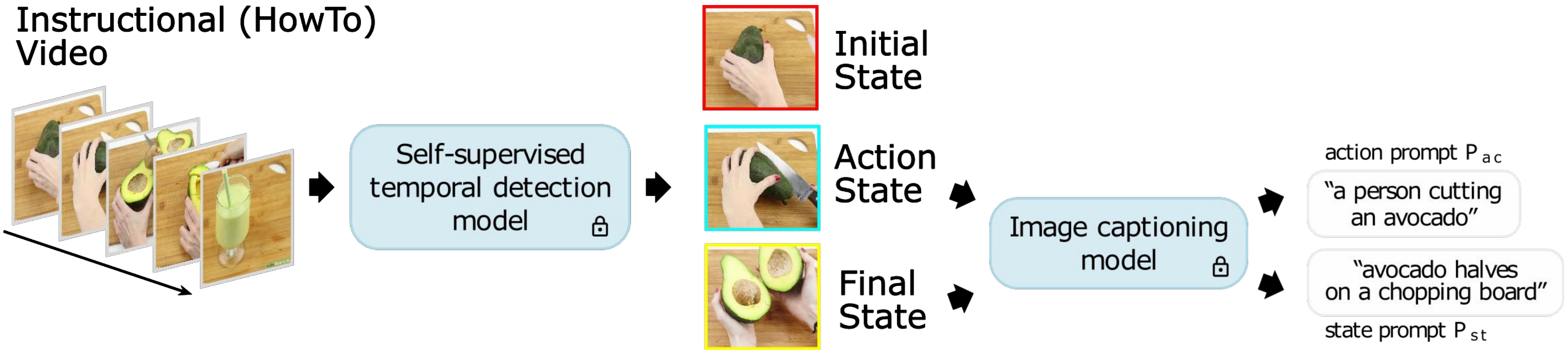
with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

- Two contributions…. Dataset & Method

T Soucek et al (2023). GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. ArXiv

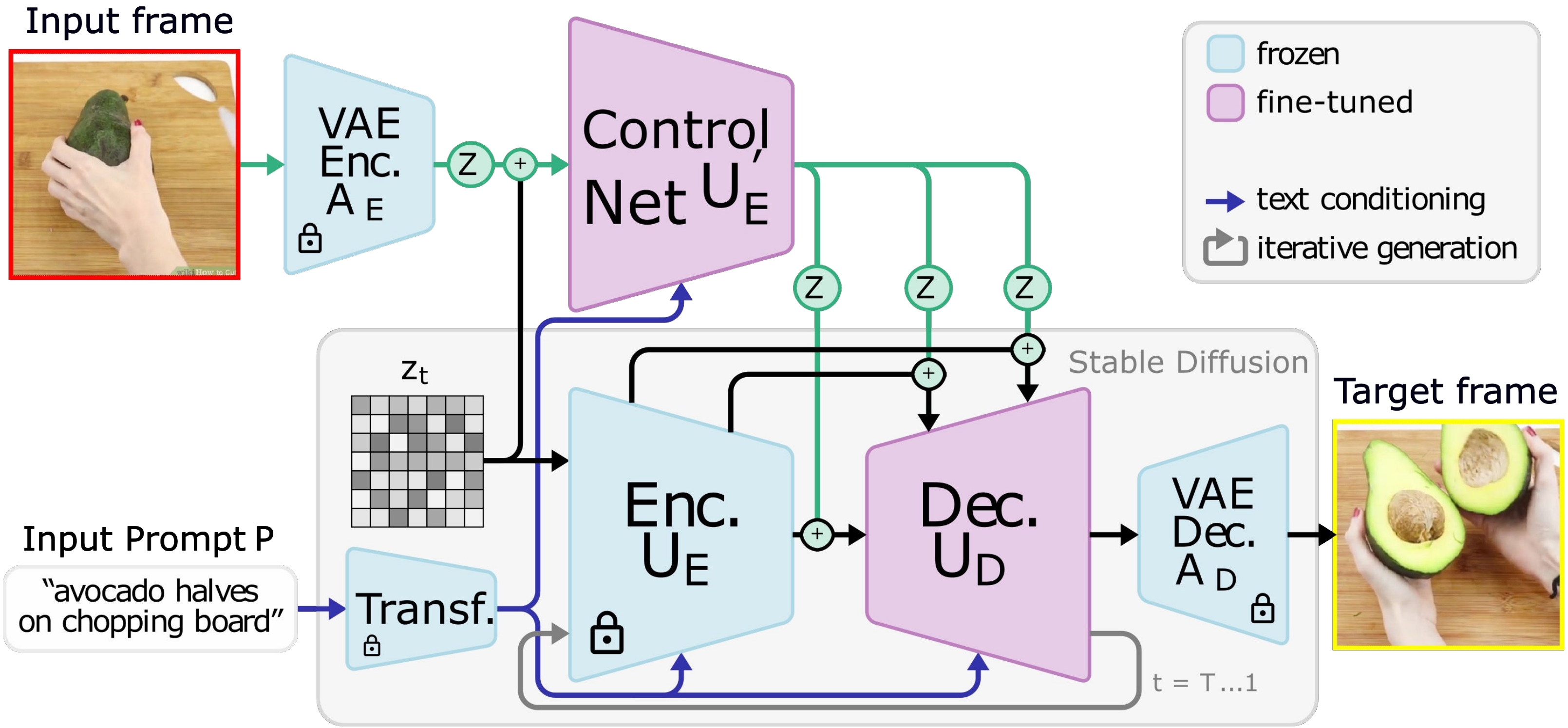
with: Tomas Soucek Ivan Laptev Michael Wray Josef Sivic

- Two contributions…. Dataset & Method



Tomas Soucek, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic (2022). Multi-task learning of object state changes from uncurated videos.

- Two contributions…. Dataset & Method



T Soucek et al (2023). GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. ArXiv

GenHowTo…

with: Tomas Soucek Michael Wray
Ivan Laptev Josef Sivic

Input *less noise* ⟷ *more noise*

University of BRISTOL

T Soucek et al (2023). GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. ArXiv

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

- Qualitative Evaluation…
  - Initial vs Final State
  - Binary Classifier

	Method	$\text{Acc}_{\text{ac}} \uparrow$	$\text{Acc}_{\text{st}} \uparrow$
	test set categories unseen during training		
(a)	Stable Diffusion	0.51	0.50
(b)	Edit Friendly DDPM	0.60	0.61
(c)	InstructPix2Pix	0.55	0.63
(d)	*CLIP (manual prompts)*	0.52	0.62
(e)	**GenHowTo**	**0.66**	**0.74**
	test set categories seen during training		
(f)	Edit Friendly DDPM†	0.69	0.80
(g)	**GenHowTo**†	**0.77**	**0.88**
(h)	*Real images*	0.96	0.97

† Models trained also on the test set *categories*.

T Soucek et al (2023). GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. ArXiv

GenHowTo…

with: Tomas Soucek Michael Wray
Ivan Laptev Josef Sivic

a person is wrapping a tortilla on a plate

REAL IMAGE —— GENERATED

a plate with two burritos on it

REAL IMAGE —— GENERATED

a man pouring beer into a glass

REAL IMAGE —— GENERATED

a man sitting at a table holding a glass of beer

REAL IMAGE —— GENERATED

T Soucek et al (2023). GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. ArXiv

Opportunities in Egocentric Vision

Tasks are harder

Detection, 3D Mapping, Tracking,
VOS, Hand-Object, Generative, …

Solutions prove more rewarding

Weak supervision, Domain Adap/Gen.,
Audio-Visual, long-term understanding


Videos are multi-modal

Multi-modal learning…

with: Vangelis Kazakos
Arsha Nagrani.
Andrew Zisserman
Jaesung Huh
Jacob Chalk

- The magic of audio-visual understanding…

- Object-Object interactions

Multi-modal learning…

with: Vangelis Kazakos
Arsha Nagrani.
Andrew Zisserman
Jaesung Huh
Jacob Chalk

- The magic of audio-visual understanding…
- Object-Object interactions
- Material sounds

Multi-modal learning…

with: Vangelis Kazakos
Arsha Nagrani.
Andrew Zisserman
Jaesung Huh
Jacob Chalk

- The magic of audio-visual understanding…
- Object-Object interactions
- Material sounds
- Sound-emitting objects

Harmonic vs Percussive

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

EPIC-KITCHENS

Harmonic Sounds

Percussive Sounds

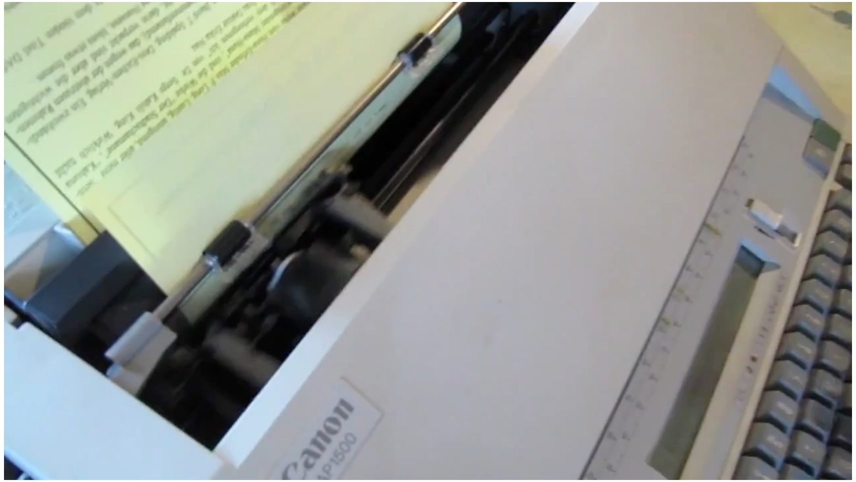
E Kazakos, A Nagrani, A Zisserman, D Damen (2021). Slow-Fast Auditory Streams For Audio Recognition. ICASSP

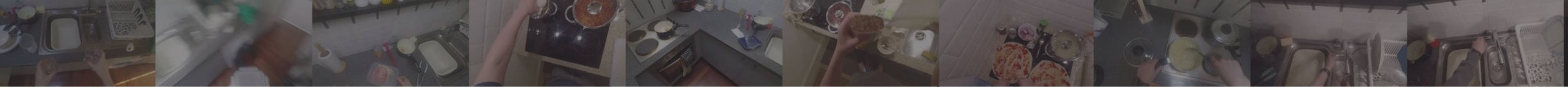
with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

VGG-Sound

Harmonic Sounds

Percussive Sounds





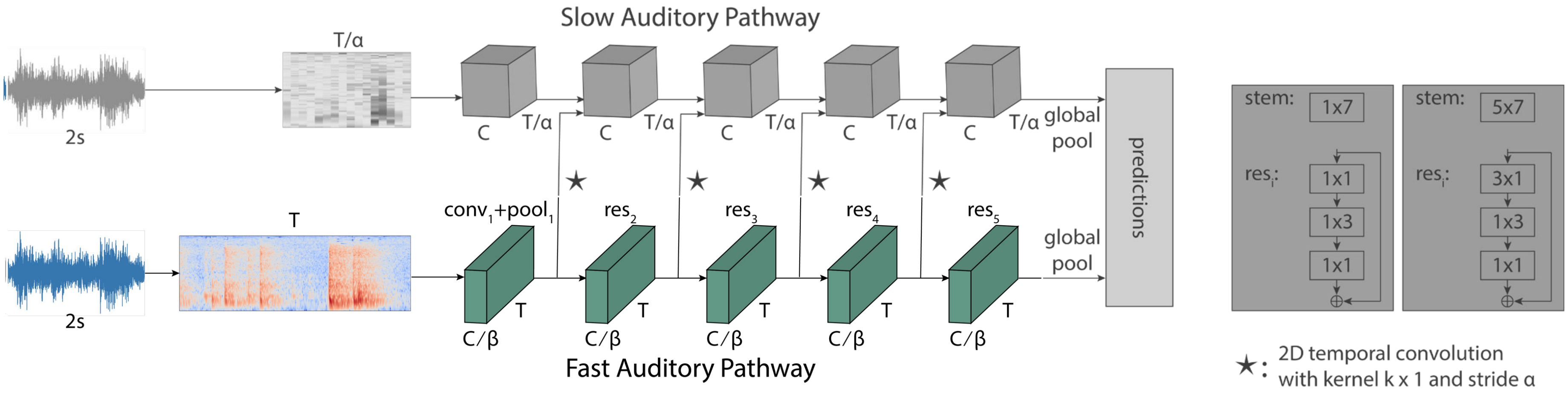
with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

Auditory Slow-Fast

Outstanding Paper Award – ICASSP 2021

Audio Slow-Fast

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



- Slow has low temporal precision and large amount of channels
- Fast has fewer channels but high temporal resolution
- Multi-level lateral connections
- Separable convolutions

Audio Slow-Fast

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

Slow stream

Animals

- baltimore oriole calling
- cheetah chirrup
- zebra braying
- dinosaurs bellowing
- horse neighing
- black capped chickadee calling
- cat hissing
- cuckoo bird calling
- mosquito buzzing
- bull bellowing
- whale calling

Scenes

- volcano explosion
- playing lacrosse
- hair dryer drying
- sea waves
- playing tympani
- blowtorch igniting
- opening/closing electric car windows
- thunder
- electric blender running
- playing shofar
- airplane flyby
- playing trumpet
- wind chime
- striking bowling

Fast stream

Percussive sounds

- footsteps on snow
- snake rattling
- tap dancing
- car engine knocking
- woodpecker pecking tree
- chopping wood
- people clapping
- lawn mowing
- typing on typewriter
- opening or closing car doors
- playing tennis
- railroad car
- playing tympani
- playing drum kit
- playing vibraphone
- popping pop corn

Voices

- singing choir
- people cheering
- people crowd
- child speech
- baby laughter

Others

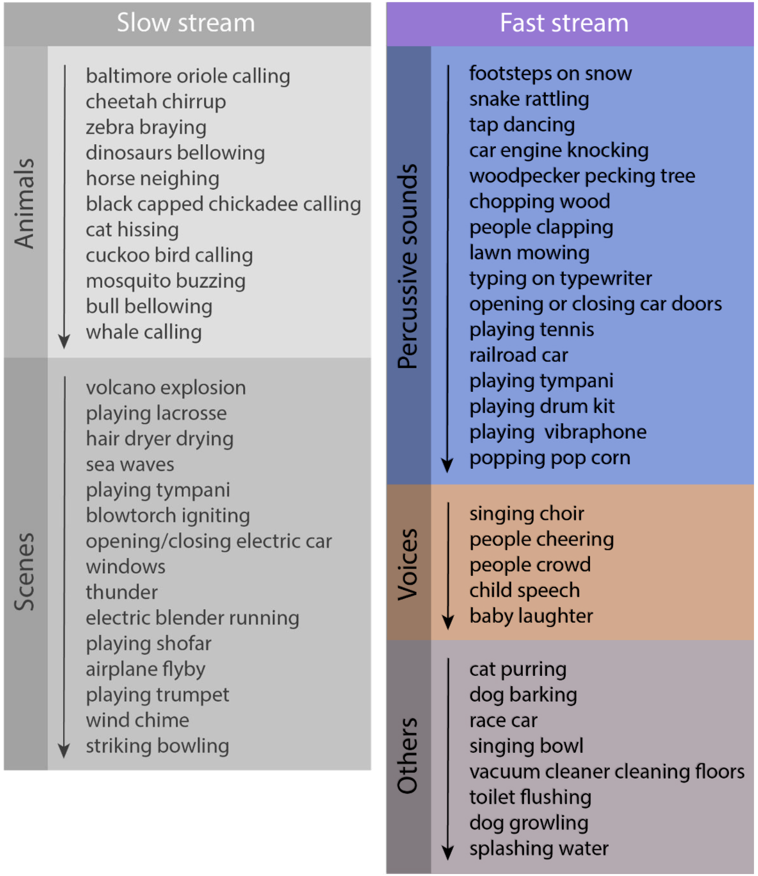
- cat purring
- dog barking
- race car
- singing bowl
- vacuum cleaner cleaning floors
- toilet flushing
- dog growling
- splashing water

University of BRISTOL

E Kazakos, A Nagrani, A Zisserman, D Damen (2021). Slow-Fast Auditory Streams For Audio Recognition. ICASSP

Audio Slow-Fast

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



E Kazakos, A Nagrani, A Zisserman, D Damen (2021). Slow-Fast Auditory Streams For Audio Recognition. ICASSP

Audio Slow-Fast

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

TOWARDS LEARNING UNIVERSAL AUDIO REPRESENTATIONS

Luyu Wang, Pauline Luc, Yan Wu, Adrià Recasens, Lucas Smaira, Andrew Brock, Andrew Jaegle,

Table 2: Evaluating frameworks and architectures on HARES. We compare the impact of architecture choice under the classification and SimCLR objective. We also show the performance of several other recent strongly performing frameworks. Average scores are reported for tasks in each domain separately, and all three combined. All models are trained on AudioSet except for bidirectional CPC and Wav2Vec2.0, for which we also show results when they are trained on LibriSpeech (LS).

Architecture	#Params	Input format	Used in	Env.	Speech	Music	HARES	AudioSet (mAP)
				Classification/SimCLR				
BYOL-A CNN	5.3m	Spectrogram	[9]	69.4/69.9	61.4/69.8	57.6/63.1	63.1/68.2	32.2/32.2
EfficientNet-B0	4.0m	Spectrogram	[8]	71.1/63.8	43.5/40.7	48.0/44.0	53.8/49.2	34.5/26.2
CNN14	71m	Spectrogram	[11, 13]	74.6/66.4	56.0/37.3	56.4/44.8	62.3/48.9	37.8/28.8
ViT-Base	86m	Spectrogram	[12]	73.3/74.6	50.4/56.5	60.3/64.2	60.5/64.5	36.8/36.8
ResNet50	23m	Spectrogram	[19]	74.8/74.4	51.7/65.0	59.6/63.7	61.4/67.8	38.4/36.2
SF ResNet50	26m	Spectrogram	[17]	74.0/74.3	56.9/73.4	59.6/65.2	63.3/71.7	37.2/36.6
NFNet-F0	68m	Spectrogram	Ours	**76.1**/76.0	59.0/65.9	61.8/65.5	65.4/69.2	**39.3**/37.6
SF NFNet-F0	63m	Spectrogram	Ours	75.2/75.8	65.6/**77.2**	64.5/**68.6**	68.5/**74.6**	38.2/37.8

achieve state-of-the-art performance across all domains.

Index Terms— audio representations, representation evaluation, speech, music, acoustic scenes

supervised contrastive learning [8, 13], and comparing them across a large set of model architectures. We find that models trained with contrastive learning tend to generalize better in the speech and music domain, while performing comparably to supervised pretraining for environment sounds. We

University of BRISTOL

E Kazakos, A Nagrani, A Zisserman, D Damen (2021). Slow-Fast Auditory Streams For Audio Recognition. ICASSP

ICASSP 2023
4 - 10 JUNE,
RHODES ISLAND, GREECE

EPIC-Sounds: A Large-scale Dataset of Actions That Sound

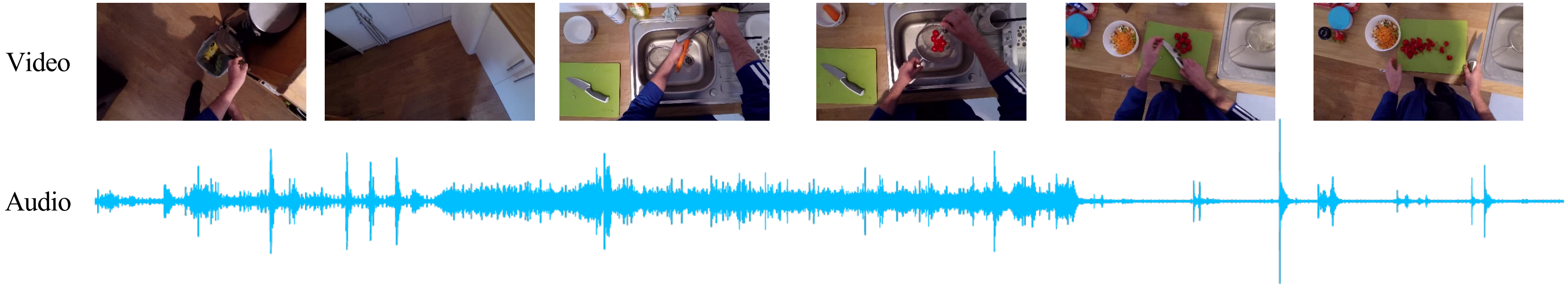
Jaesung Huh*, Jacob Chalk*, Evangelos Kazakos, Dima Damen, Andrew Zisserman
* : Equal contribution

VGG
University of Oxford

University of BRISTOL

Motivation

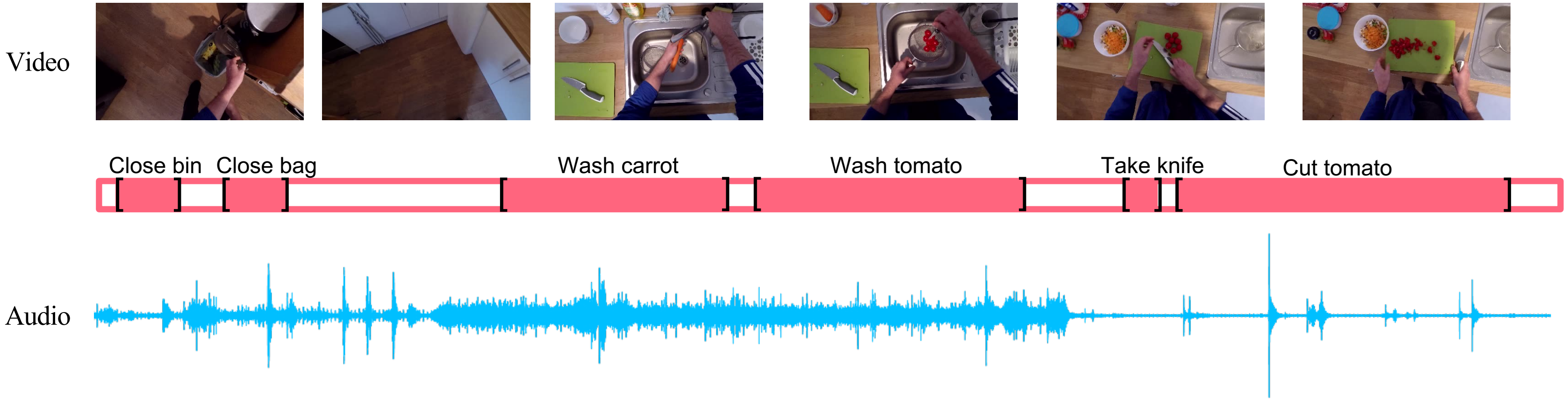
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Motivation

with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

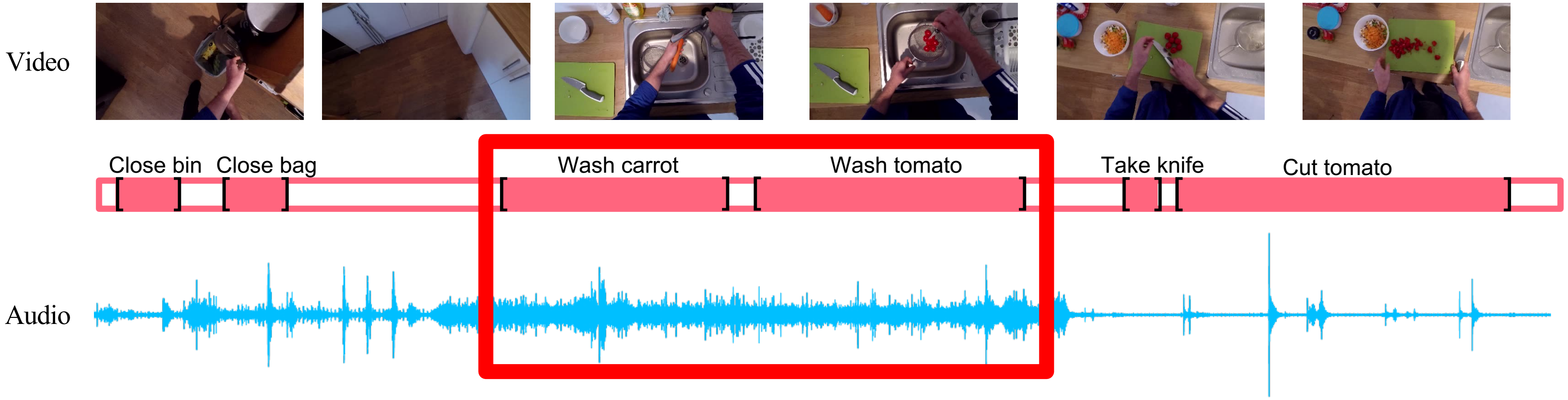
Motivation

with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman

J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Motivation

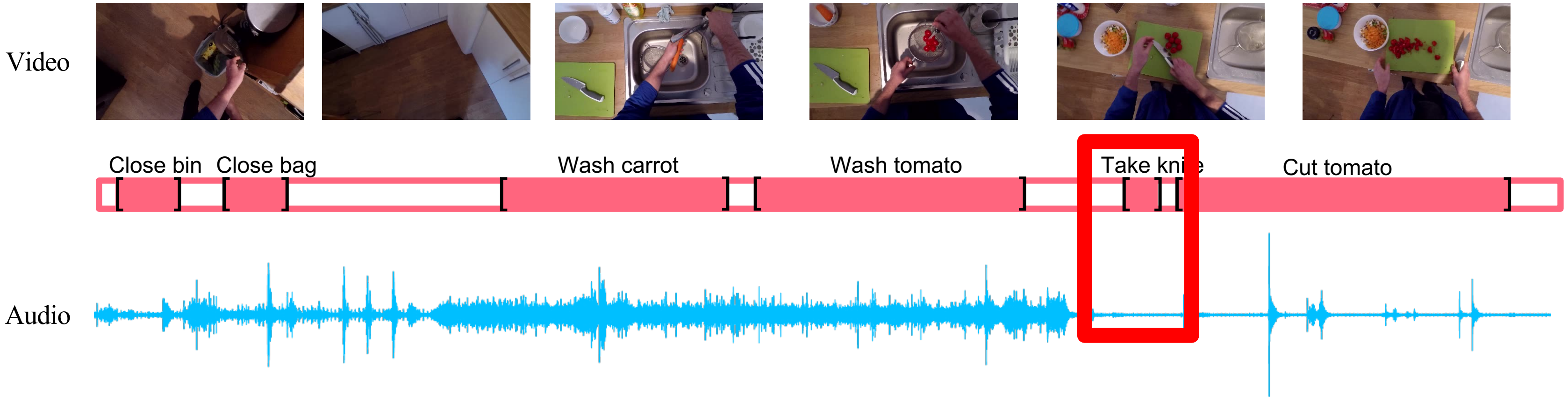
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Motivation

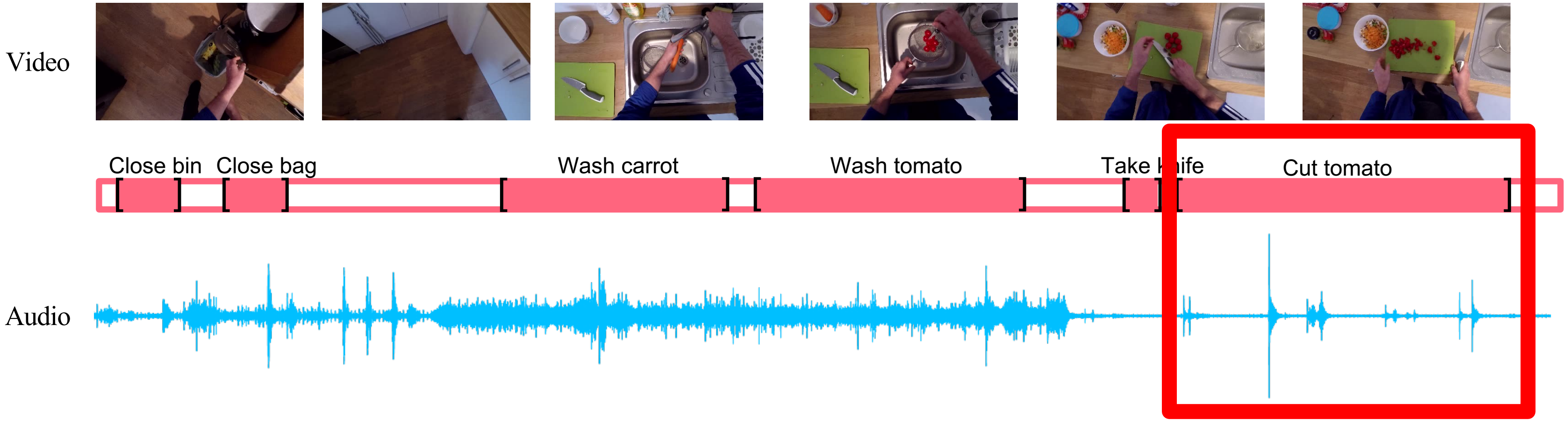
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Motivation

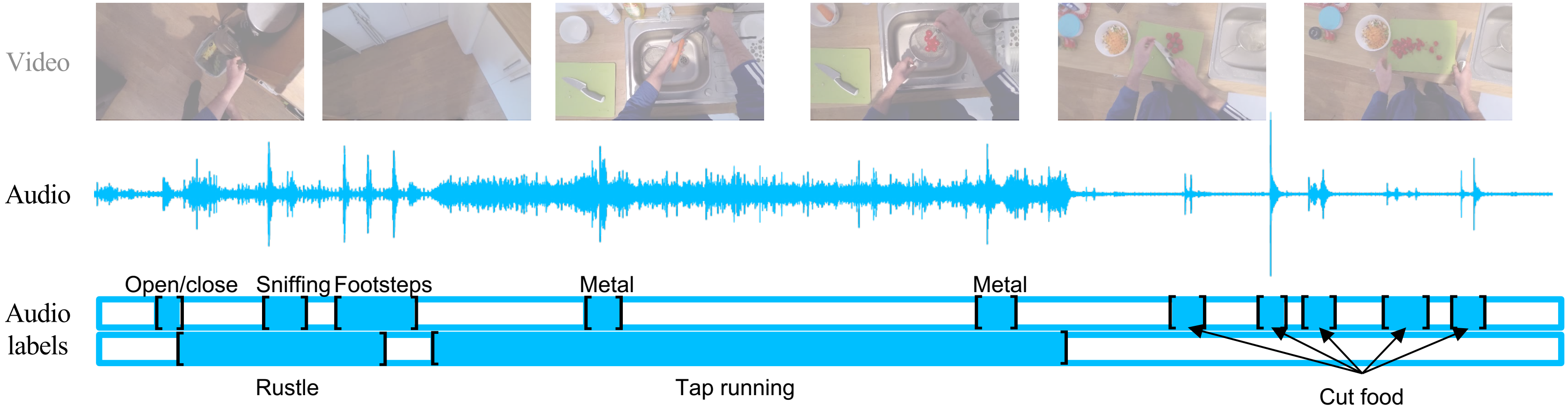
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Motivation

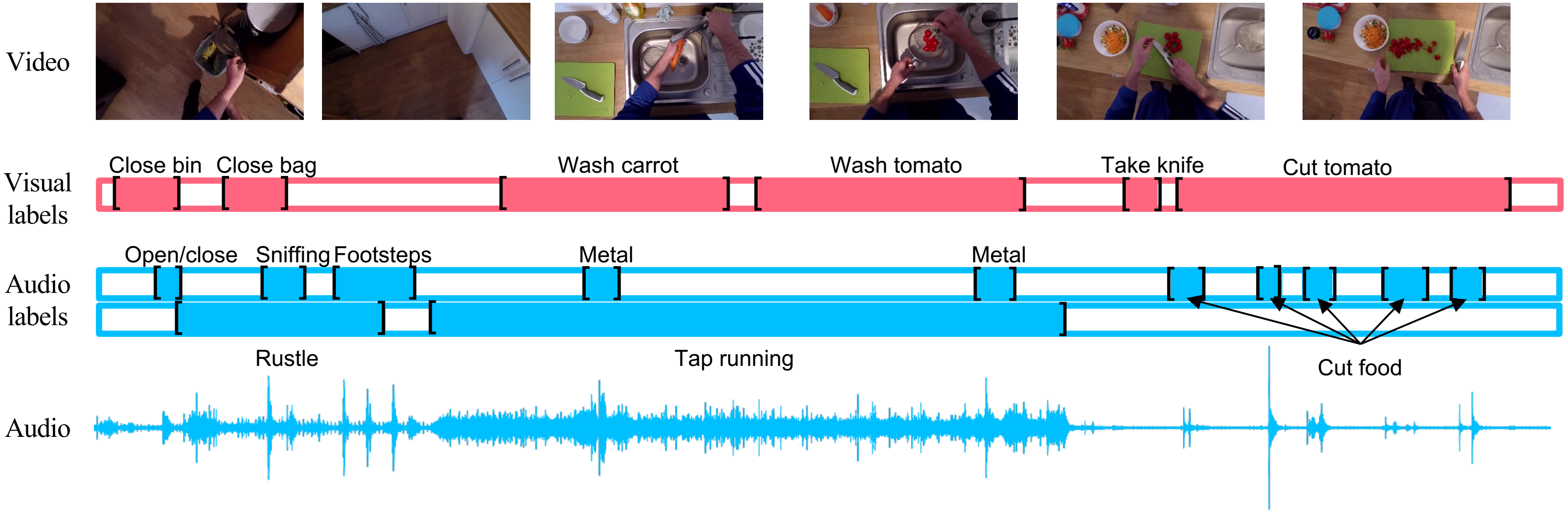
with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Motivation

with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman



J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

EPIC-SOUNDS

with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman

EPIC-KITCHENS VIDEOS
100 hours
45 kitchens

Visual Action Annotations
90K visual actions
97 verb classes
300 noun classes

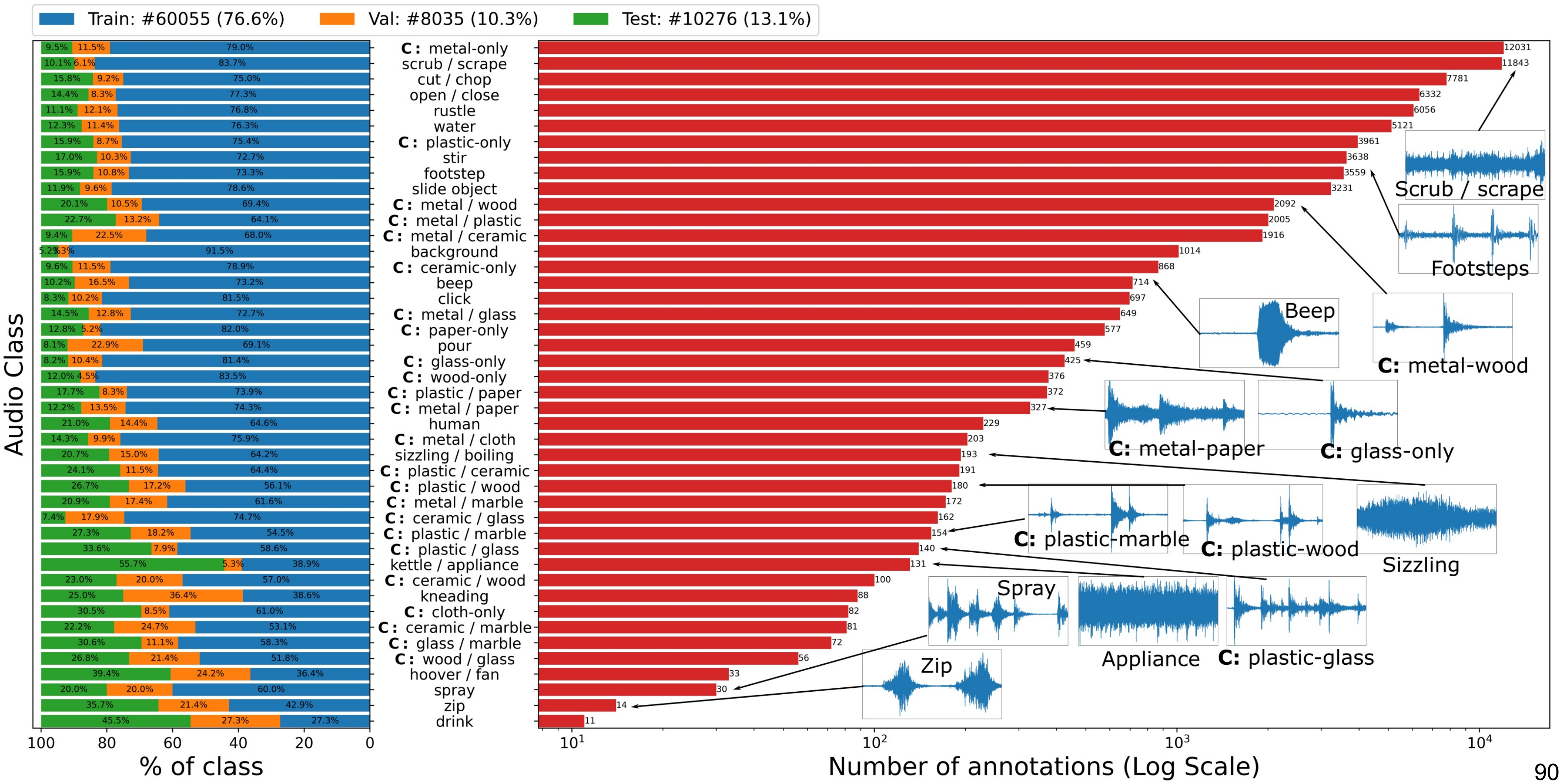
EPIC-Sounds
Audio-Based Annotations
79K categorised audio events
44 sound categories
39K uncategorised events

J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP


spray

EPIC-SOUNDS

with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman

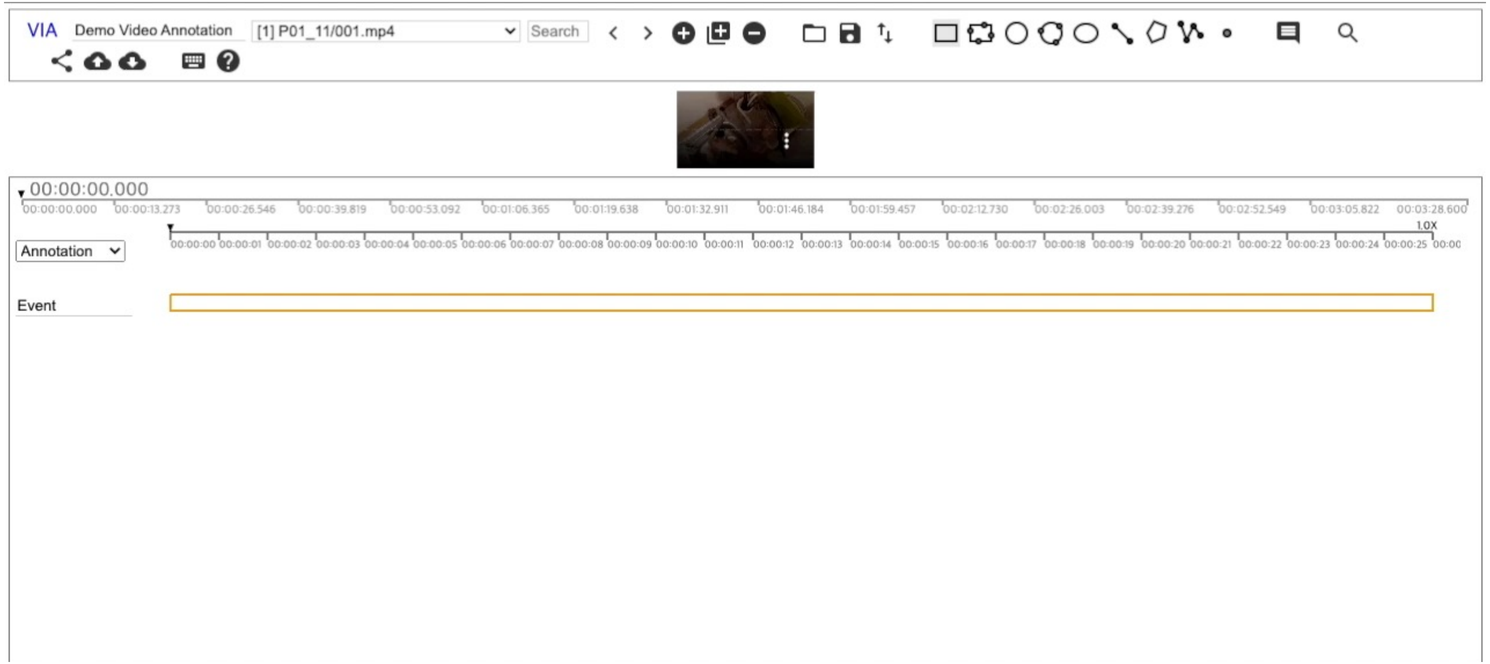


WACV2024 – Waikoloa, Hawaii

Annotations Pipeline

with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman

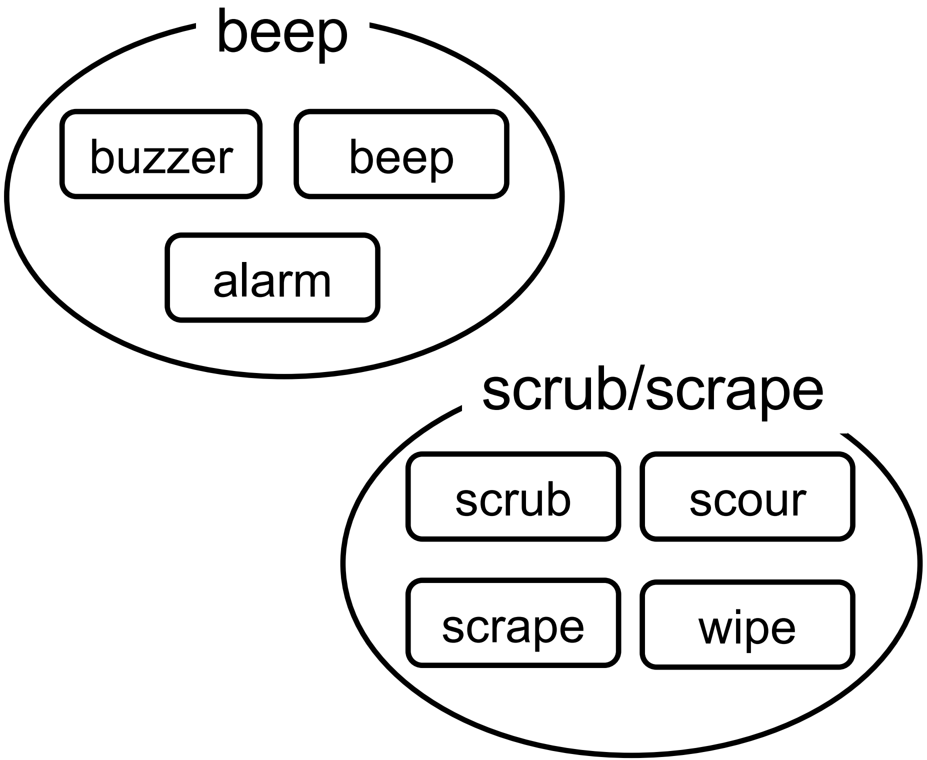
- We annotate all the distinctive sound events which consist of temporal intervals using free-form sound descriptions.
- Using VGG VIA annotation tool



Post Processing

with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman

- From free-form descriptions to categories



J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Collision Sounds

with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman

- For collision sounds, we annotate the **materials** of the objects that colliding.
- Materials example

Ceramic

Cloth

Metal

Plastic

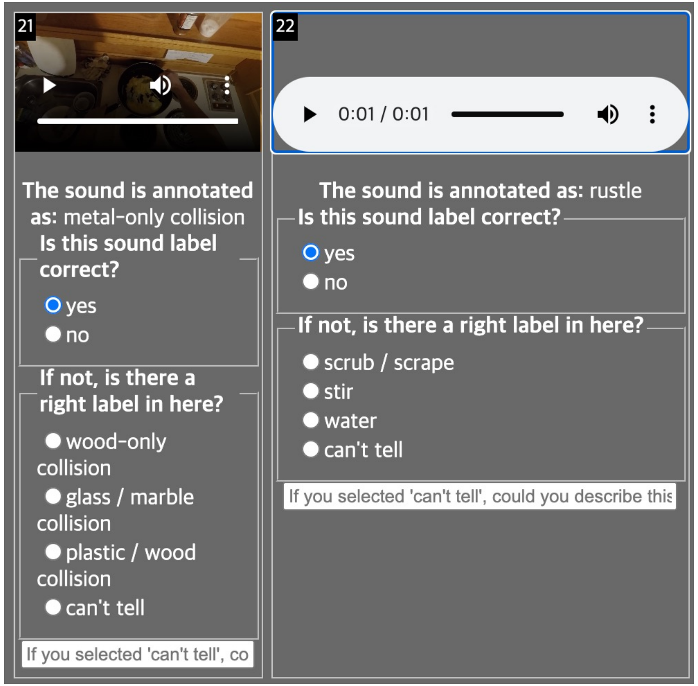
Glass

J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

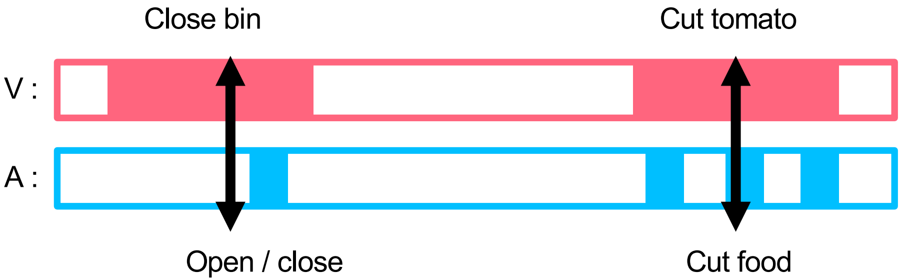
Post Processing

with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman

- Manual check on validation / test set



- We use the overlaps between audio and visual segments for reviewing train set.



EPIC-SOUNDS

with: Jaesung Huh* & Jacob Chalk*
Vangelis Kazakos Andrew Zisserman

EPIC-KITCHENS VIDEOS
100 hours
45 kitchens

Visual Action Annotations
90K visual actions
97 verb classes
300 noun classes

EPIC-Sounds
Audio-Based Annotations
79K categorised audio events
44 sound categories
39K uncategorised events

J Huh*, J Chalk* E Kazakos, D Damen, A Zisserman (2023). EPIC-Sounds: A Large-scale Dataset of Actions That Sound. ICASSP

Search or jump to... / Pull requests Issues Codespaces Marketplace Explore

epic-kitchens / epic-sounds-annotations Public

Edit Pins | Unwatch 5 | Fork 3 | Starred 47

Code | Issues 1 | Pull requests | Actions | Projects | Wiki | Security | Insights | Settings

111 lines (91 sloc) | 10.3 KB

Raw | Blame

EPIC-SOUNDS Dataset

We introduce EPIC-SOUNDS, a large scale dataset of audio annotations capturing temporal extents and class labels within the audio stream of the egocentric videos from EPIC-KITCHENS-100. EPIC-SOUNDS includes 78.4k categorised and 39.2k non-categorised segments of audible events and actions, distributed across 44 classes. In this repository, we provide labelled temporal timestamps for the train / val split, and just the timestamps for the recognition test split. We also provided the temporal timestamps for annotations that could not be clustered into one of our 44 classes, along with the free-form description used during the initial annotation. We train and evaluate two state-of-the-art audio recognition models on our dataset, which we also provide the code and pretrained models for.

Download the Data

A download script is provided for the videos here. You will have to extract the untrimmed audios from these videos. Instructions on how to extract and format the audio into a HDF5 dataset can be found on the Auditory SlowFast GitHub repo. Alternatively, you can email uob-epic-kitchens@bristol.ac.uk for access to an existing HDF5 file.

Contact: uob-epic-kitchens@bristol.ac.uk

Citing

When using the dataset, kindly reference our ICASSP 2023 Paper:

Opportunities in Egocentric Vision

Tasks are harder

Detection, 3D Mapping, Tracking, VOS, Hand-Object, Generative, …

Solutions prove more rewarding

Weak supervision, Domain Adap/Gen., Audio-Visual, long-term understanding

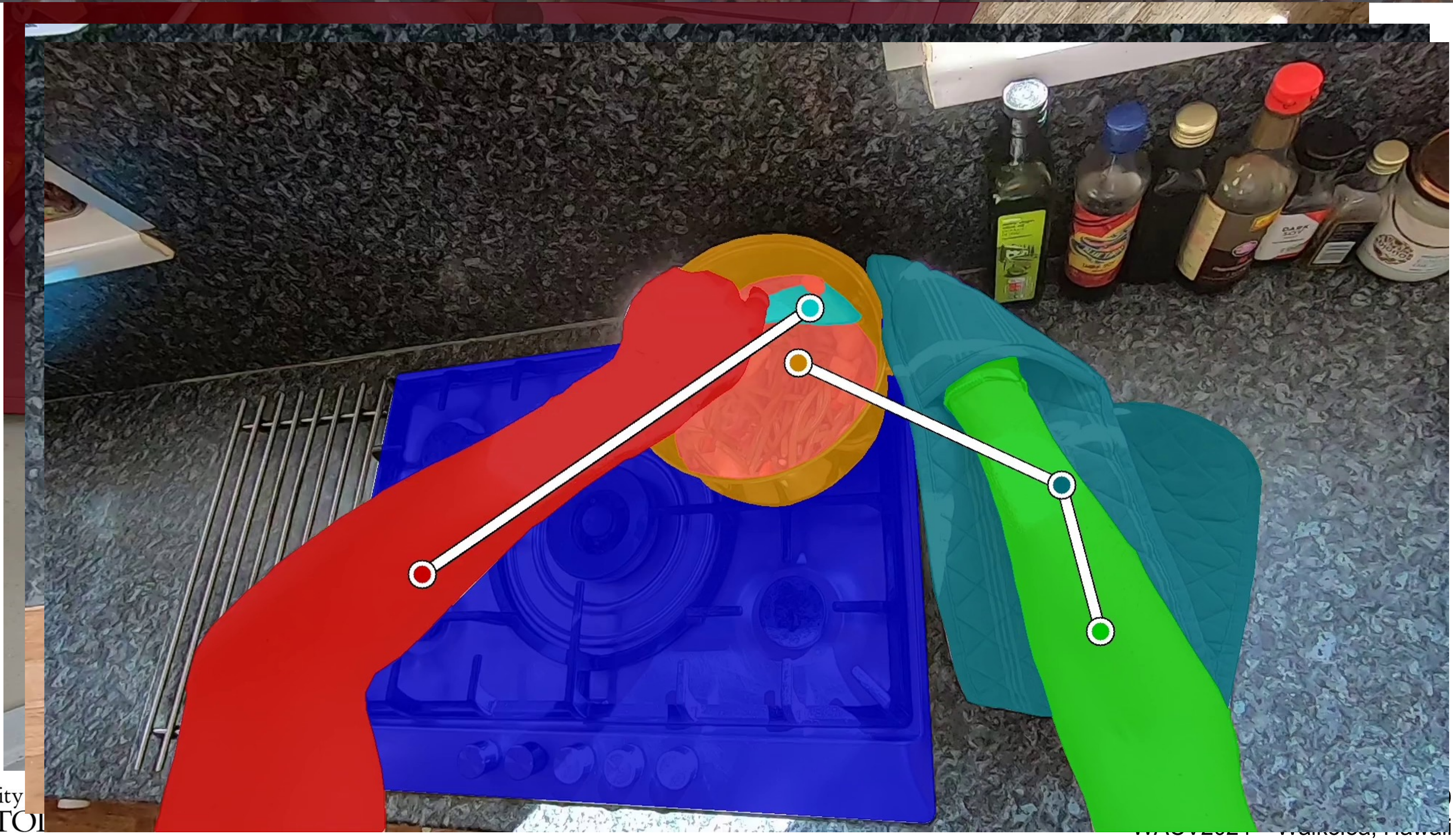
EPIC-KITCHENS VISOR

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma,
Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

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

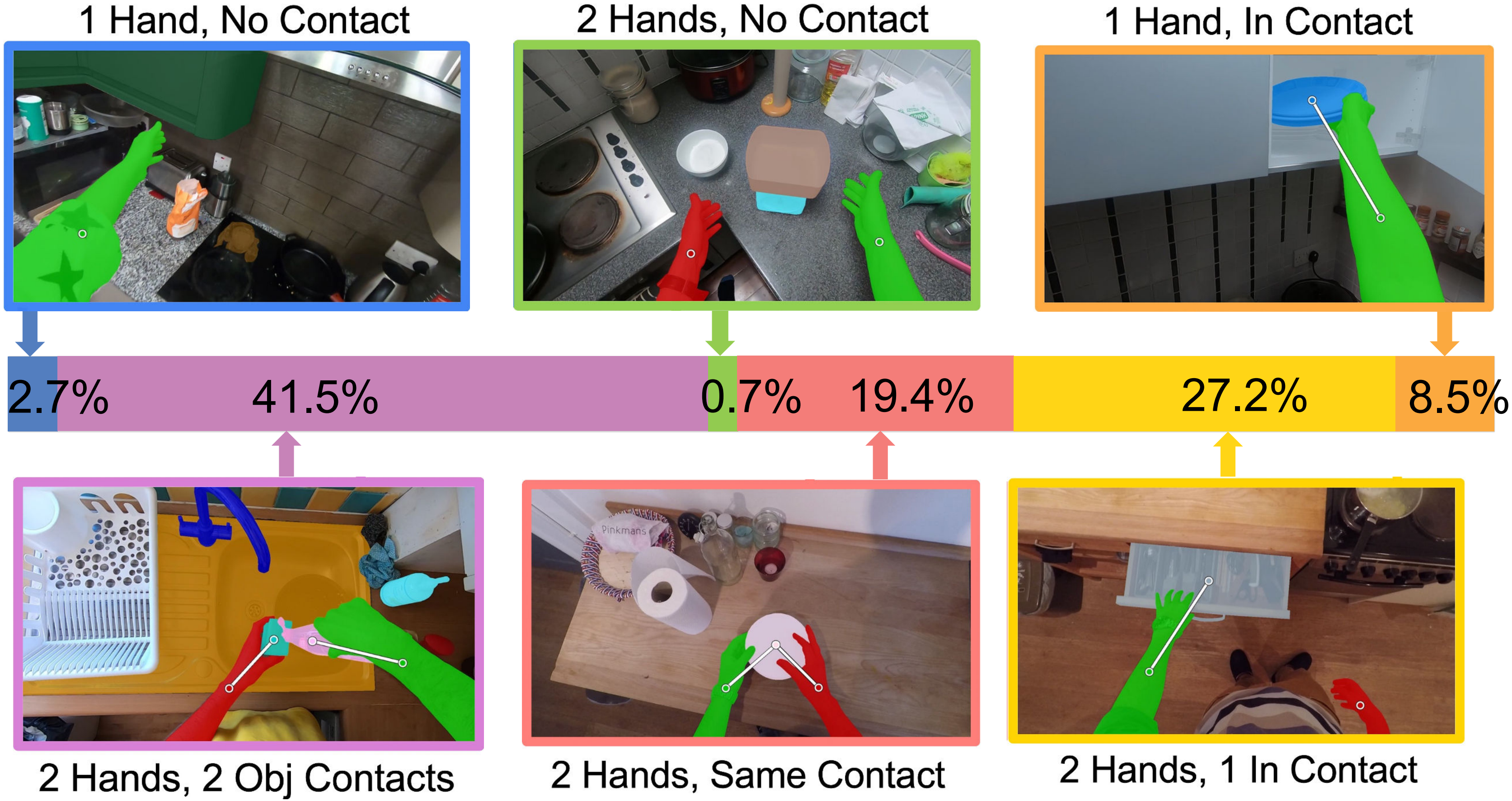
VISOR Relations

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler, Dima Damen



Object relation stats

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler, Dima Damen



University of BRISTOL

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

EPIC-KITCHENS VISOR

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma,
Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

EPIC-KITCHENS VISOR

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma,
Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

University of BRISTOL

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

Opportunities in Egocentric Vision

Tasks are harder

Detection, 3D Mapping, Tracking, VOS, Hand-Object, Generative, …

Solutions prove more rewarding

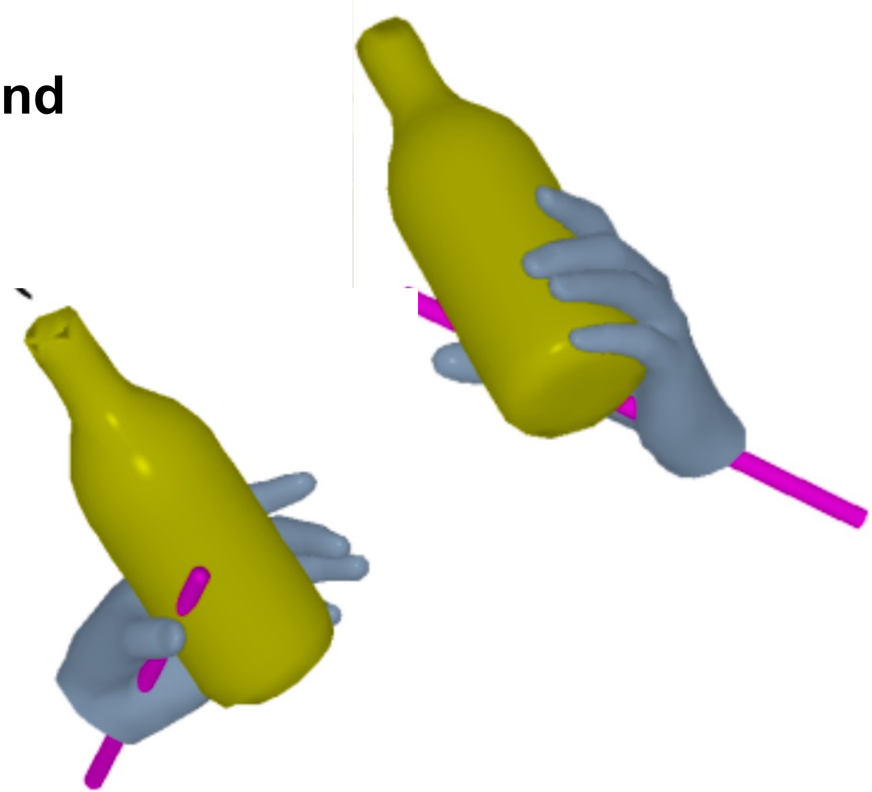
Weak supervision, Domain Adap/Gen., Audio-Visual, long-term understanding

Get a Grip

with: Zhifan Zhu

left hand

bottle



Z Zhu and D Damen (2023). Get a Grip: Reconstructing Hand-Object Stable Grasps in Egocentric Videos. *ArXiv*

Get a Grip

with: Zhifan Zhu

Non-Ego Views

Ego Views

Z Zhu and D Damen (2023). Get a Grip: Reconstructing Hand-Object Stable Grasps in Egocentric Videos. *ArXiv*

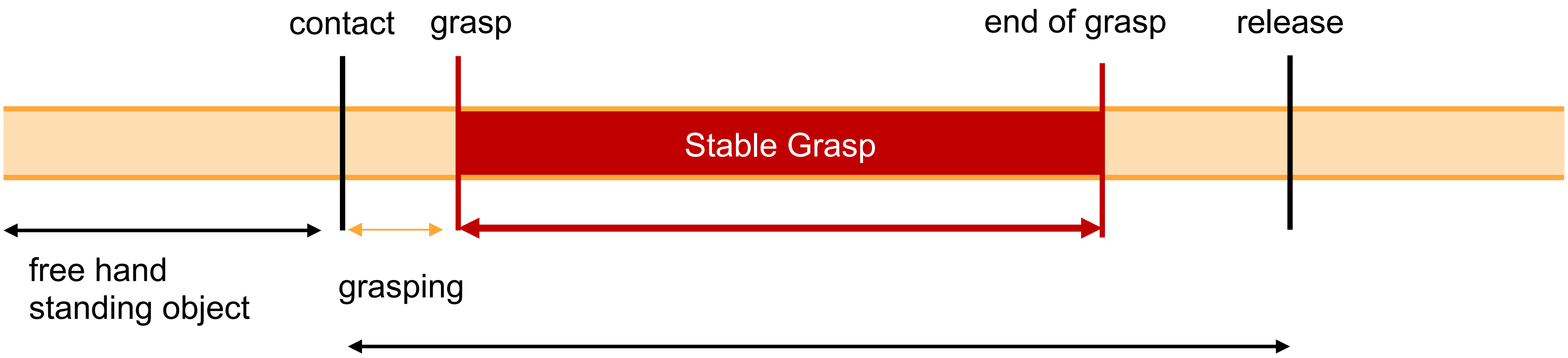
Get a Grip

with: Zhifan Zhu

Z Zhu and D Damen (2023). Get a Grip: Reconstructing Hand-Object Stable Grasps in Egocentric Videos. *ArXiv*

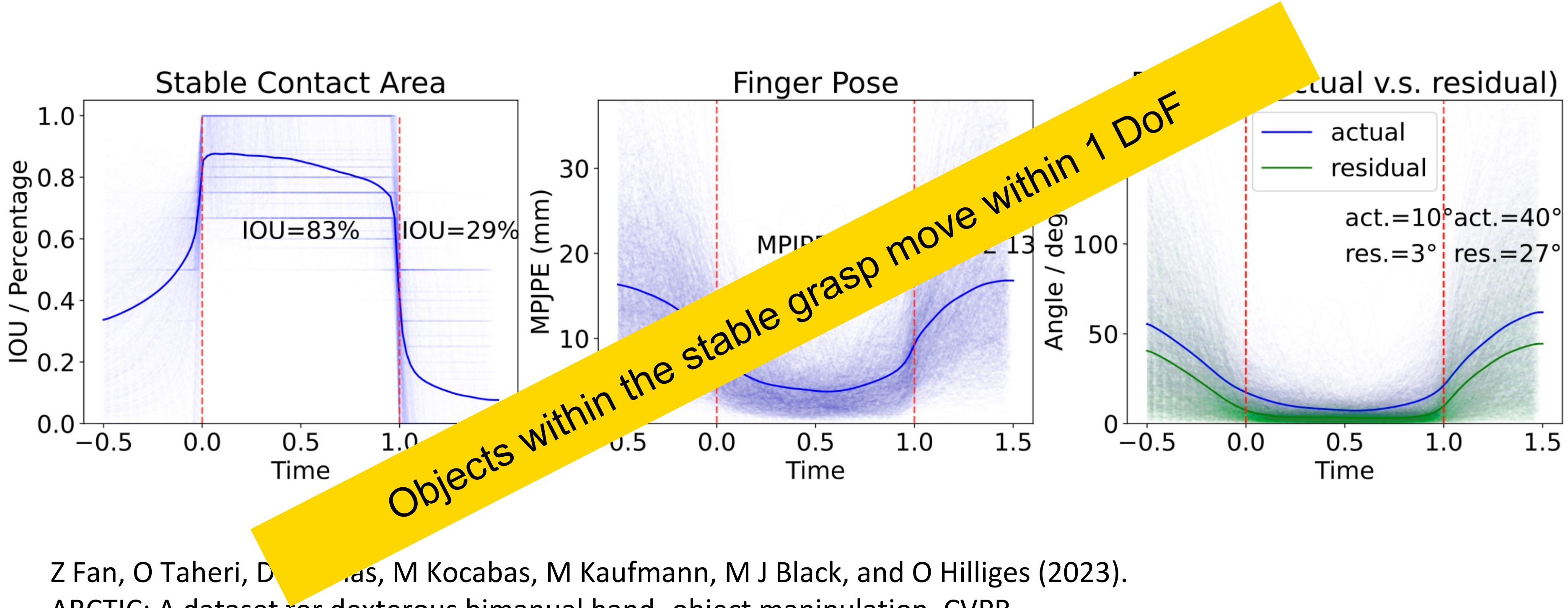
Get a Grip

with: Zhifan Zhu



Get a Grip

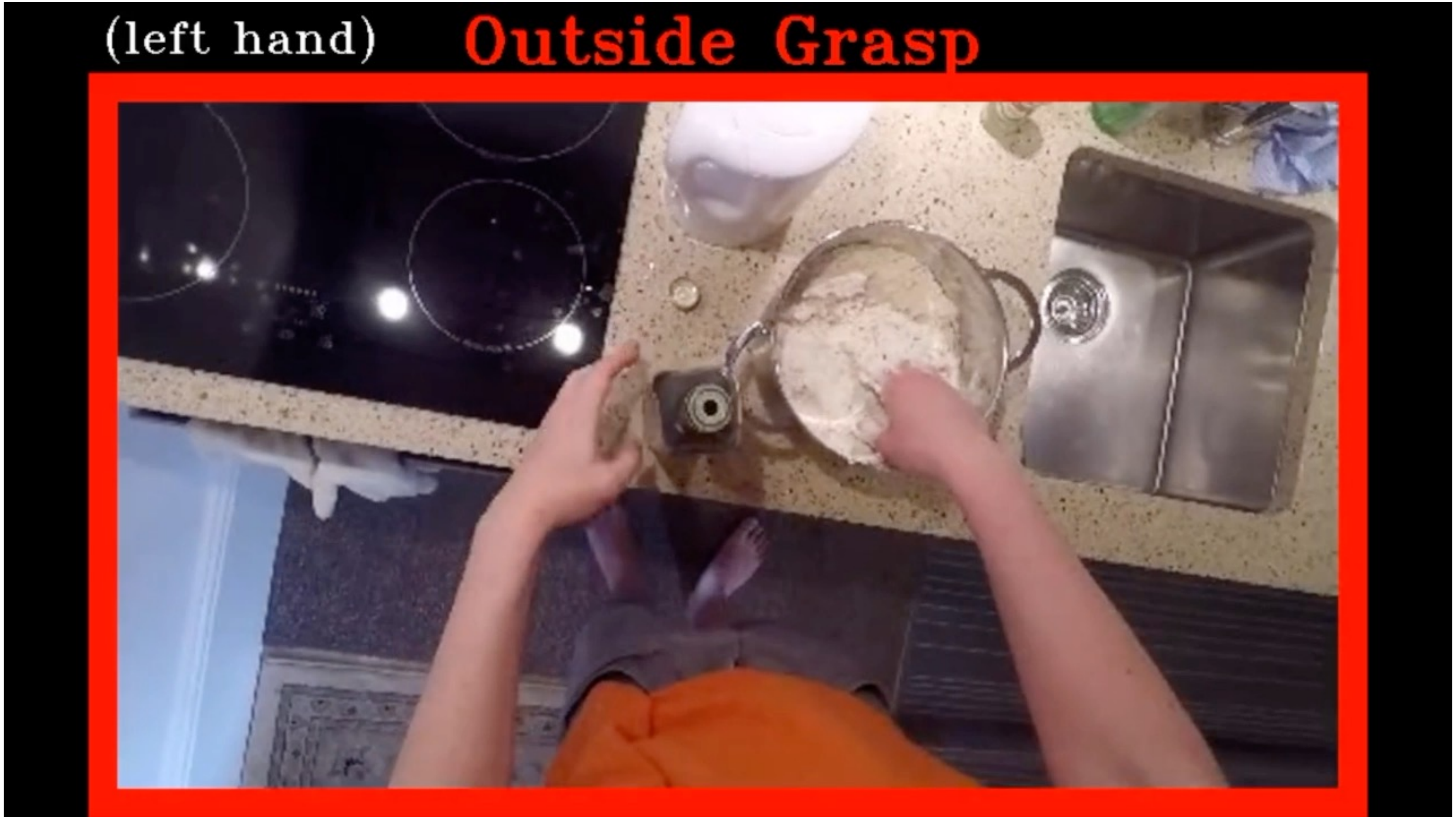
with: Zhifan Zhu



Z Fan, O Taheri, D [...]as, M Kocabas, M Kaufmann, M J Black, and O Hilliges (2023). ARCTIC: A dataset for dexterous bimanual hand- object manipulation. CVPR

Get a Grip

with: Zhifan Zhu



Z Zhu and D Damen (2023). Get a Grip: Reconstructing Hand-Object Stable Grasps in Egocentric Videos. *ArXiv*

Z Zhu and D Damen (2023). Get a Grip: Reconstructing Hand-Object Stable Grasps in Egocentric Videos. *ArXiv*

Get a Grip

with: Zhifan Zhu

Z Zhu and D Damen (2023). Get a Grip: Reconstructing Hand-Object Stable Grasps in Egocentric Videos. *ArXiv*

Opportunities in Egocentric Vision

Tasks are harder

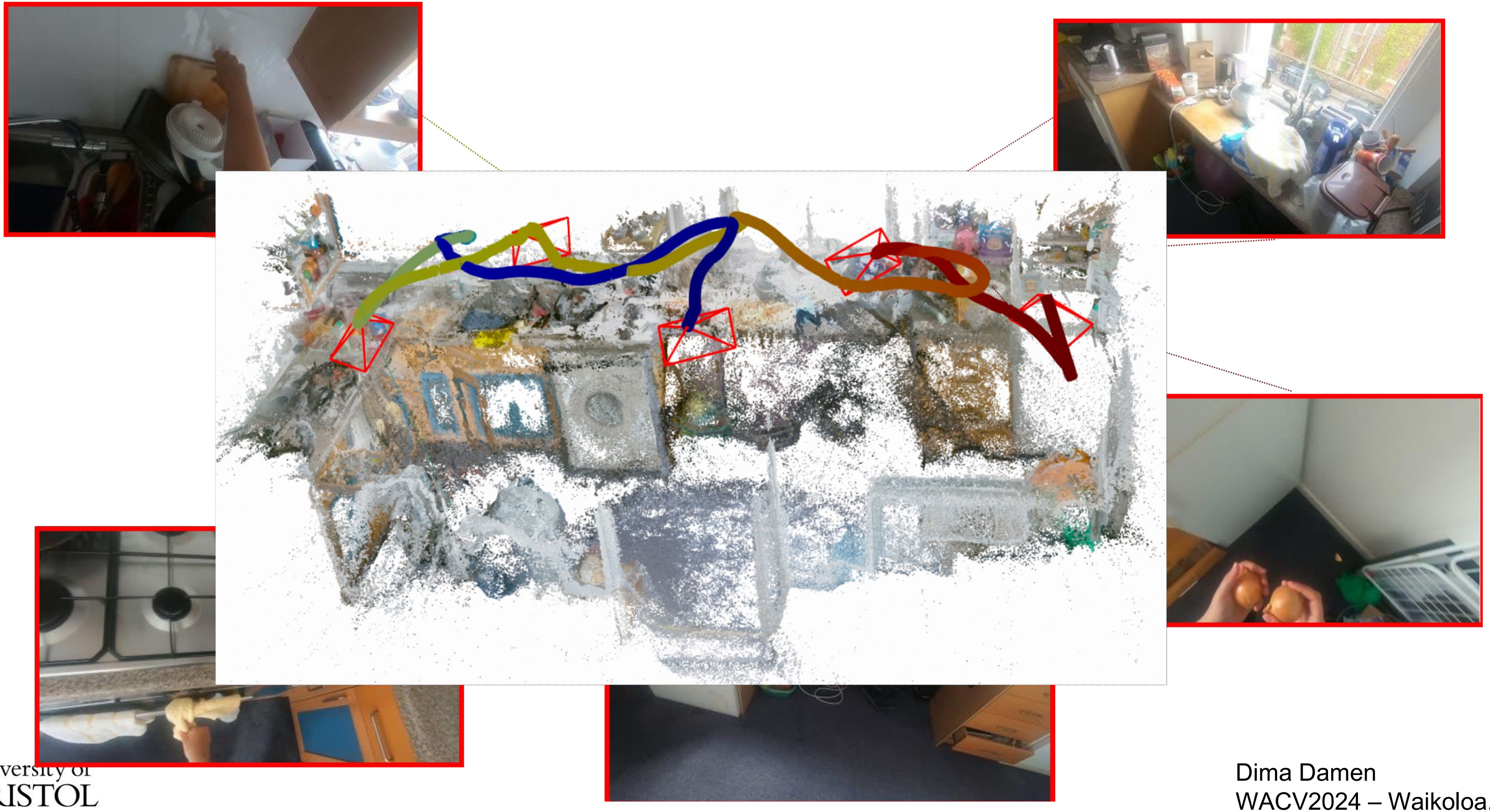
Detection, 3D Mapping, Tracking, VOS, Hand-Object, Generative, …

Solutions prove more rewarding

Weak supervision, Domain Adap/Gen., Audio-Visual, long-term understanding

EPIC Fields

with: V Tschernezki*, A Darkhalil*, Z Zhu*,
D Fouhey, I Laina, D Larlus, A Vedaldi



EPIC-KITCHENS

EPIC Fields

with: V Tschernezki*, A Darkhalil*, Z Zhu*,
D Fouhey, I Laina, D Larlus, A Vedaldi

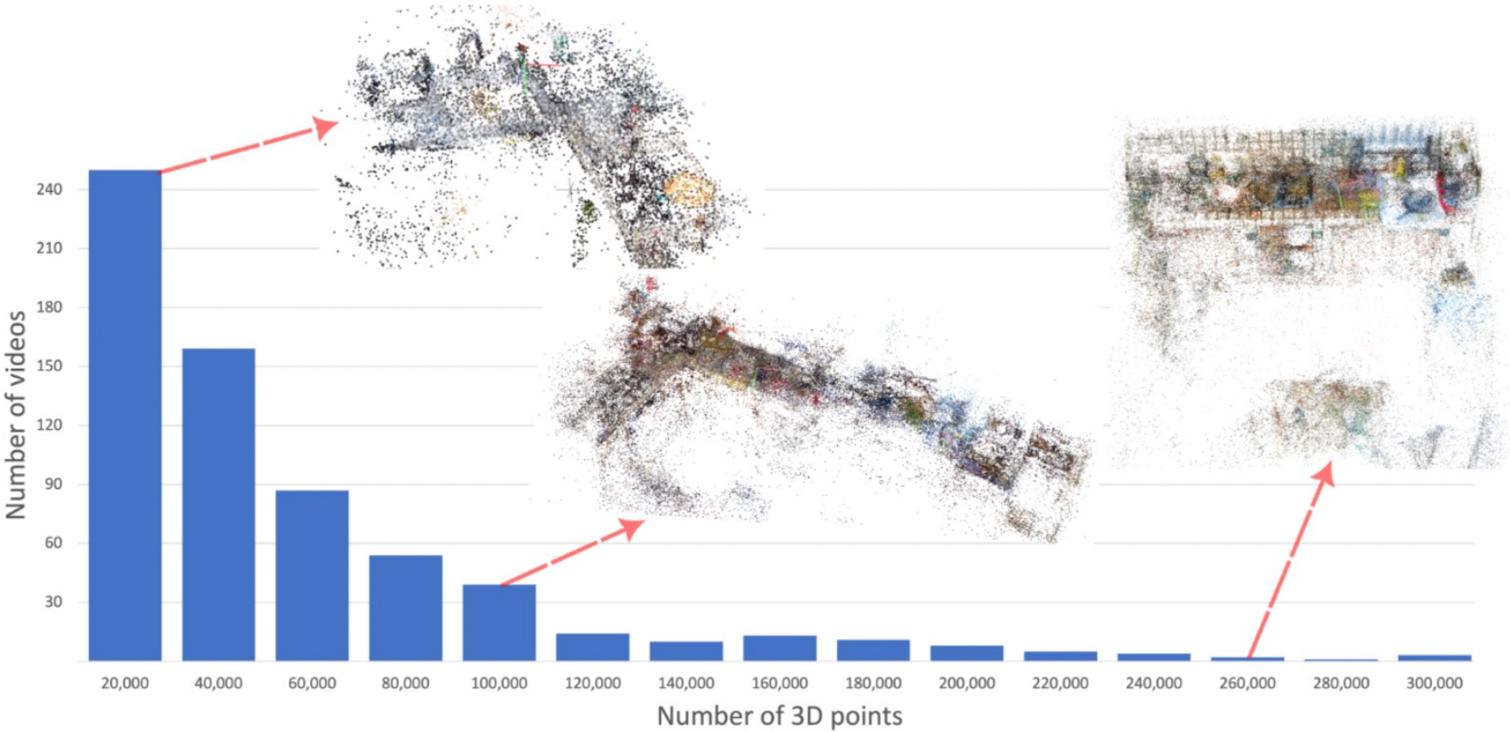


Figure 4: **Number of 3D points histogram.** The majority of our reconstructions generate less than 40,000 points that are enough to represent the kitchen. However, some reconstructions have more than 100,000, we include the point clouds for each points range showing the fine details covered by having more points

EPIC Fields

with: V Tschernezki*, A Darkhalil*, Z Zhu*,
D Fouhey, I Laina, D Larlus, A Vedaldi

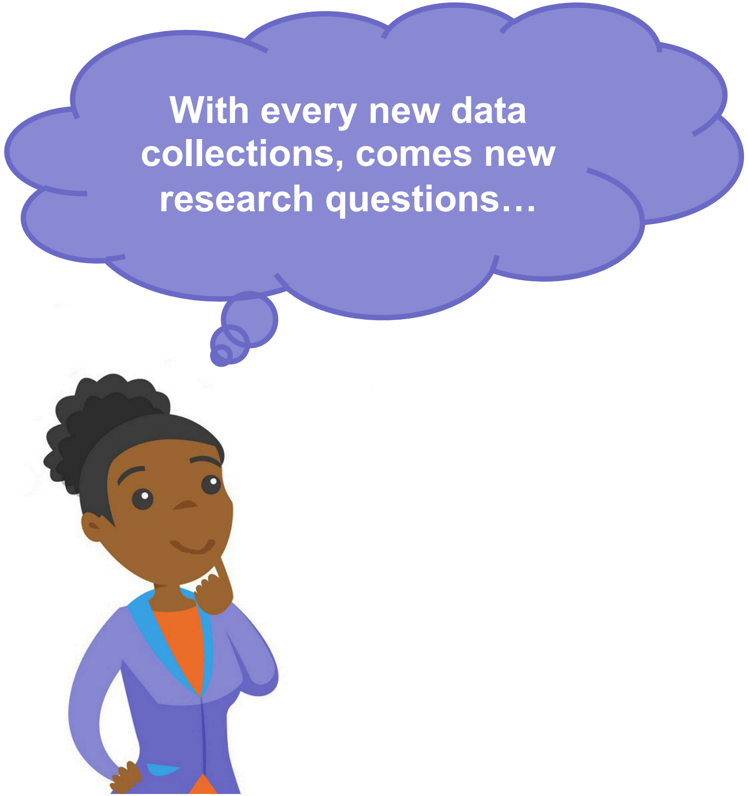
Table 1: Comparison of datasets commonly used in dynamic new-view synthesis.

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	✓

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

Ego-Exo4D

with: Kristen Grauman
+102 authors



K Grauman et al (2023). Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives.. *ArXiv*

Ego-Exo4D

with: Kristen Grauman
+102 authors

University of BRISTOL

K Grauman et al (2023). Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives.. *ArXiv*

Ego-Exo4D

with: Kristen Grauman
+102 authors

Ego-Exo Relation

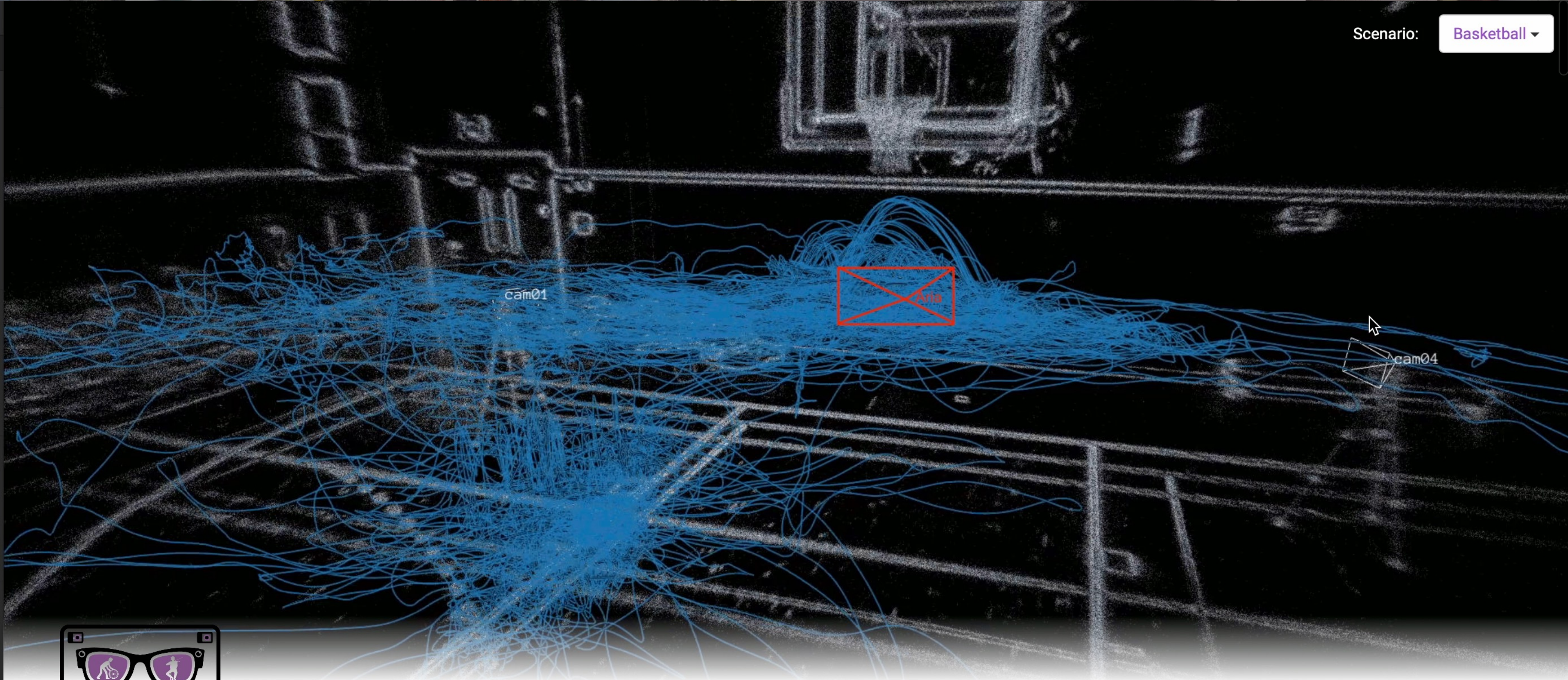
K Grauman et al (2023). Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives.. *ArXiv*

Ego-Exo4D

with: Kristen Grauman
+102 authors

K Graurnan et al (2023). Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives.. *ArXiv*

Scenario: Basketball



EGO-EXO4D

Hover your mouse over scene cameras above to see a sample video for the chosen scenario.

A **diverse**, large-scale **multi-modal**, **multi-view**, video dataset and benchmark collected across 13 cities worldwide by 839 camera wearers, capturing **1422 hours** of video of skilled human activities.

Learn More ↓ | Watch Video ↗ | Start Here ↗

An Outlook into the Future of Egocentric Vision

Chiara Plizzari*, Gabriele Goletto*, Antonino Furnari*, Siddhant Bansal*, Francesco Ragusa*, Giovanni Maria Farinella[†], Dima Damen[†], Tatiana Tommasi[†]

Politecnico di Torino

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

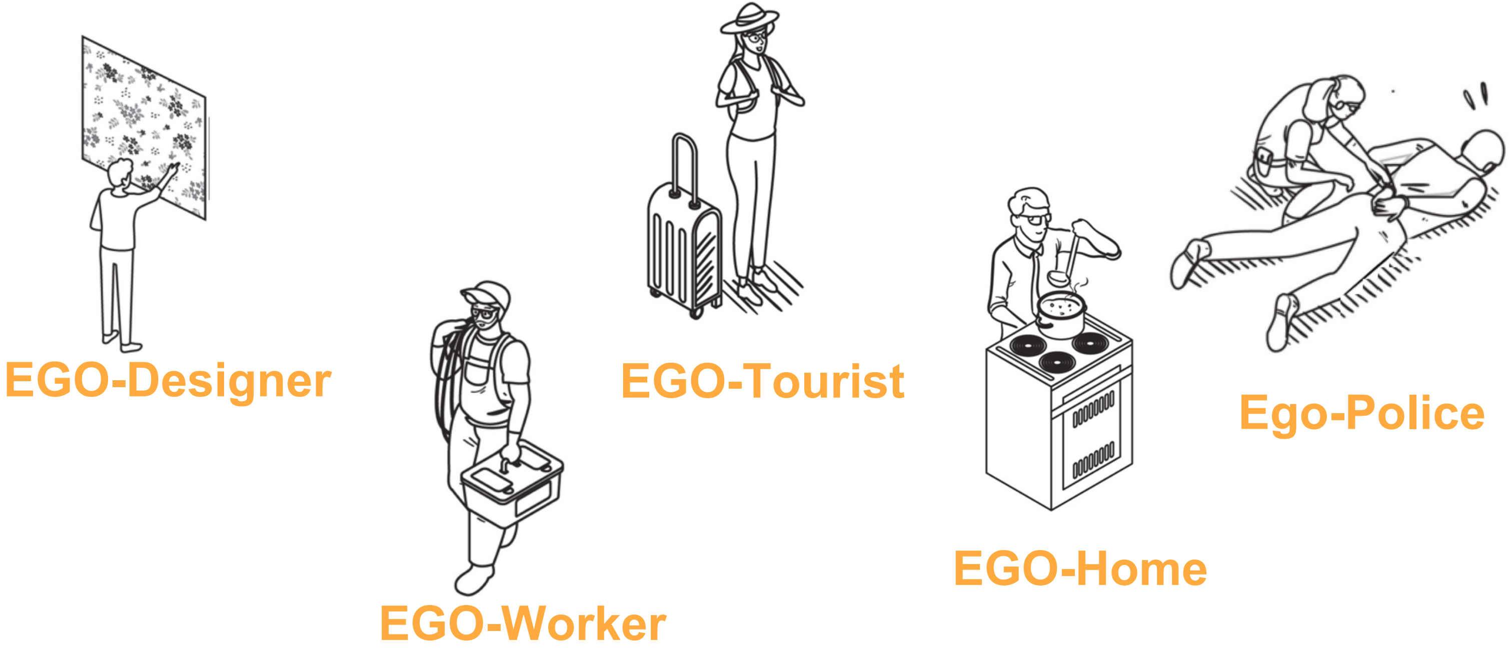
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

Envisioning an Ambitious Future and **Analysing** the Current Status of Egocentric Vision

How did we do this?

with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

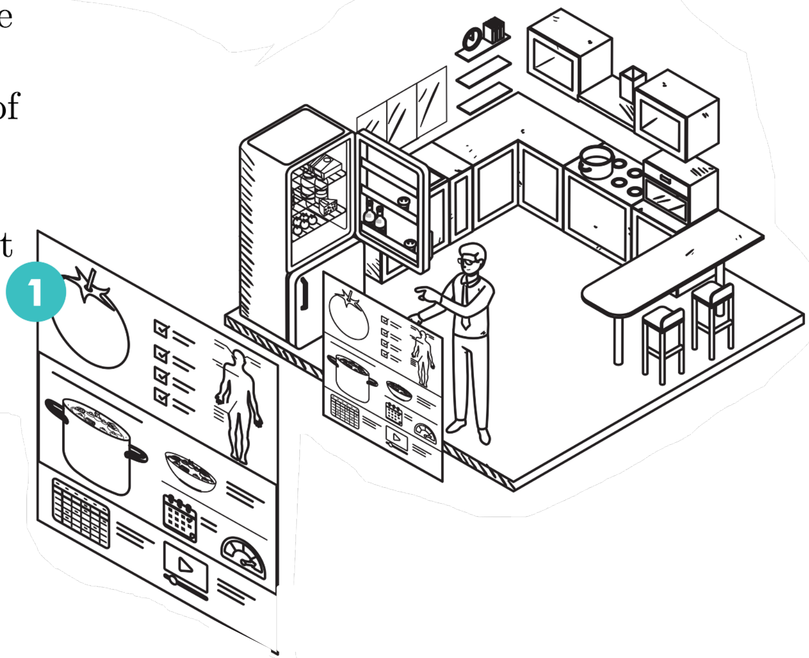
We imagined a device – *EgoAI* and envisioned its utility in multiple scenarios



EGO-Home

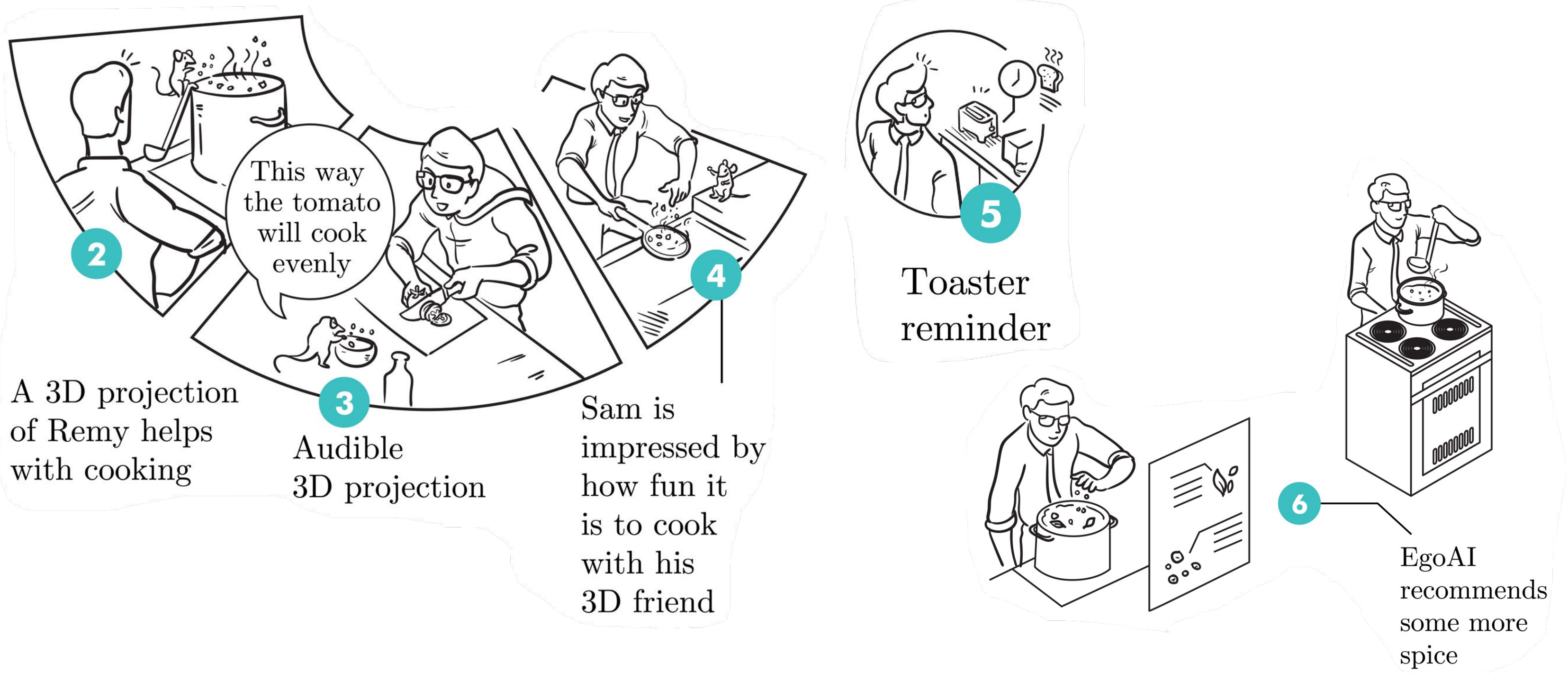
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

Sam is finally home after a long day. EgoAI kept track of Sam's food intake and a tomato soup sounds like the best complementary nutrition



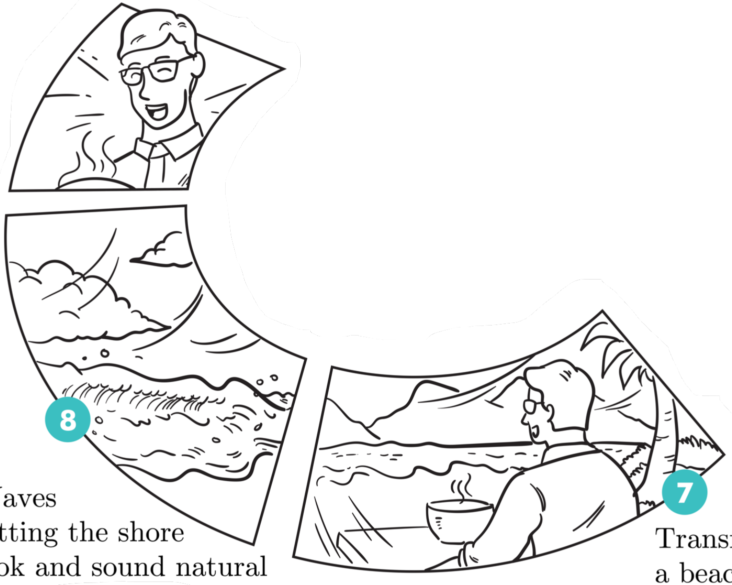
EGO-Home

with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi



EGO-Home

with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi



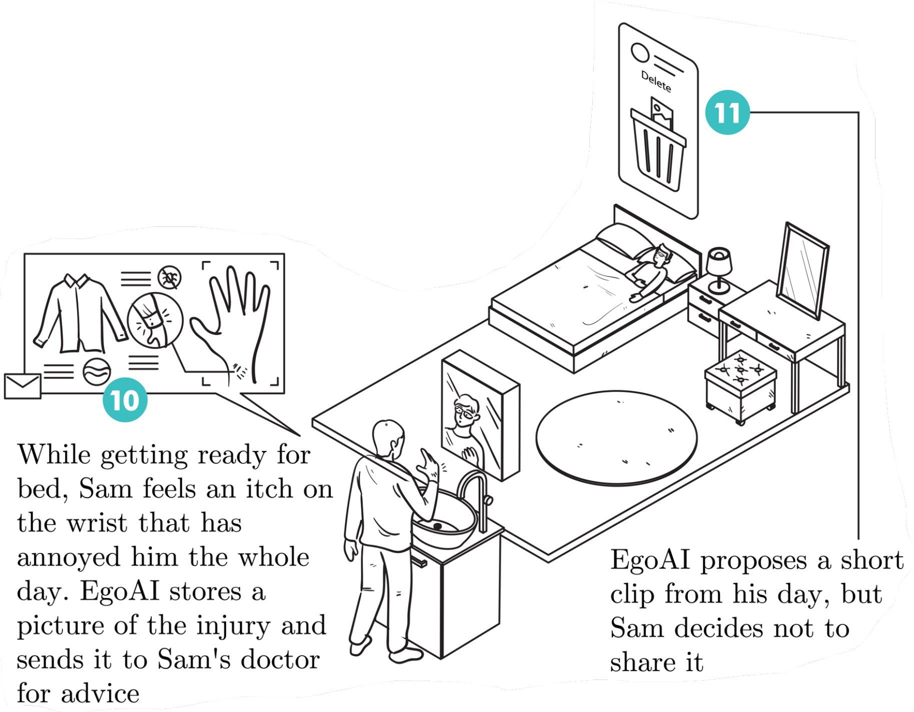
Waves
hitting the shore
look and sound natural

Transferred to
a beach he
visited last
summer

After dinner, Sam enjoys a group card game with his friends, who are connected through their own EgoAI

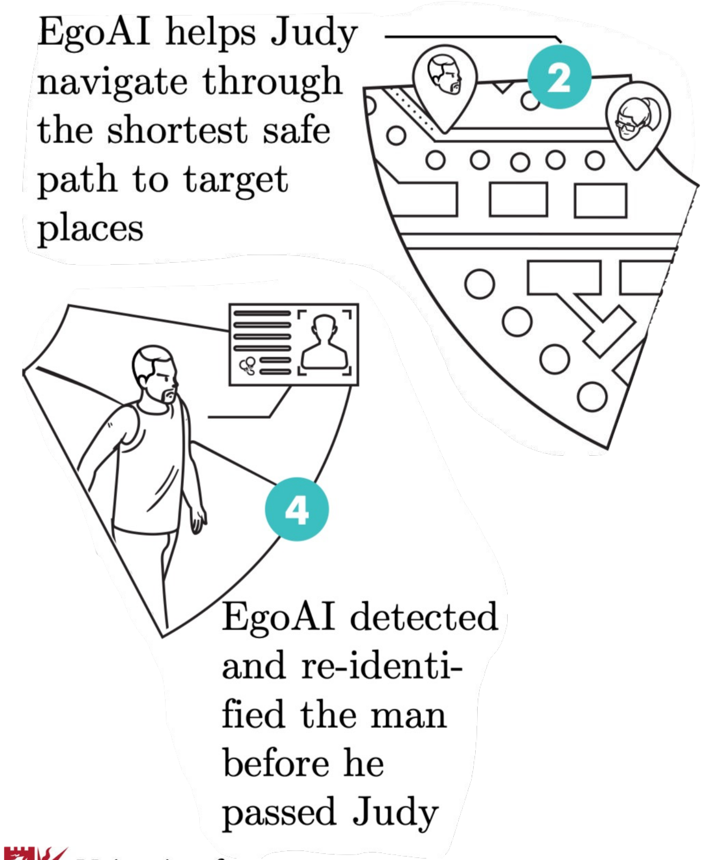
EGO-Home

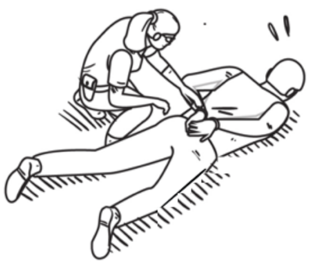
with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi



EGO-Home

with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi





EGO-Police

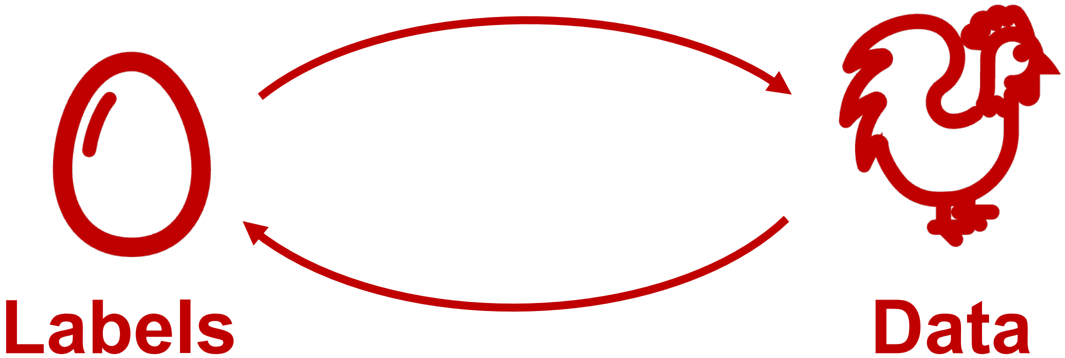
- Localisation and Navigation 1 2
- Messaging 1 3 11
- Action Recognition 2 13
- Person Re-ID 2 4
- Object Detection and Retrieval 7
- Measuring System 8 9
- Decision Making 9
- 3D Scene Understanding 10
- Hand-Object Interaction 12
- Summarisation 13
- Privacy 14

EGO-Home

with: Chiara Plizzari, Gabriele Goletto, Antonio Furnari, Siddhant Bansal, Francesco Ragusa, Giovanni Maria Farinella, Tatiana Tommasi

- 12 tasks
  - Seminal Works
  - SOTA methods
  - Datasets
  - Future Perspective
- 44 pages
- 385 references

In this talk…



EGO-EXO4D

Tasks are harder

Detection, 3D Mapping, Tracking, VOS, Hand-Object, Generative, …

Solutions prove more rewarding

Weak supervision, Domain Adap/Gen., Audio-Visual, long-term understanding

University of BRISTOL

Dima Damen
WACV2024 – Waikoloa, Hawaii

The Team

Visit us…

2024 Summer of Research@Bristol

Enjoy a research-oriented and fun summer!

Get supervision and support to develop cutting-edge research projects with a three-month internship with one of the leading researchers in Machine Learning and Computer Vision at the University of Bristol.

Open to PhD students in any related discipline.

MaVi
Machine Learning and Computer Vision

Click to apply online by:
Fri 19 Jan 2024

Thank you

For further info, datasets, code, publications…

http://dimadamen.github.io

@dimadamen

http://www.linkedin.com/in/dimadamen

Q&A