Video Understanding

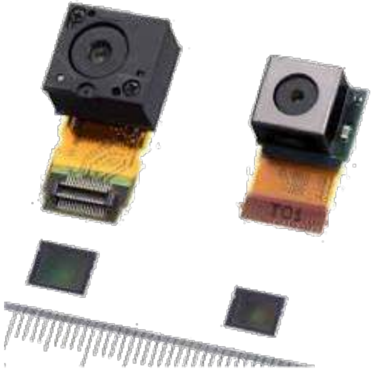
An Egocentric Perspective

University of BRISTOL

Dima Damen

January 2, 2022

Visual Sensing – the landscape



Visual Sensing – the landscape

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

Darnella Frazier, the teenager who shot the harrowing video of George Floyd under the knee of the Minneapolis police officer now charged in his death, testified Tuesday that she began recording because "it wasn't right, he was suffering, he was in pain."

The present…

The future…





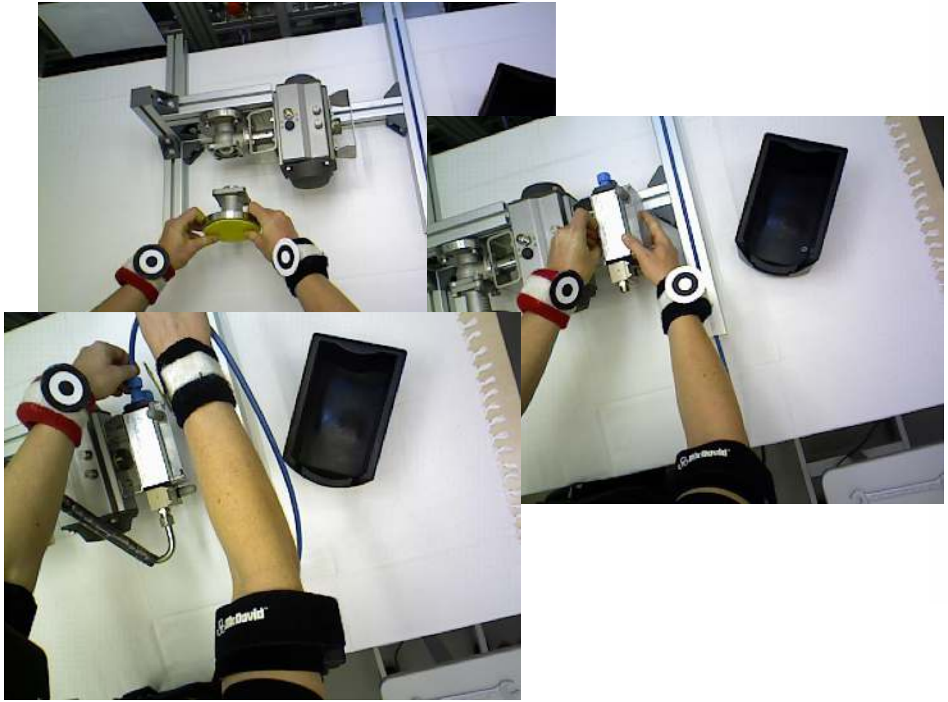
Wearable Sensing

Wearable Sensing

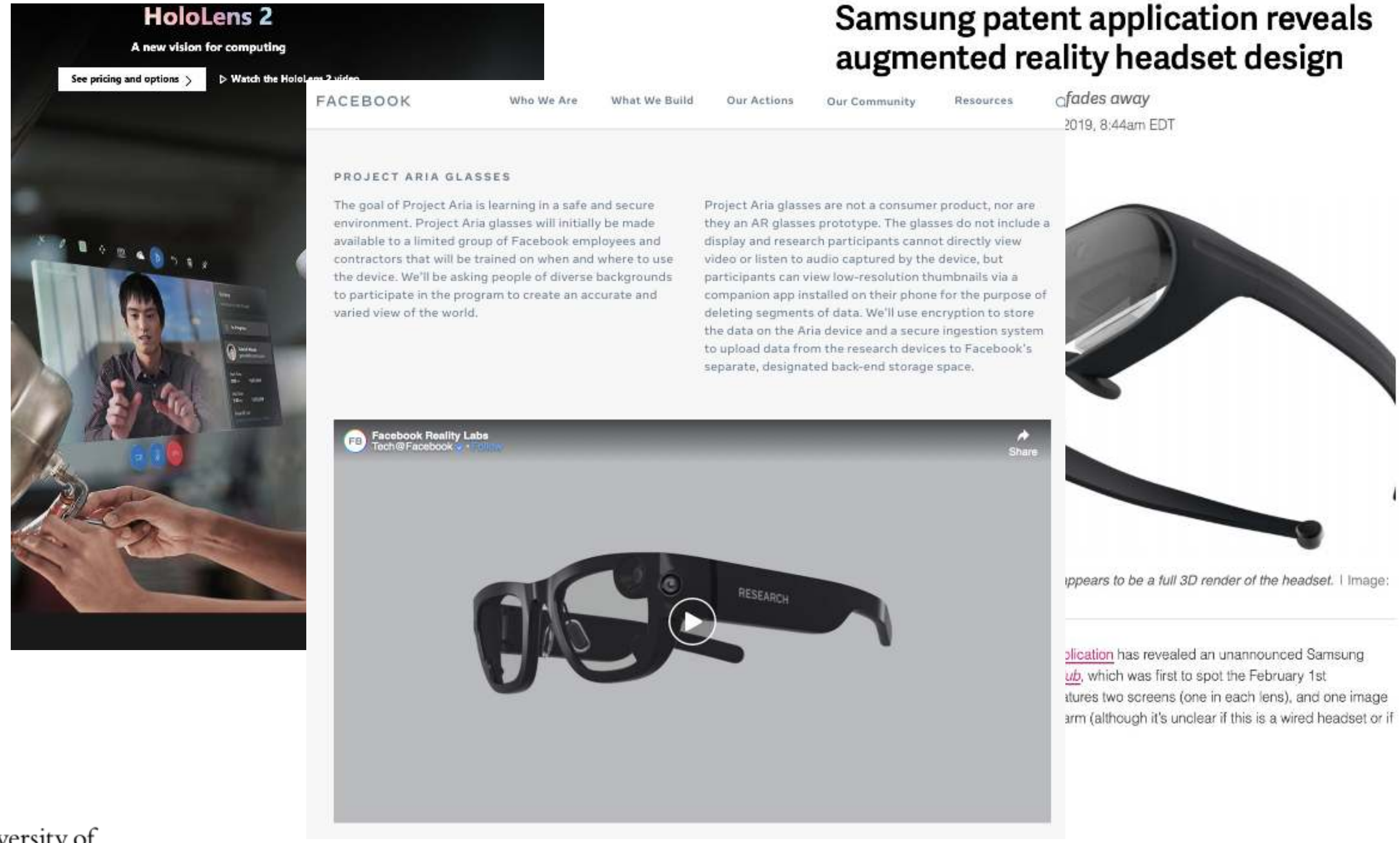
Wearable Sensing

Egocentric Vision

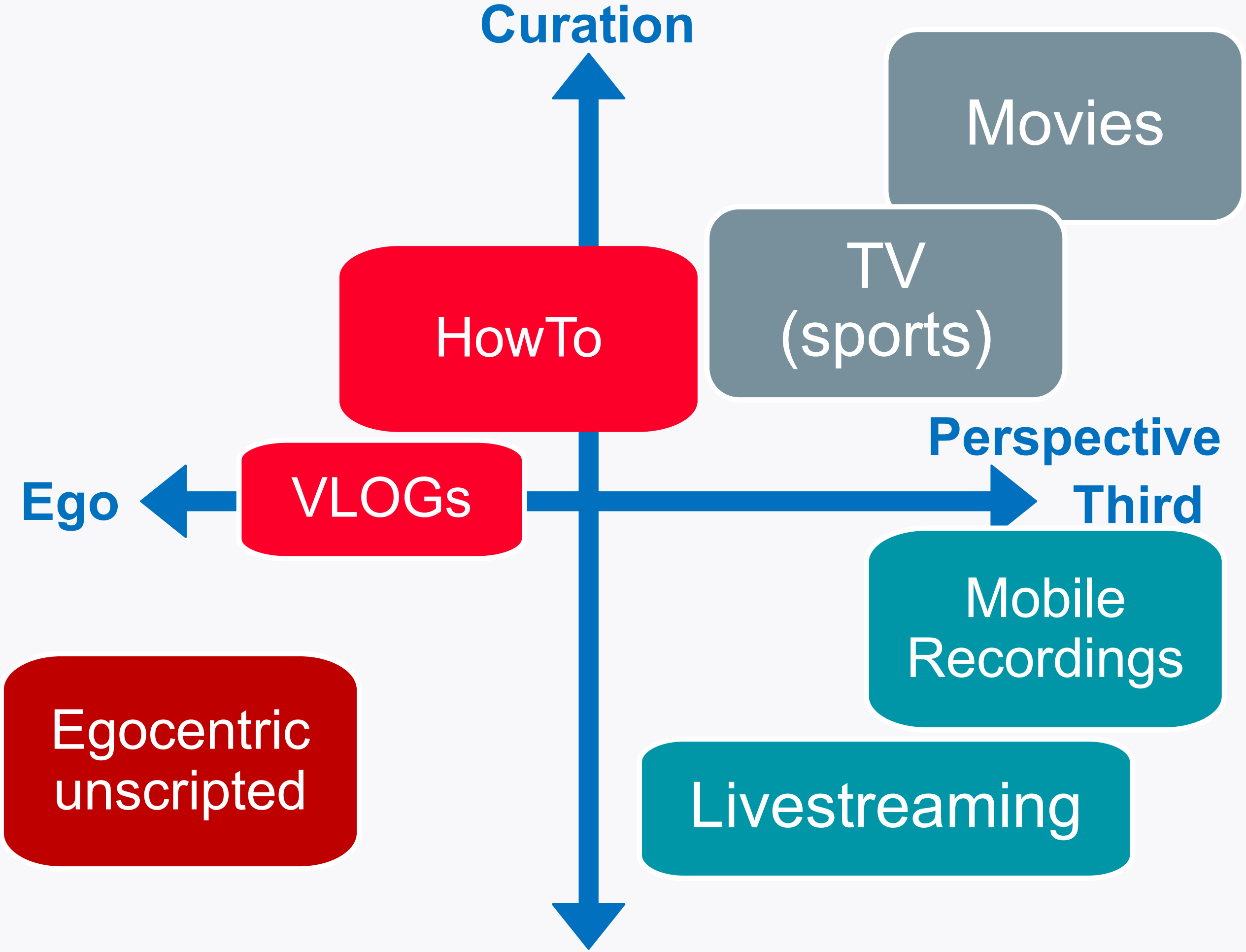
- Starting 2010…



Why Egocentric Vision?



Video Sources



Egocentric Videos?

45 seconds

Egocentric Videos?

And so?

- What? – Actions & Objects,
- How? – Methods, Tools, Skills
- Why? – Goals and next steps
- When? – Exact moments
- Where? – Tracking and localization
- ~~Who?~~

Scaling and Rescaling Egocentric Vision: The EPIC-KITCHENS Dataset


Dima Damen


Hazel Doughty


Giovanni M. Farinella


Sanja Fidler


Antonino Furnari


Evangelos Kazakos


Jian Ma


Davide Moltisanti

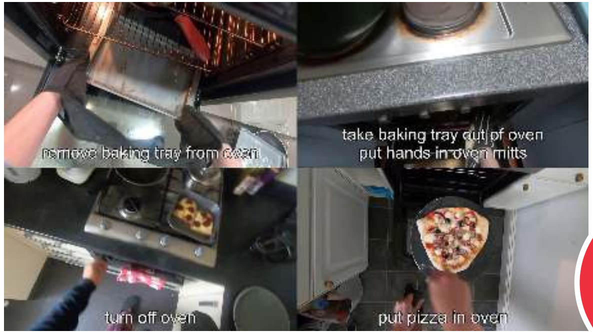

Jonathan Munro


Toby Perrett


Will Price


Michael Wray

Scaling and Rescaling Egocentric Vision



Data Collection

Live Narrations

Dense Action Segments

Extension Data Collection

Pause-and-talk Narrator

Improved Annotations

EPIC-KITCHENS-100

EPIC-KITCHENS-55

EPIC-KITCHENS-55

Avg actions per video

91.3

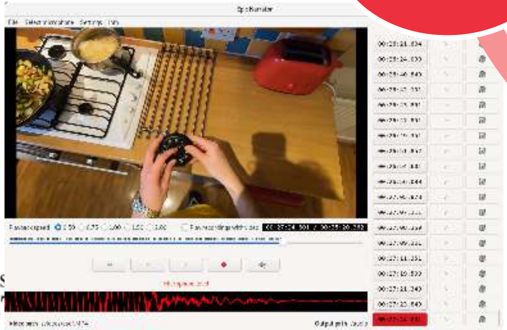
188.6

EPIC-KITCHENS-100

Avg actions per minute

13

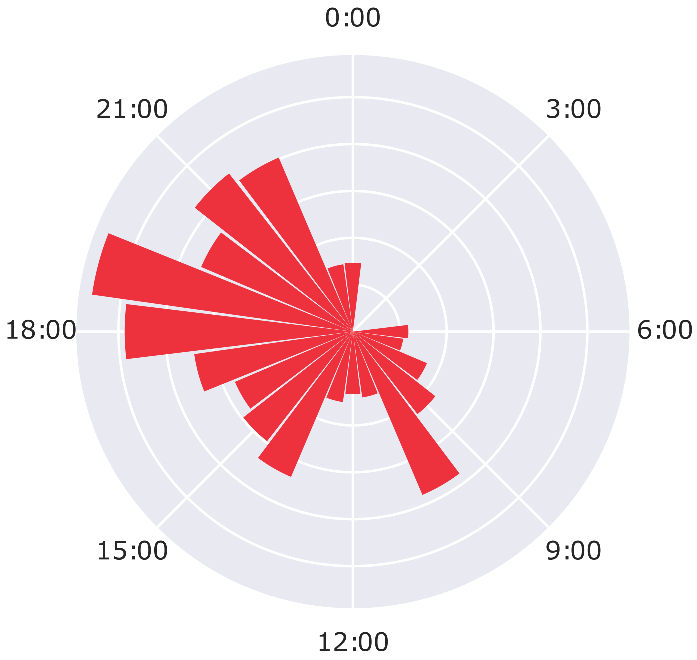
20



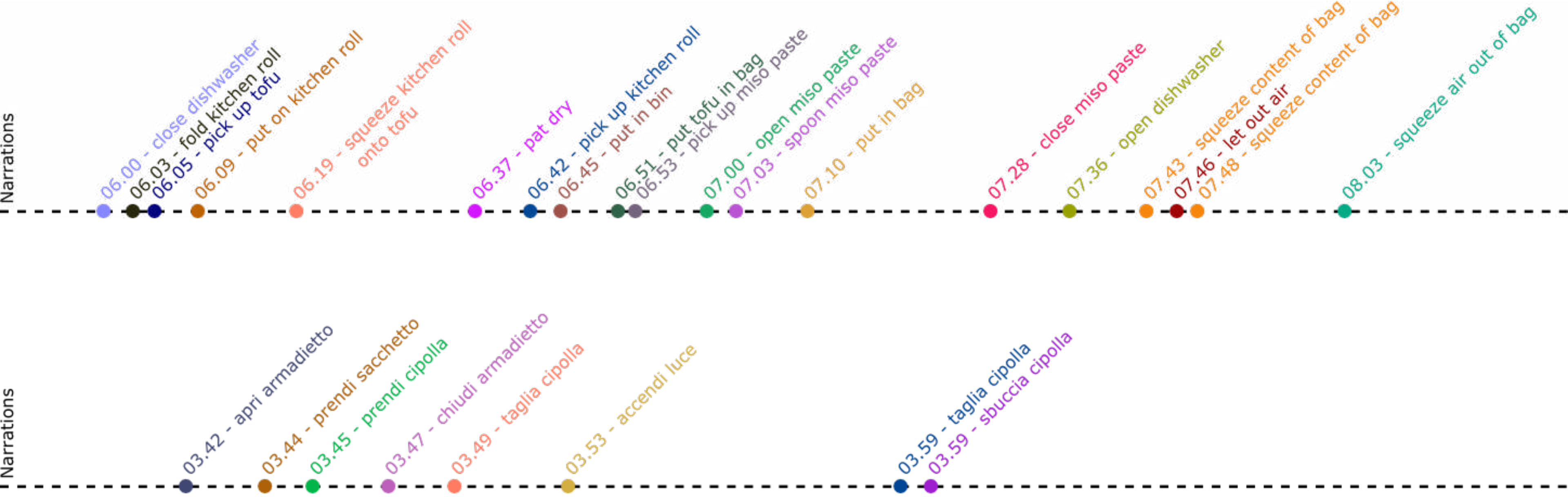
- Head-Mounted Go-Pro, adjustable mounting
- Recording starts immediately before entering the kitchen
- Only stopped before leaving the kitchen

Scaling and Rescaling Egocentric Vision

- 45 kitchens
- Single-person environments
- 4 cities
- May – Nov 2017 – 55 hours
- May – Dec 2019 – 45 hours
- 10 nationalities
- 3 days - all kitchen activities



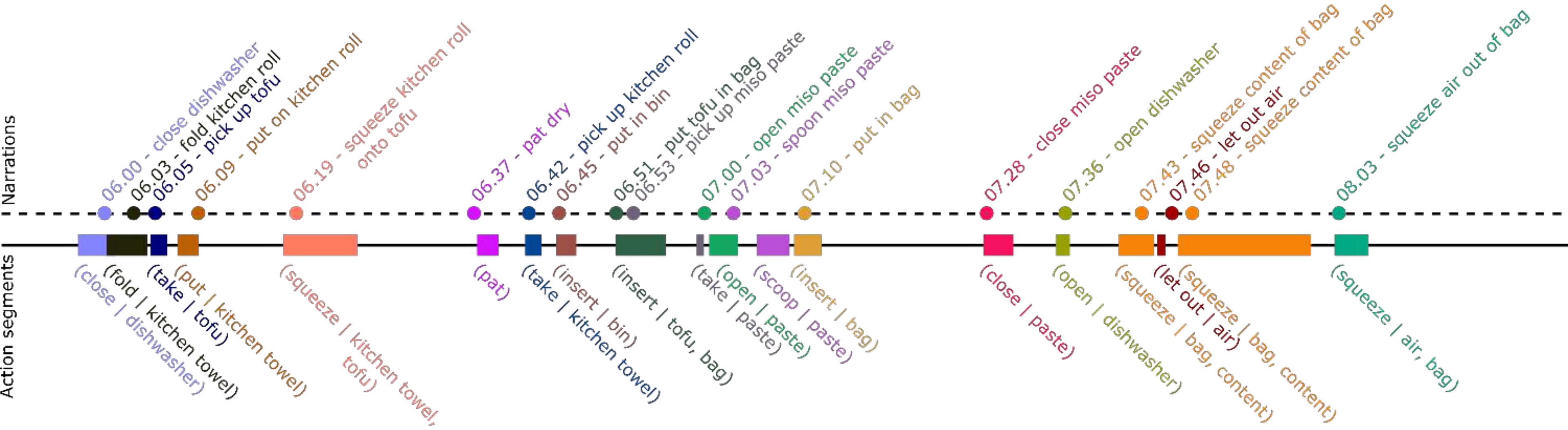
Scaling and Rescaling Egocentric Vision



University of BRISTOL

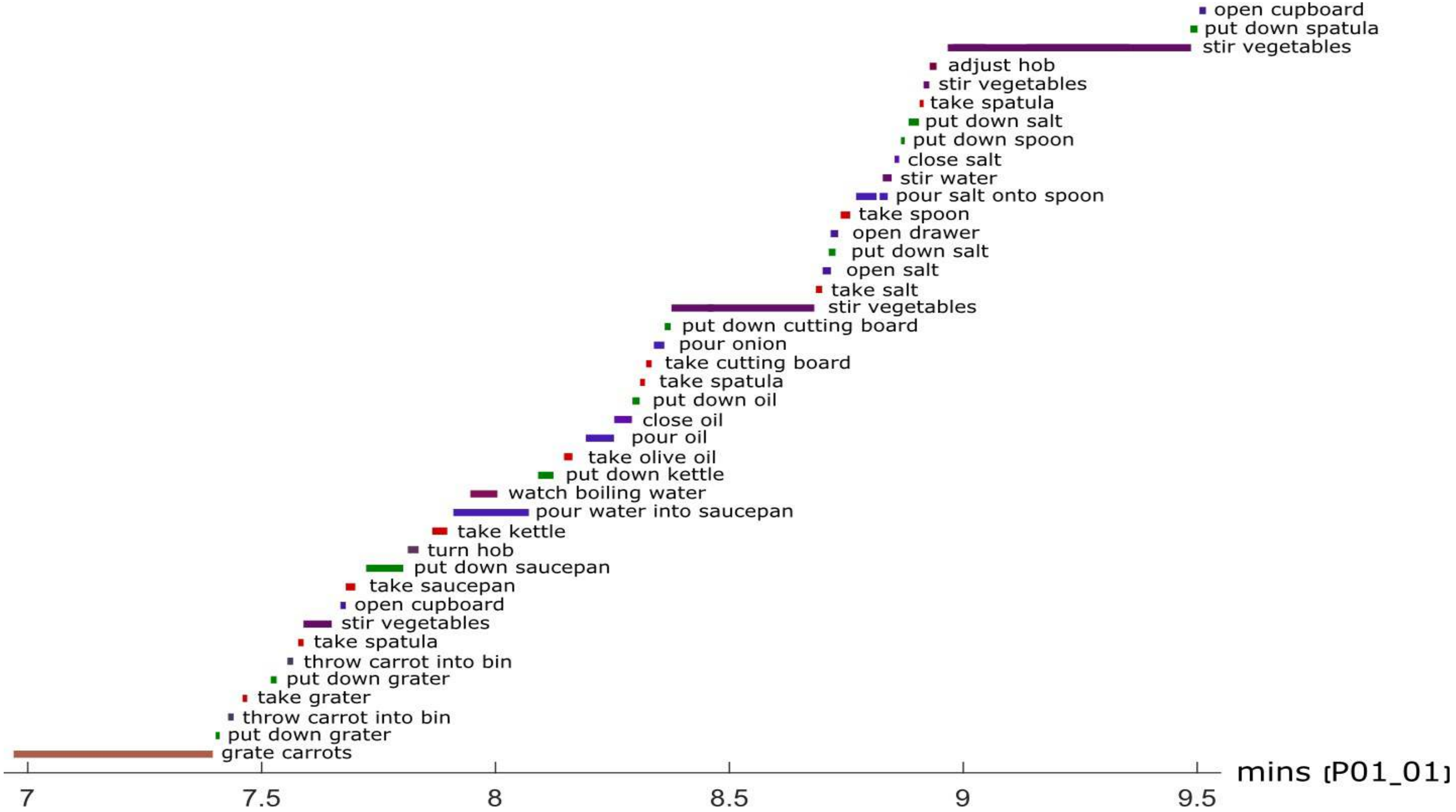
Scaling and Rescaling Egocentric Vision

Scaling and Rescaling Egocentric Vision



University of BRISTOL

Scaling and Rescaling Egocentric Vision

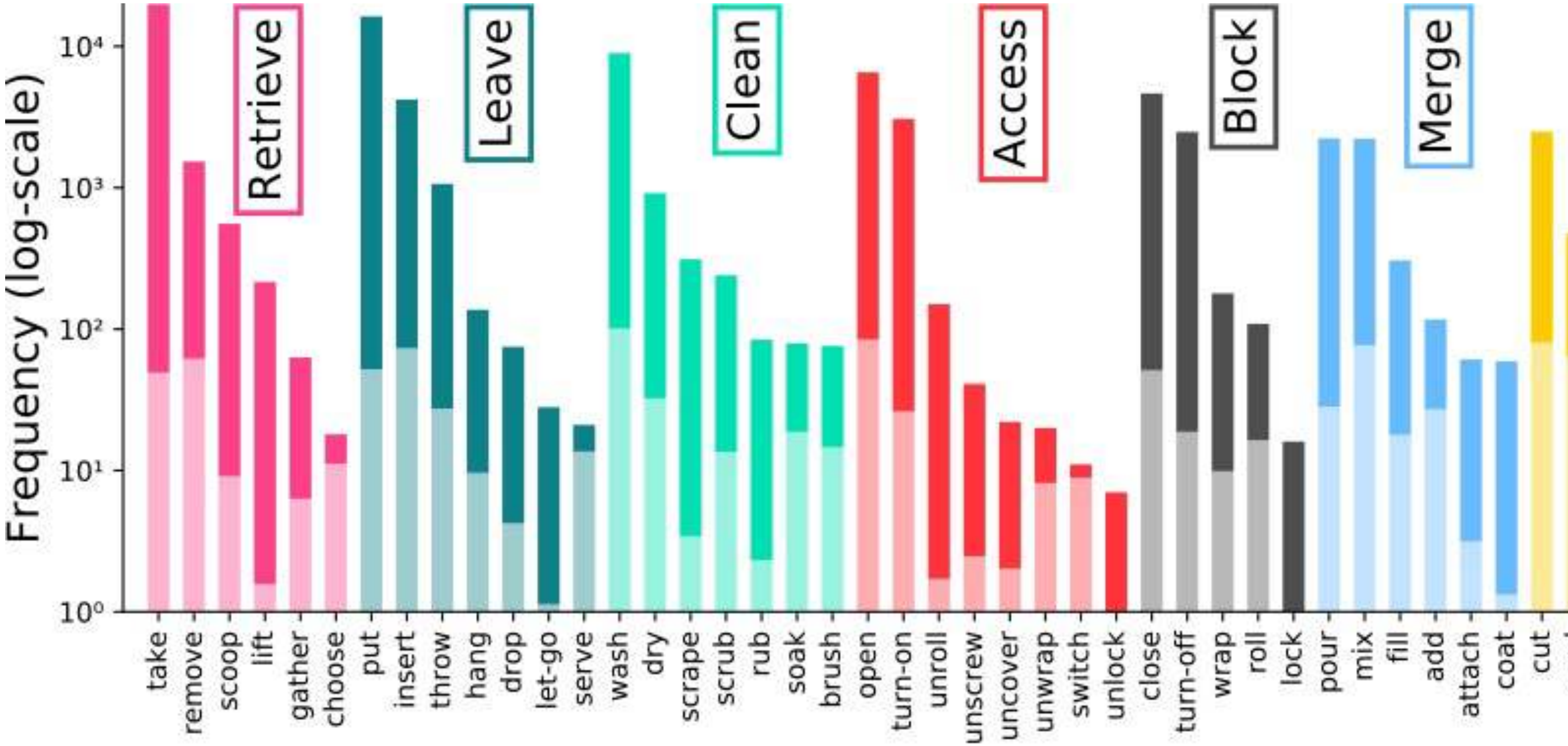


University of BRISTOL


open oven
put spoon on counter
put on glove
pick up fork


put down glass
pick up glass
put down plate
put down spoon
pick up aeropress filter

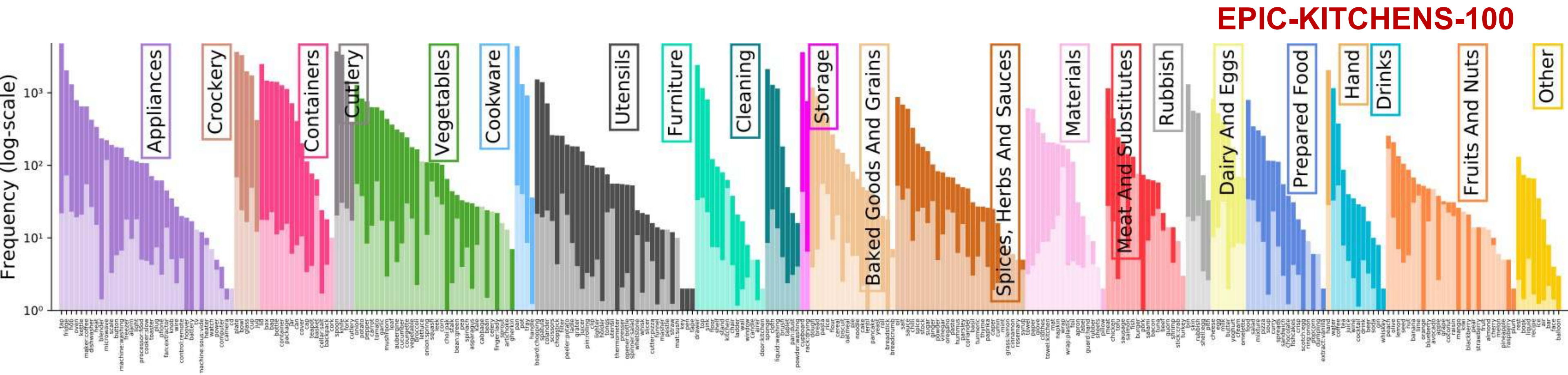
EPIC-KITCHENS-100 Statistics



University of BRISTOL

Dima Damen
January 2, 2022

EPIC-KITCHENS-100 Statistics



University of BRISTOL

Dima Damen
January 2, 2022

EPIC-KITCHENS-100 Release

FHD video:
- 1920x1080 px
- 60FPS / 50 FPS

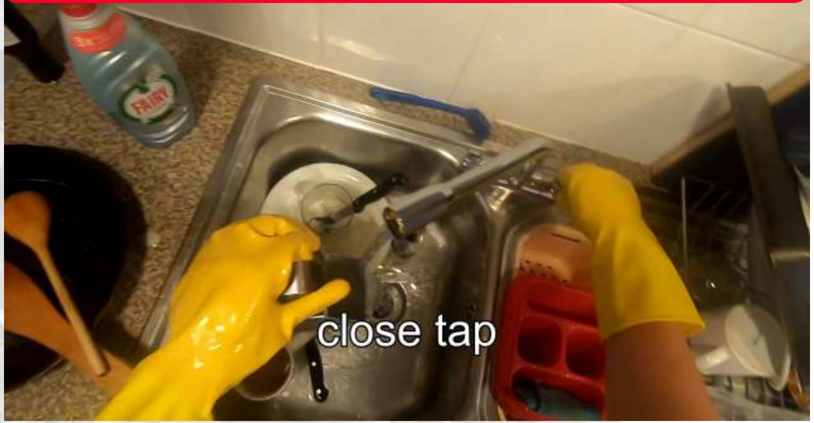
RGB frames:
- 456x256 px
- 60FPS

TVL_1 optical flow (u, v) frames:
- 456x256 px
- 30FPS

37 Participants – 8 in the same kitchen

EPIC-KITCHENS

2 years later



37 Participants – 8 in a different kitchen

EPIC-KITCHENS

2 years later

Action Recognition Challenge

Action Recognition Challenge

Given a trimmed action segment:

$(t_{\text{start}}, t_{\text{stop}})$

classify the action within.

$\hat{y}_{\text{verb}} =$ open

$\hat{y}_{\text{noun}} =$ oven

$\hat{y}_{\text{action}} =$ (open, oven)

Action Recognition Challenge

Seen Kitchens (S1)																
#	User	Entries	Date of Last Entry	Team Name	Top-1 Accuracy (%)			Top-5 Accuracy (%)			Precision (%)			Recall (%)		
					Verb ▲	Noun ▲	Action ▲	Verb ▲	Noun ▲	Action ▲	Verb ▲	Noun ▲	Action ▲	Verb ▲	Noun ▲	Action ▲
1	wasun	14	05/28/20	UTS-Baidu	70.41 (1)	52.85 (1)	42.57 (1)	90.78 (4)	76.62 (2)	63.55 (2)	60.44 (4)	47.11 (1)	24.94 (3)	45.82 (4)	50.02 (1)	26.93 (2)
2	action_banks	18	05/29/20	NUS_CVML	66.56 (6)	49.60 (4)	41.59 (2)	90.10 (5)	77.03 (1)	64.11 (1)	59.43 (7)	45.62 (3)	25.37 (1)	41.65 (8)	46.25 (4)	26.98 (1)
3	Sudhakaran	50	05/29/20	FBK_HuPBA	68.68 (3)	49.35 (5)	40.00 (3)	90.97 (3)	72.45 (5)	60.23 (4)	60.63 (3)	45.45 (4)	21.82 (6)	47.19 (2)	45.84 (5)	24.34 (4)
4	tnet	34	05/27/20	SAIC_Cambridge	69.43 (2)	49.71 (3)	40.00 (3)	91.23 (2)	73.18 (3)	60.53 (3)	60.01 (5)	45.74 (2)	24.95 (2)	47.40 (1)	46.78 (3)	25.27 (3)
5	aptx4869lm	12	01/30/20	GT-WISC-MPI	68.51 (4)	49.96 (2)	38.75 (4)	89.33 (8)	72.30 (6)	58.99 (5)	51.04 (16)	44.00 (6)	23.70 (5)	43.70 (7)	47.32 (2)	23.92 (5)
6	weiyaowang	14	05/28/20		66.67 (5)	48.48 (6)	37.12 (5)	88.90 (9)	71.36 (7)	56.21 (8)	51.86 (14)	41.26 (7)	20.97 (7)	44.33 (6)	44.92 (6)	21.48 (8)
7	TBN_Ensemble	1	07/20/19	Bristol-Oxford	66.10 (7)	47.88 (7)	36.66 (6)	91.28 (1)	72.80 (4)	58.62 (6)	60.73 (2)	44.89 (5)	24.01 (4)	46.81 (3)	43.88 (7)	22.92 (6)
8	cvg_uni_bonn	21	05/27/20	CVG Lab Uni Bonn	62.86 (8)	43.44 (10)	34.53 (7)	89.64 (6)	69.24 (8)	56.73 (7)	52.82 (13)	38.81 (11)	19.21 (10)	44.72 (5)	39.50 (10)	21.80 (7)
9	antoninofurnari	1	07/19/19		56.93 (16)	43.05 (11)	33.06 (8)	85.68 (20)	67.12 (11)	55.32 (9)	50.42 (17)	39.84 (9)	18.91 (11)	37.82 (14)	38.11 (11)	19.12 (11)
10	Wenda	12	04/25/20	Wenda Go!	61.10 (12)	43.73 (8)	31.54 (9)	89.45 (7)	68.45 (10)	52.62 (10)	55.79 (10)	41.24 (8)	20.67 (8)	40.25 (10)	40.49 (9)	19.33 (10)
11	EPIC_TSM_FUSION	1	03/30/20		62.37	41.88	29.90	88.55	66.43	49.81	59.51	39.50	18.38	34.44	36.04	15.80

Open (

Five recently closed chall

- Action Recognition
- Action Detection
- Action Anticipation
- Unsupervised Domain
- Multi-Instance Retriev

University of BRISTOL

More?

http://epic-kitchens.github.io

EPIC-KITCHENS-100 2021 CHALLENGES

Challenge and Leaderboard Details with links to Codalab Leaderboards

For Challenge Results and winners on EPIC-KITCHENS-55, go to: Challenge 2020 Details.
Note that these are NEW leaderboards, and results are not directly comparable to last year's results.

EPIC-Kitchens 2021 Challenges - Dates

Aug 23rd, 2020	EPIC-Kitchens Challenges 2021 Launched alongisde EPIC@ECCV Workshop
May 28, 2021	Server Submission Deadline at 23:59:59 GMT
Jun 4, 2021	Deadline for Submission of Technical Reports
TBC	Results announcement dates will be confirmed later

Challenges Guidelines

The five challenges below and their test sets and evaluation servers are available via CodaLab. The leaderboards will decide the winners for each individual challenge. For each challenge, the CodaLab server page details submission format and evaluation metrics.

To **enter any of the five competitions**, you need to register an account for that challenge using a valid institute (university/company) email address. A single registration per research team is allowed. We perform a manual check for each submission, and expect to accept registrations within 2 working days.

For all challenges the maximum submissions per day is limited to 1, and the overall maximum number of submissions per team is limited to 50 overall, submitted once a day. This includes any failed submissions due to formats - please do not contact us to ask for increasing this limit.

To **submit** your results, follow the JSON submission format, upload your results and give time for the evaluation to complete (in the order of several minutes). **Note our new rules on declaring the supervision level, given our proposed scale, for each submission.** After the evaluation is complete, the results automatically appear on the public leaderboards but you are allowed to withdraw these at any point in time.

To **participate** in the challenge, you need to have your results on the public leaderboard, along with an informative team name (that represents your institute or the collection of institutes participating in the work), as well as brief information on your method. You are also required to submit a report (details TBC).

Make the most of the starter packs available with the challenges, and should you have any questions, please use our info email uob-epic-kitchens@bristol.ac.uk

University of BRISTOL

EPIC KITCHENS

ABOUT STATS DOWNLOADS CHALLENGES TEAM

NEWS

- 1st of July 2020: EPIC-KITCHENS-100 is now Released! Watch release webinar recording
- Watch the dataset's trailer and video demonstration on YouTube

What is EPIC-KITCHENS-100?

The *extended* **largest dataset in first-person (egocentric) vision**; multi-faceted, audio-visual, **non-scripted** recordings in native environments - i.e. the wearers' homes, capturing all daily activities in the kitchen over multiple days. Annotations are collected using a novel 'Pause-and-Talk' narration interface.

Characteristics

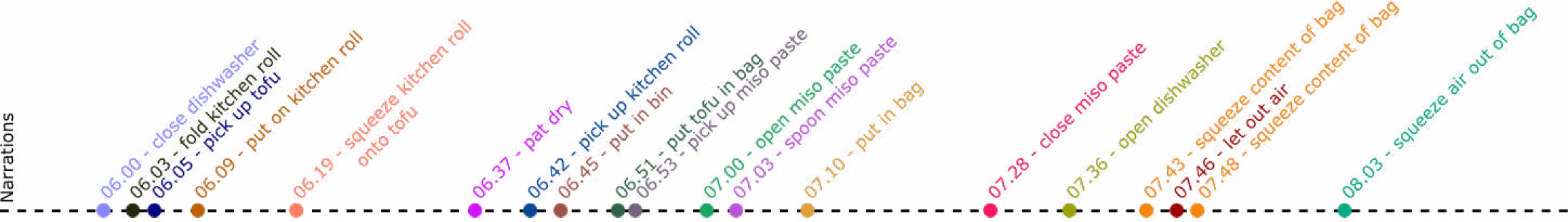
- 45 kitchens - 4 cities
- Head-mounted camera
- **100** hours of recording - Full HD
- 20M frames
- Multi-language narrations
- 90K action segments
- 20K unique narrations
- 97 verb classes, 300 noun classes
- 6 challenges

Previous versions...

- The previous version of the dataset (55 hours) was released in April 2018
- Refer to EPIC-KITCHENS-55 for details
- 2020 Challenges: Results, Tech Report
- 2019 Challenges: Results, Tech Report
- EPIC-KITCHENS-55 leaderboards remain open until the end of 2020

Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler



D Moltisanti et al (2019). Action Recognition from Single Timestamp Supervision in Untrimmed Videos. *Computer Vision and Pattern Recognition (CVPR)*

Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler

D Moltisanti et al (2019). Action Recognition from Single Timestamp Supervision in Untrimmed Videos. *Computer Vision and Pattern Recognition (CVPR)*

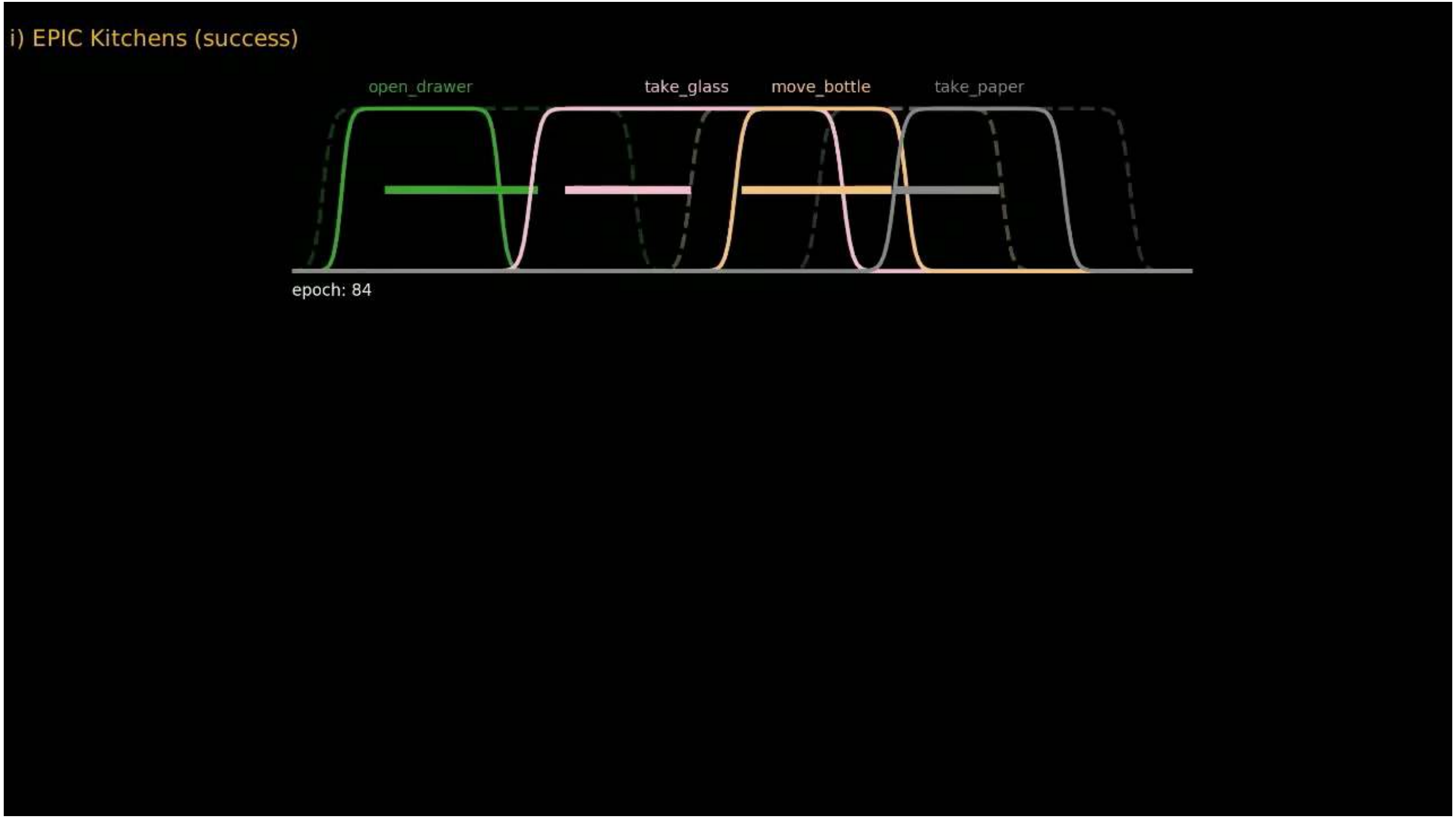
Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler

D Moltisanti et al (2019). Action Recognition from Single Timestamp Supervision in Untrimmed Videos. *Computer Vision and Pattern Recognition (CVPR)*

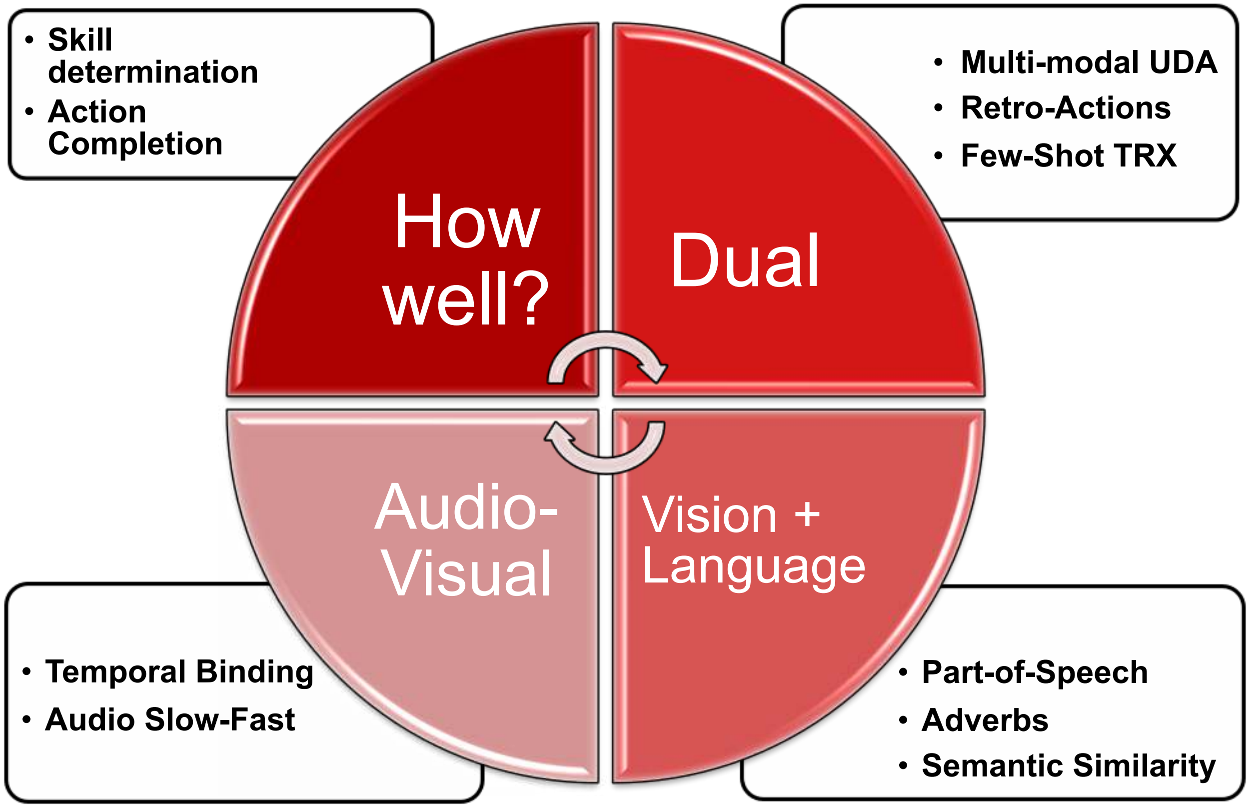
Learning from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler



D Moltisanti et al (2019). Action Recognition from Single Timestamp Supervision in Untrimmed Videos. *Computer Vision and Pattern Recognition (CVPR)*

VU - An Egocentric Perspective



VU - An Egocentric Perspective

How well?

CVPR18, CVPR19

BMVC18, ICCVW19

- Skill determination
- Action Completion

Dual

- Multi-modal UDA
- Retro-Actions
- Few-Shot TRX

CVPR20
ICCVW19
CVPR21

Audio-Visual

ICCV19
ICASSP21

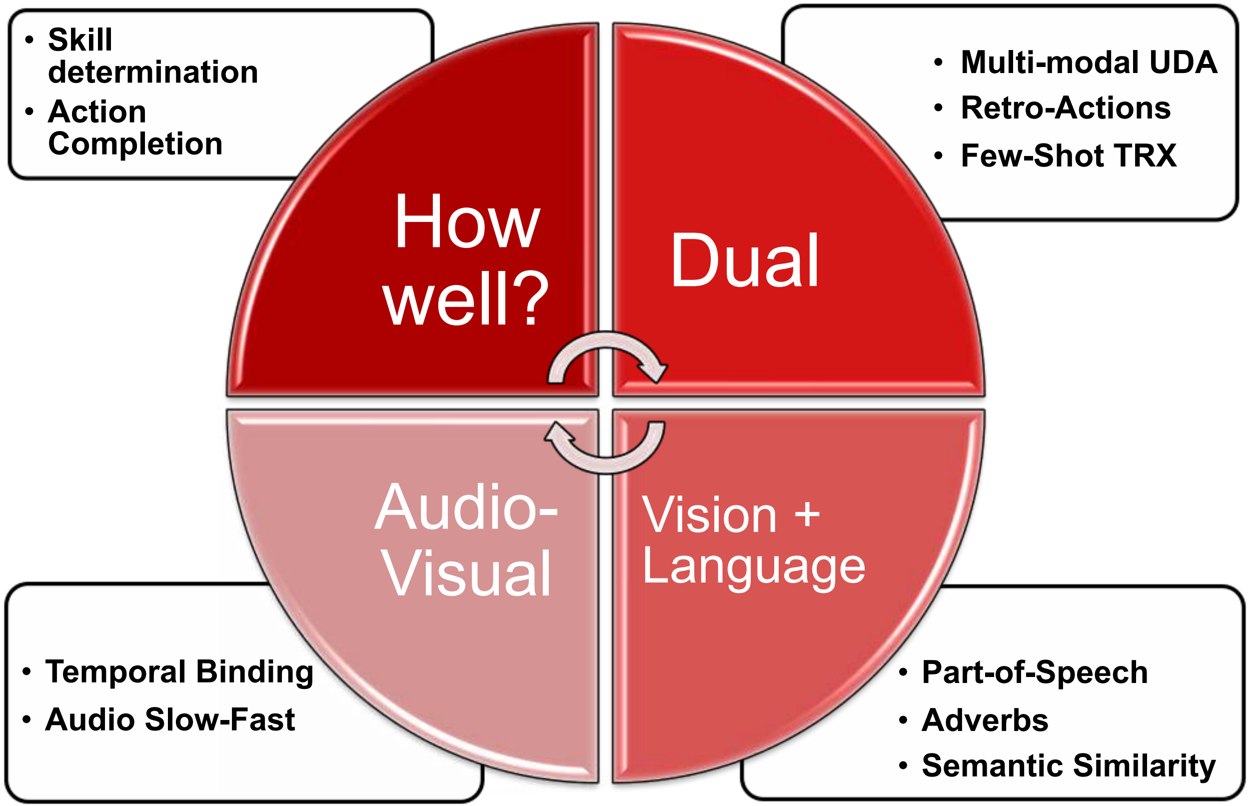
- Temporal Binding
- Audio Slow-Fast

Vision + Language

- Part-of-Speech
- Adverbs
- Semantic Similarity

ICCV19
CVPR20
CVPR21

VU - An Egocentric Perspective

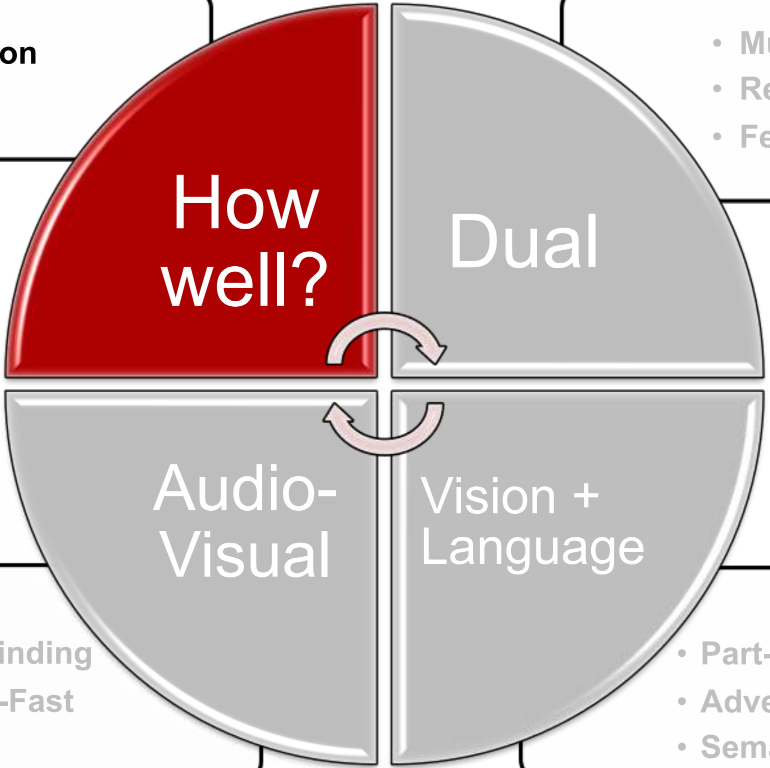


University of BRISTOL

VU - An Egocentric Perspective

- **Skill determination**
- Action Completion

- Multi-modal UDA
- Retro-Actions
- Few-Shot TRX



- Temporal Binding
- Audio Slow-Fast

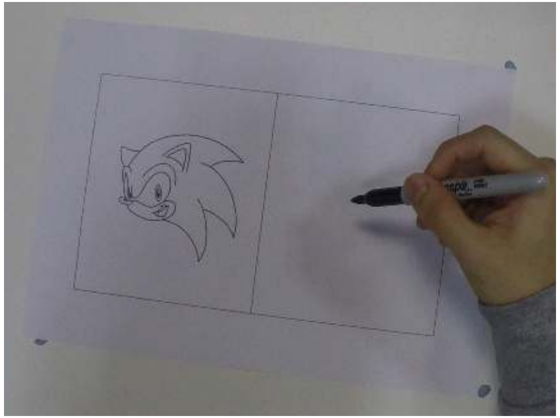
- Part-of-Speech
- Adverbs
- Semantic Similarity

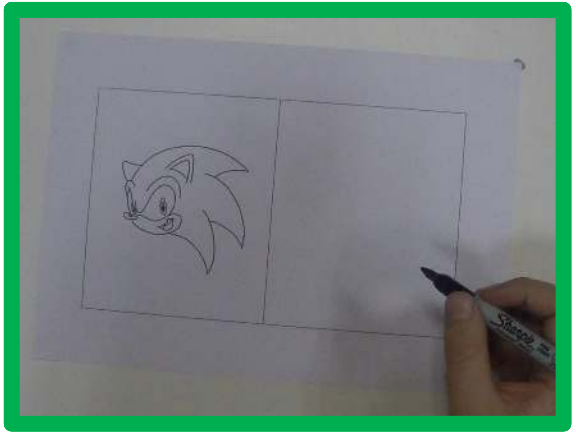
Skill determination in video

with: Hazel Doughty
Walterio Mayol-Cuevas

Assess relative skill for a collection of video sequences, applicable to a variety of tasks.

Input: Pairwise annotations of videos, indicating higher skill or no skill preference

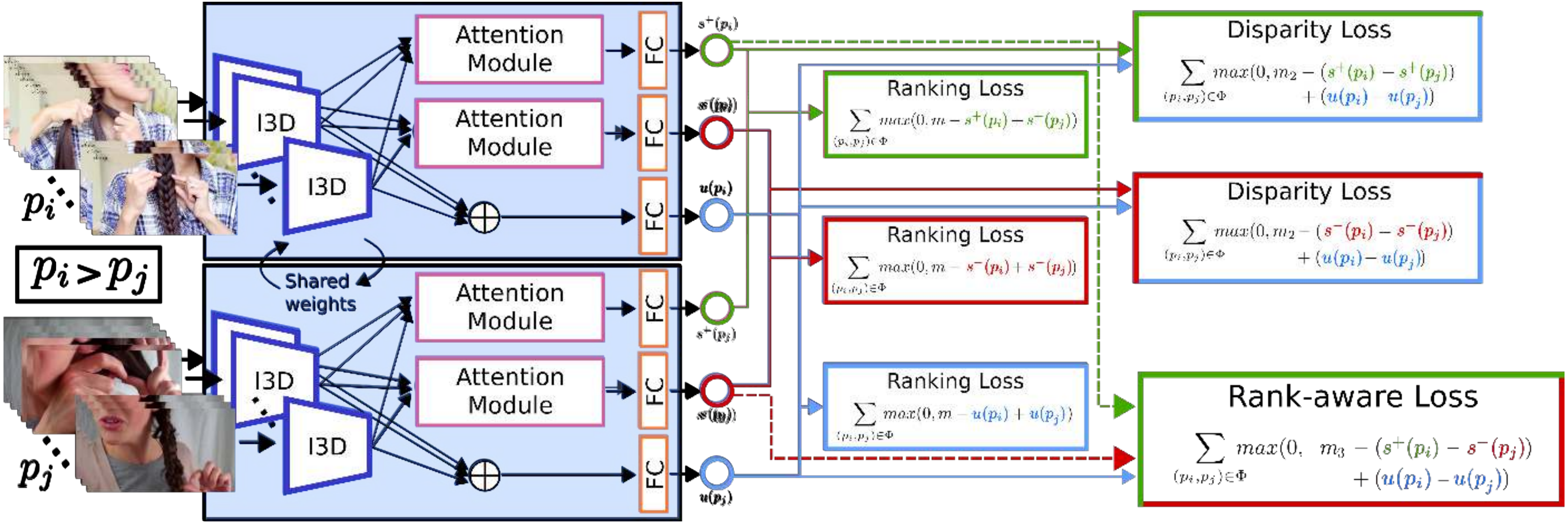




H Doughty, W Mayol-Cuevas, D Damen (2019). The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. *Computer Vision and Pattern Recognition (CVPR)*

Skill determination in video

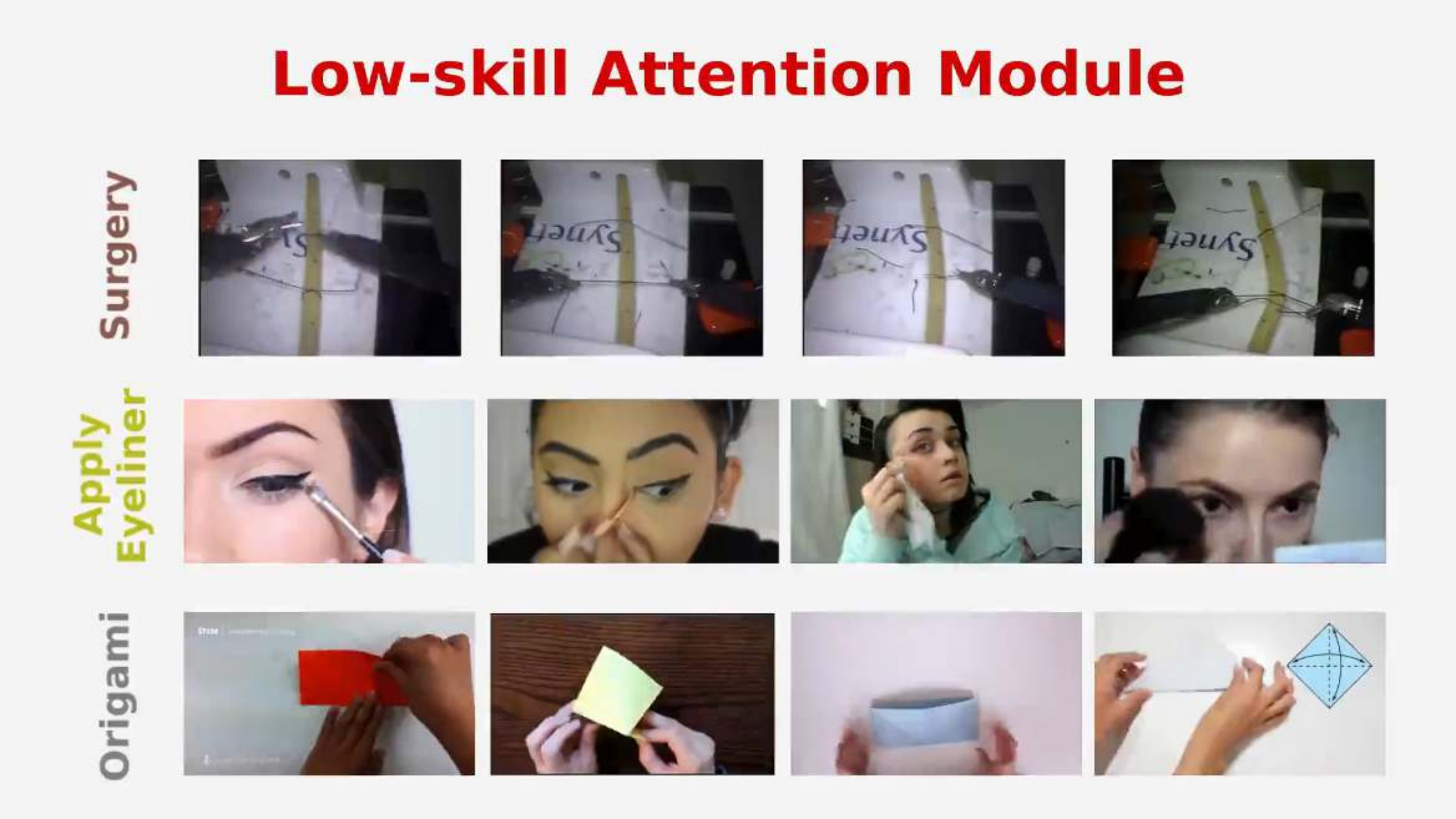
with: Hazel Doughty
Walterio Mayol-Cuevas



H Doughty, W Mayol-Cuevas, D Damen (2019). The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. *Computer Vision and Pattern Recognition (CVPR)*

Skill determination in video

with: Hazel Doughty
Walterio Mayol-Cuevas



H Doughty, W Mayol-Cuevas, D Damen (2019). The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. *Computer Vision and Pattern Recognition (CVPR)*

Skill determination in video

with: Hazel Doughty
Walterio Mayol-Cuevas

University of BRISTOL

H Doughty, W Mayol-Cuevas, D Damen (2019). The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. *Computer Vision and Pattern Recognition (CVPR)*

Skill determination in video

with: Hazel Doughty
Walterio Mayol-Cuevas

Computer Vision and Pattern Recognition (CVPR) 2019

The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos

Hazel Doughty   Walterio Mayol-Cuevas   Dima Damen

University of Bristol

ABSTRACT   VIDEO   DOWNLOADS   BIBTEX   RELATED

Abstract

We present a new model to determine relative skill from long videos, through learnable temporal attention modules. Skill determination is formulated as a ranking problem, making it suitable for common and generic tasks. However, for long videos, parts of the video are irrelevant for assessing skill, and there may be variability in the skill exhibited throughout a video. We therefore propose a method which assesses the relative overall level of skill in a long video by attending to its skill-relevant parts.

Our approach trains temporal attention modules, learned with only video-level supervision, using a novel rank-aware loss function. In addition to attending to task-relevant video parts, our proposed loss jointly trains two attention modules to separately attend to video parts which are indicative of higher (pros) and lower (cons) skill. We evaluate our approach on the EPIC-Skills dataset and additionally annotate a larger dataset from YouTube videos for skill determination with five previously unexplored tasks. Our method outperforms previous approaches and classic softmax attention on both datasets by over 4% pairwise accuracy, and as much as 12% on individual tasks. We also demonstrate our model's ability to attend to

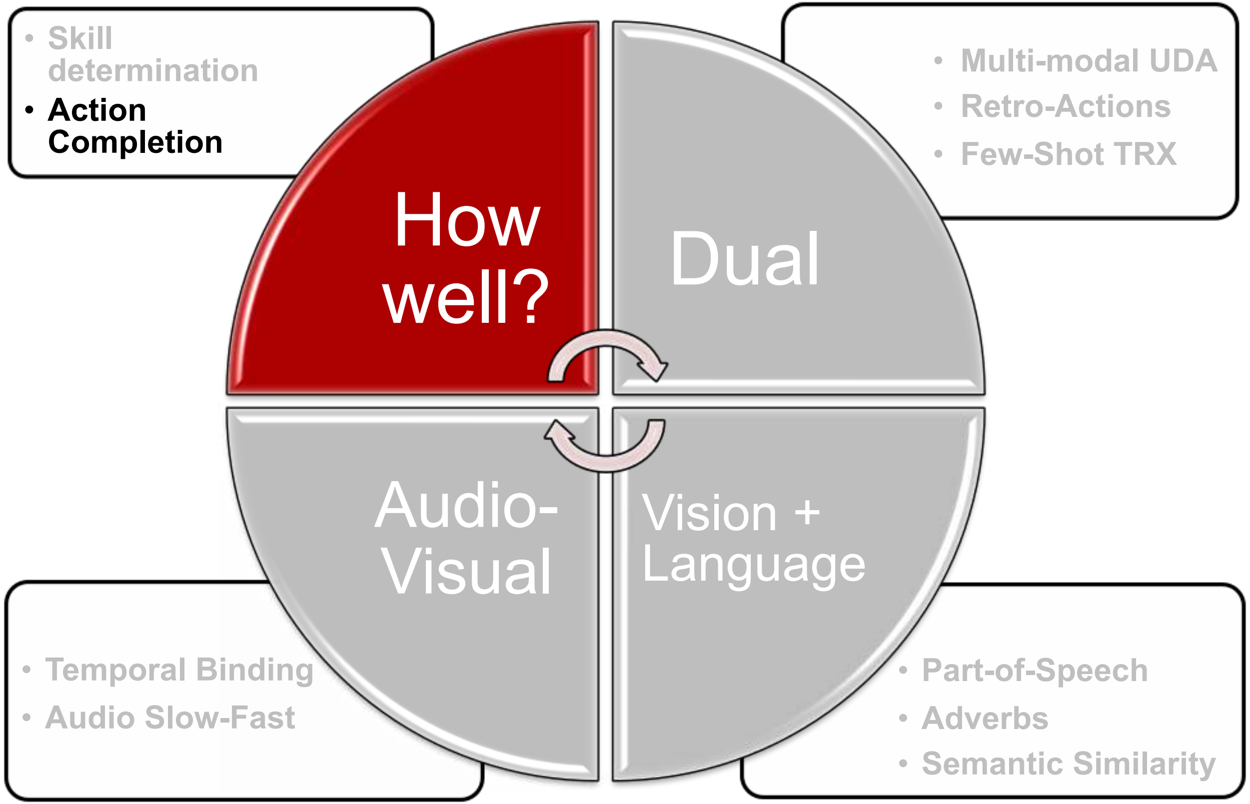
Downloads

- Paper [PDF] [ArXiv]
- Supplementary [Video]
- Code and data [GitHub - Available Now]

University of BRISTOL

H Doughty, W Mayol-Cuevas, D Damen (2019). The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. *Computer Vision and Pattern Recognition (CVPR)*

Fine-grained in Video?



Action Completion Detection

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Ours

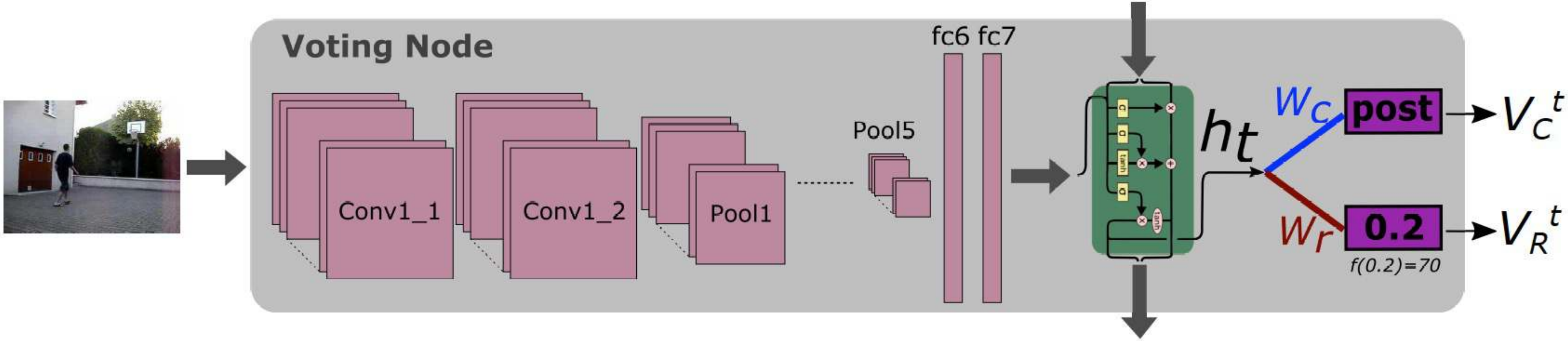
Ground truth

F Heidarivincheh, M Mirmehdi, D Damen (2018). Action Completion: A Temporal Model for Moment Detection. BMVC

Action Completion Detection

with: Farnoosh Heidarivincheh
Majid Mirmehdi

- Each frame in the sequence, contributes to the completion moment detection via ‘voting’



F Heidarivincheh, M Mirmehdi, D Damen (2018). Action Completion: A Temporal Model for Moment Detection. BMVC

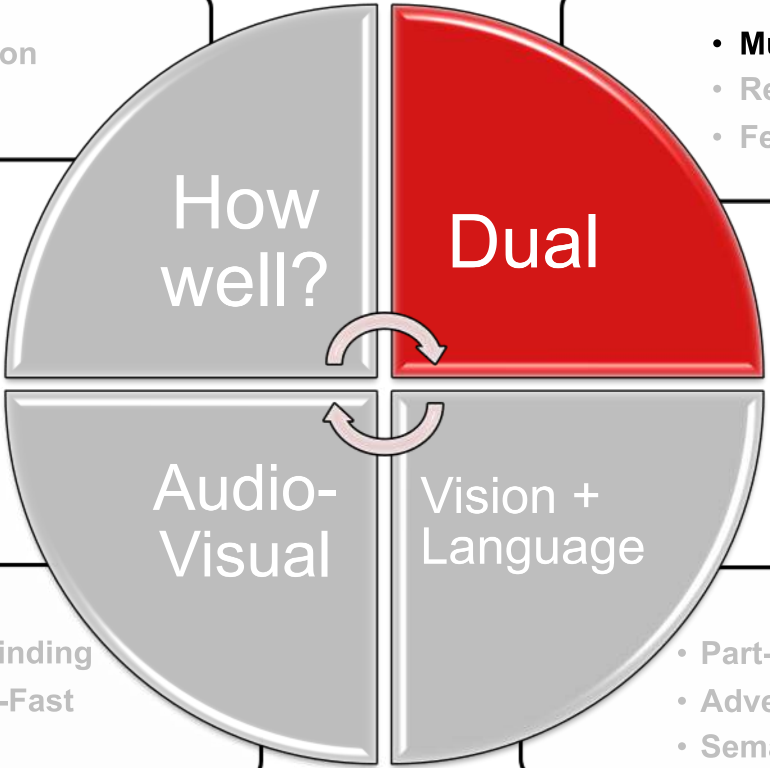
Action Completion Detection

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Ours
GT

F Heidarivincheh, M Mirmehdi, D Damen (2018). Action Completion: A Temporal Model for Moment Detection. BMVC

VU - An Egocentric Perspective



- Skill determination
- Action Completion

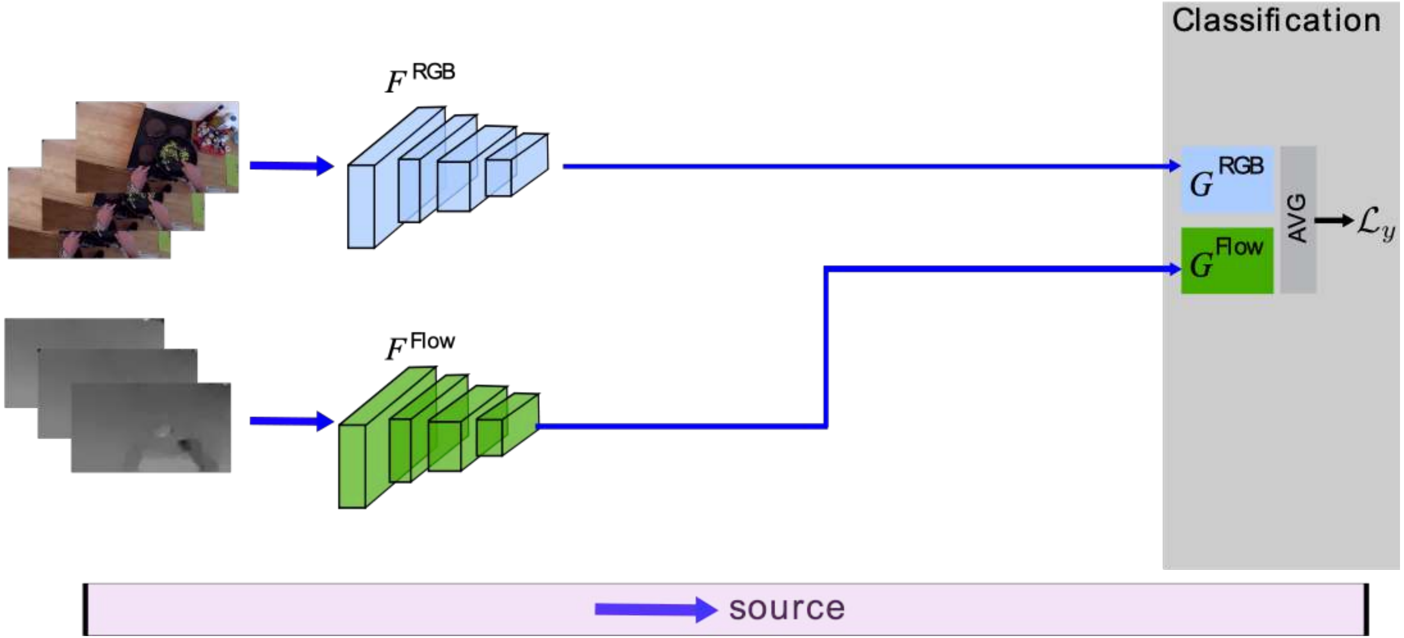
- **Multi-modal UDA**
- Retro-Actions
- Few-Shot TRX

- Temporal Binding
- Audio Slow-Fast

- Part-of-Speech
- Adverbs
- Semantic Similarity

Multi-modal UDA

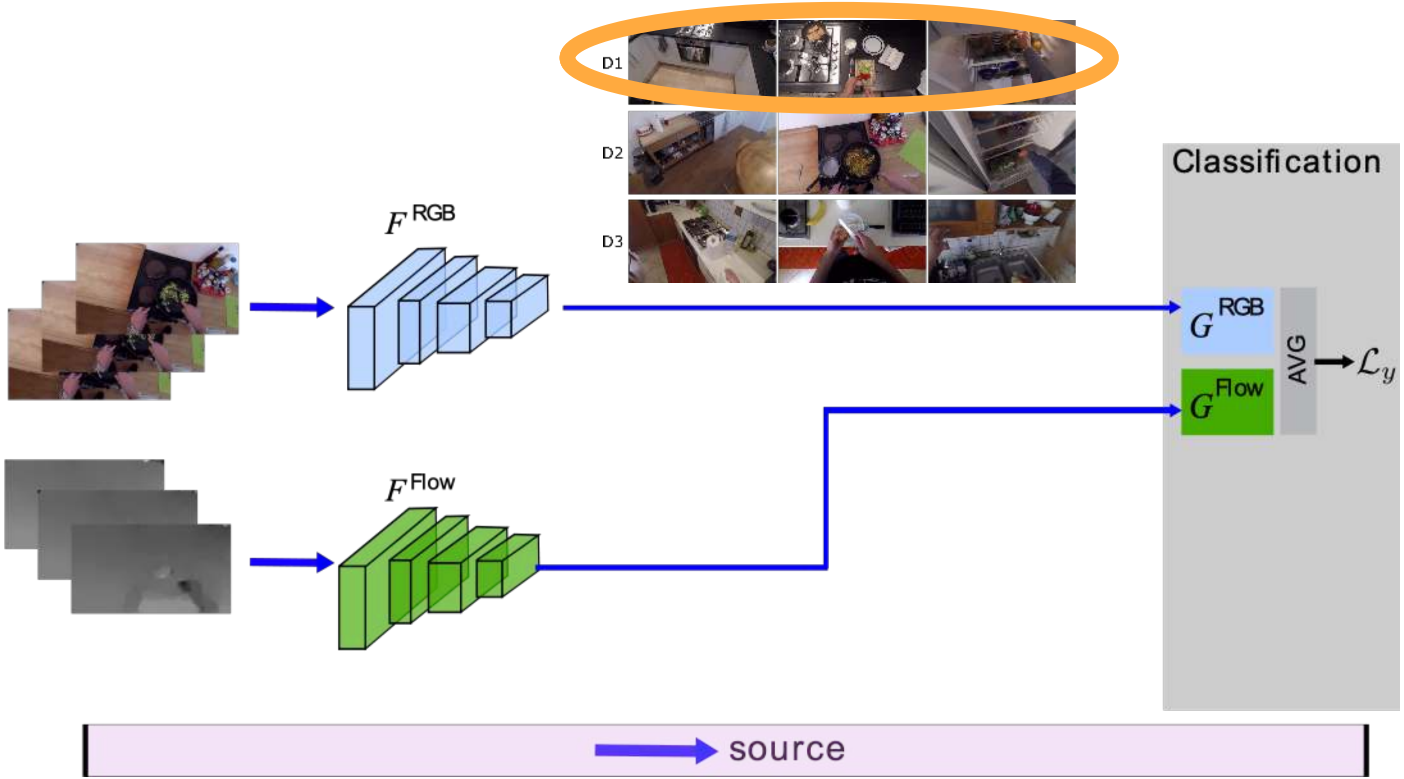
with: Jonathan Munro



J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

Multi-modal UDA

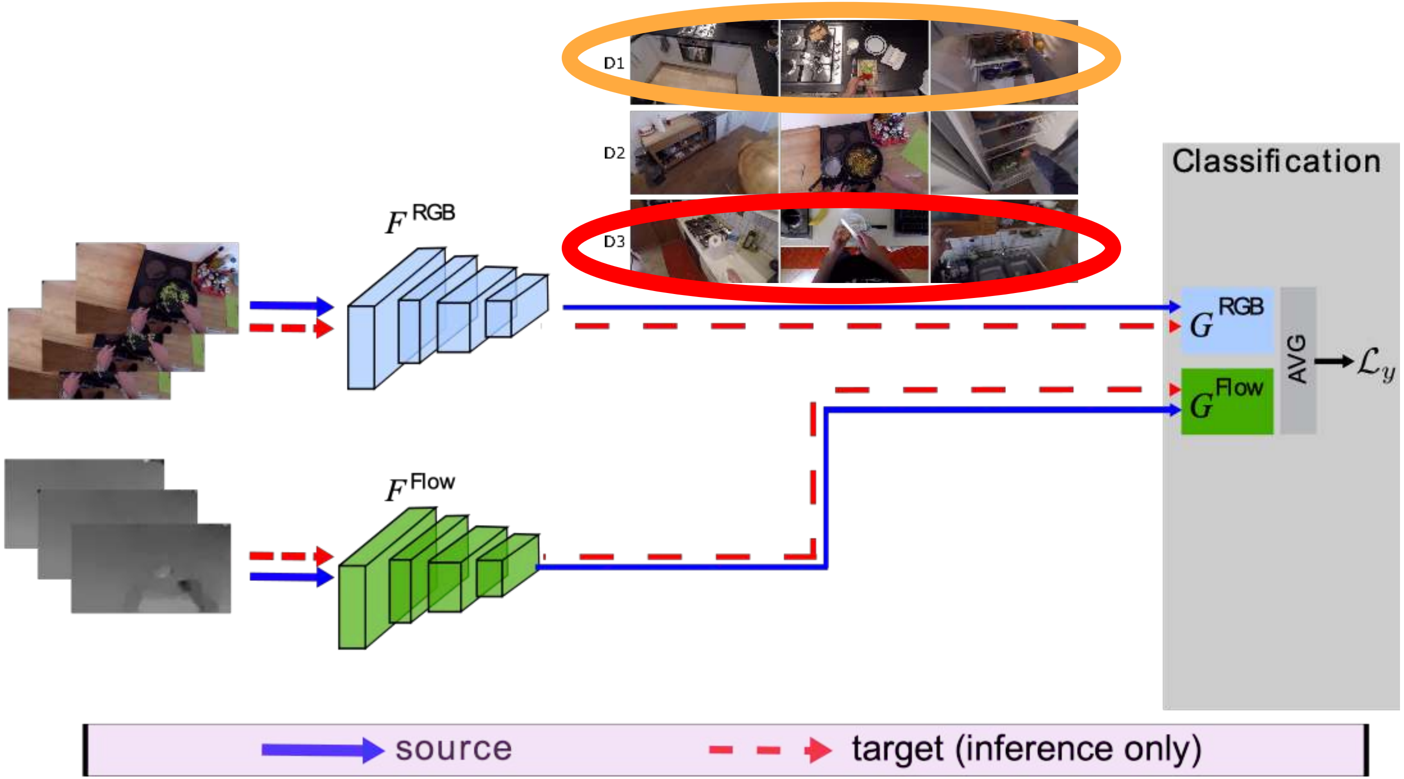
with: Jonathan Munro



J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

Multi-modal UDA

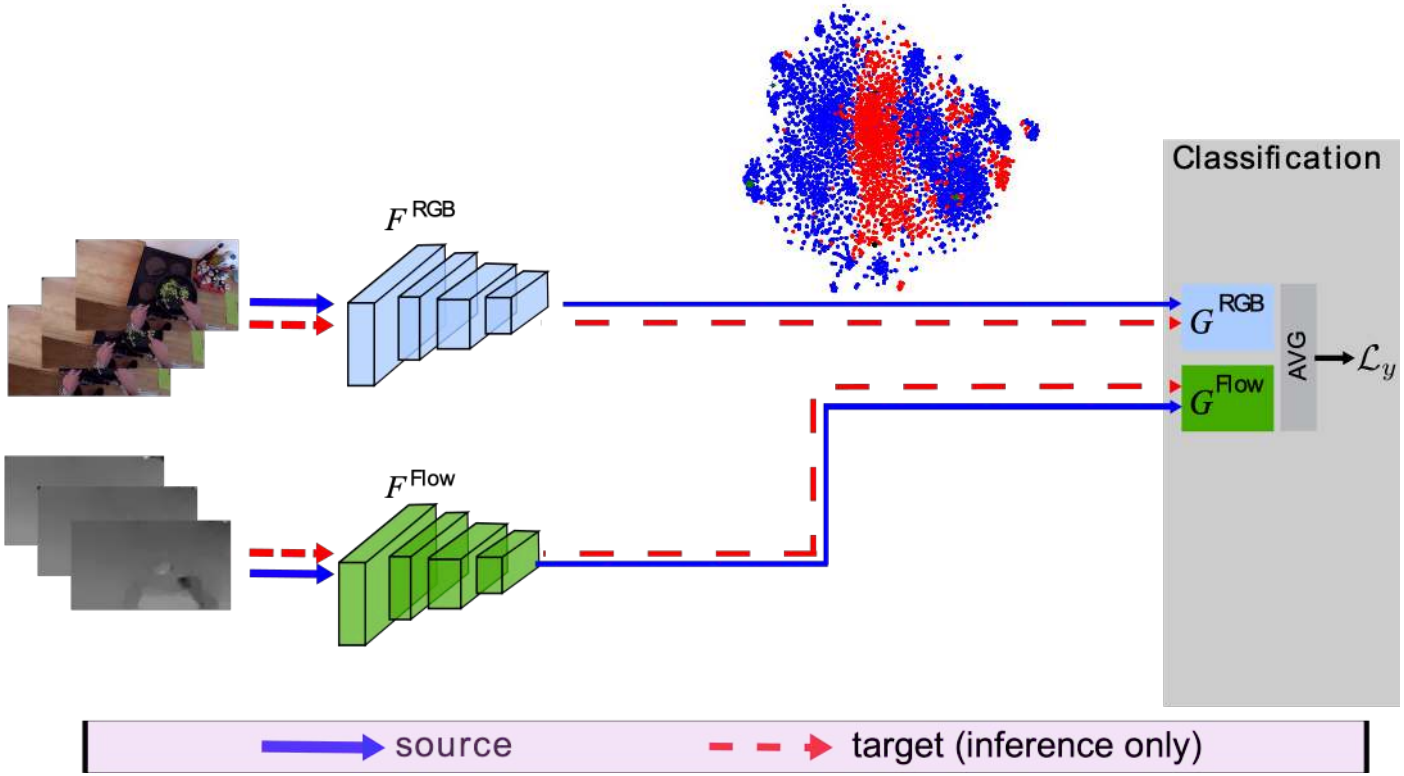
with: Jonathan Munro



J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

Multi-modal UDA

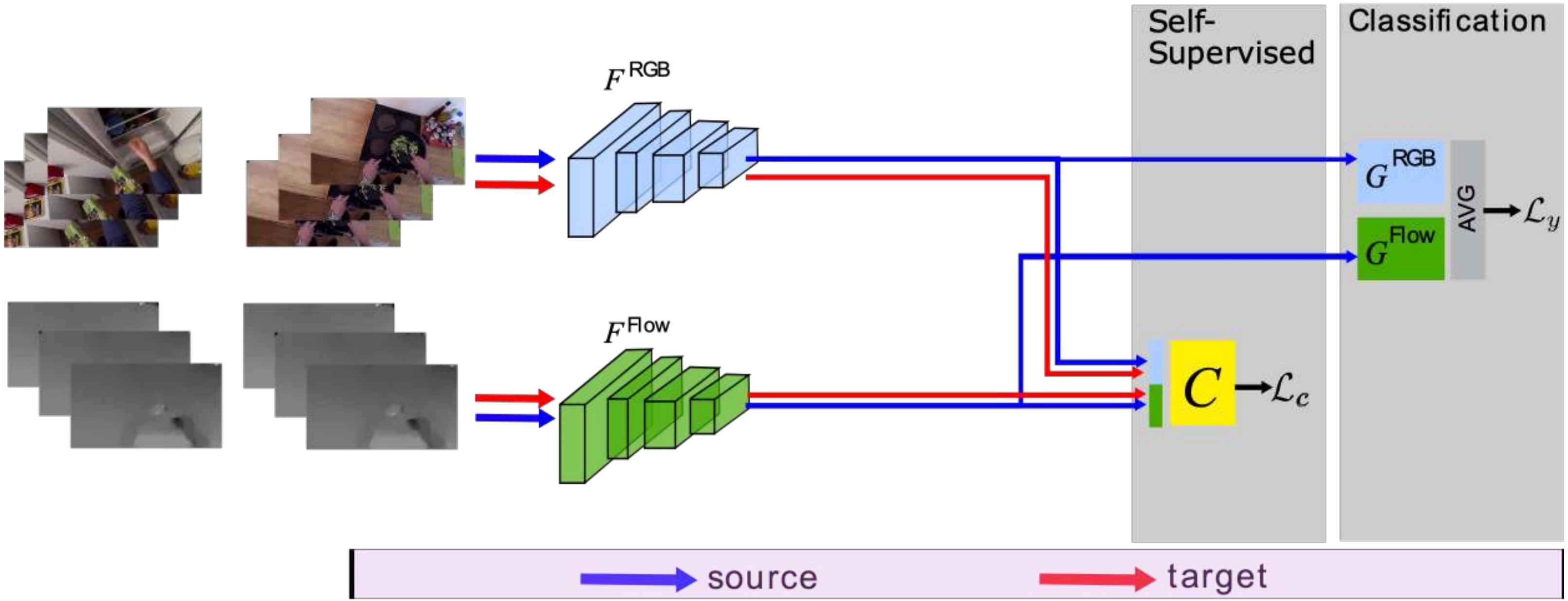
with: Jonathan Munro



J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

Multi-modal UDA

with: Jonathan Munro



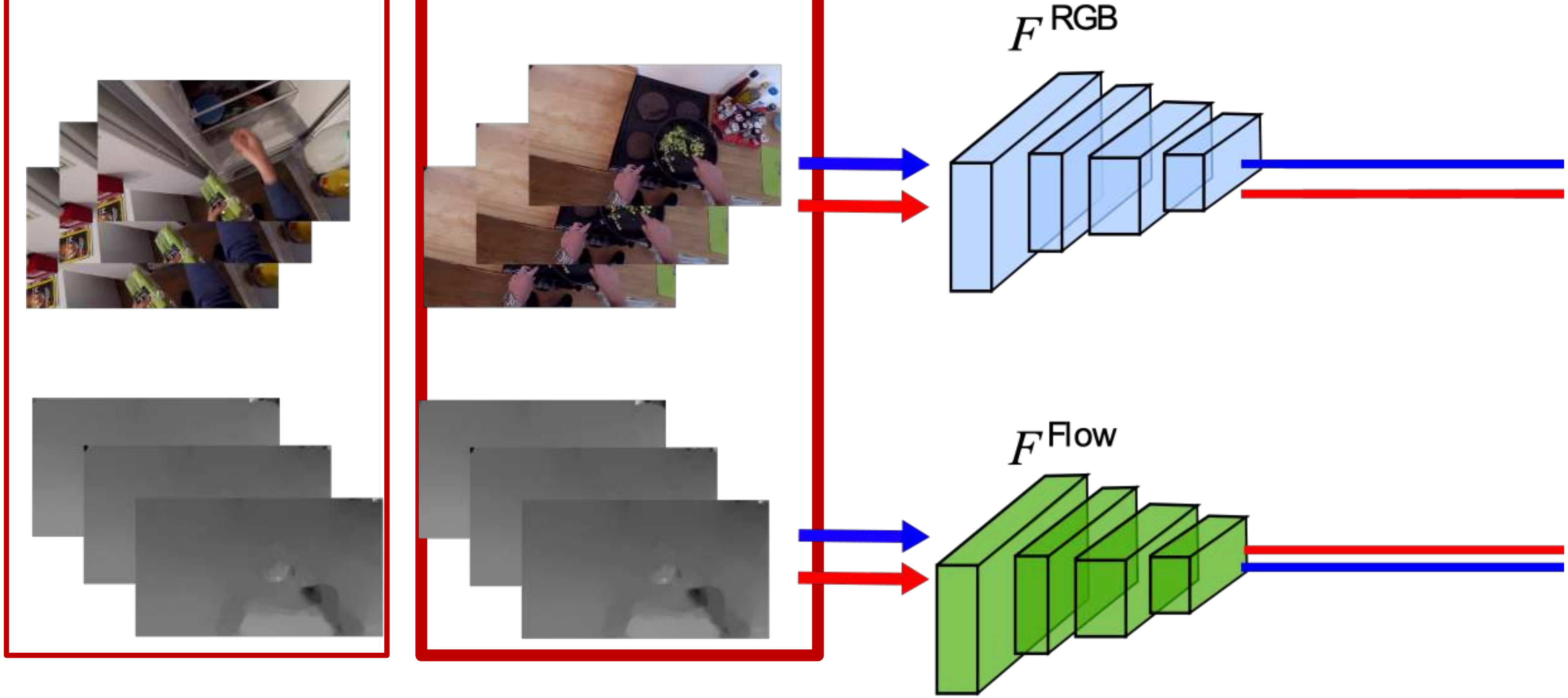
J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

Multi-modal UDA

with: Jonathan Munro

Modalities from diff action

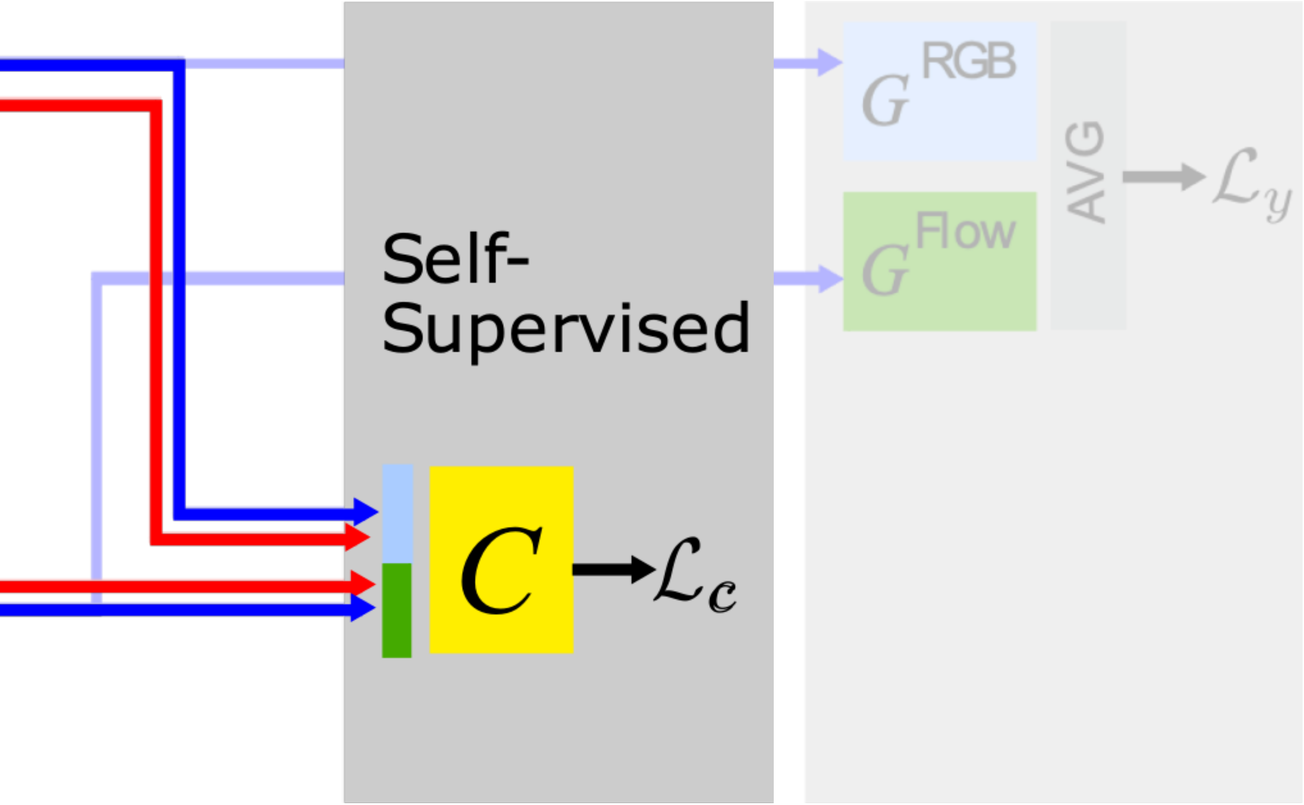
Modalities from same action



J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

Multi-modal UDA

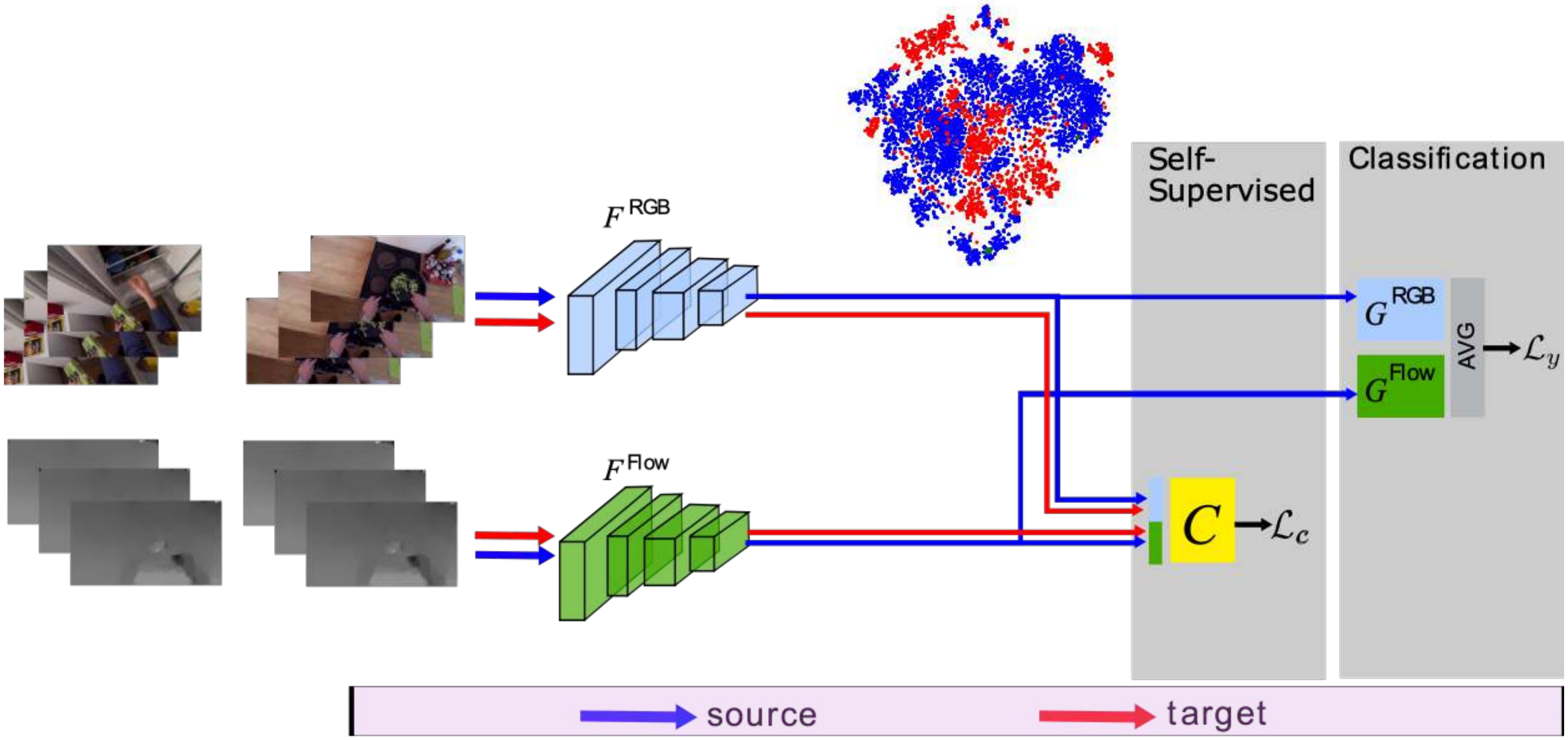
with: Jonathan Munro



J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

Multi-modal UDA

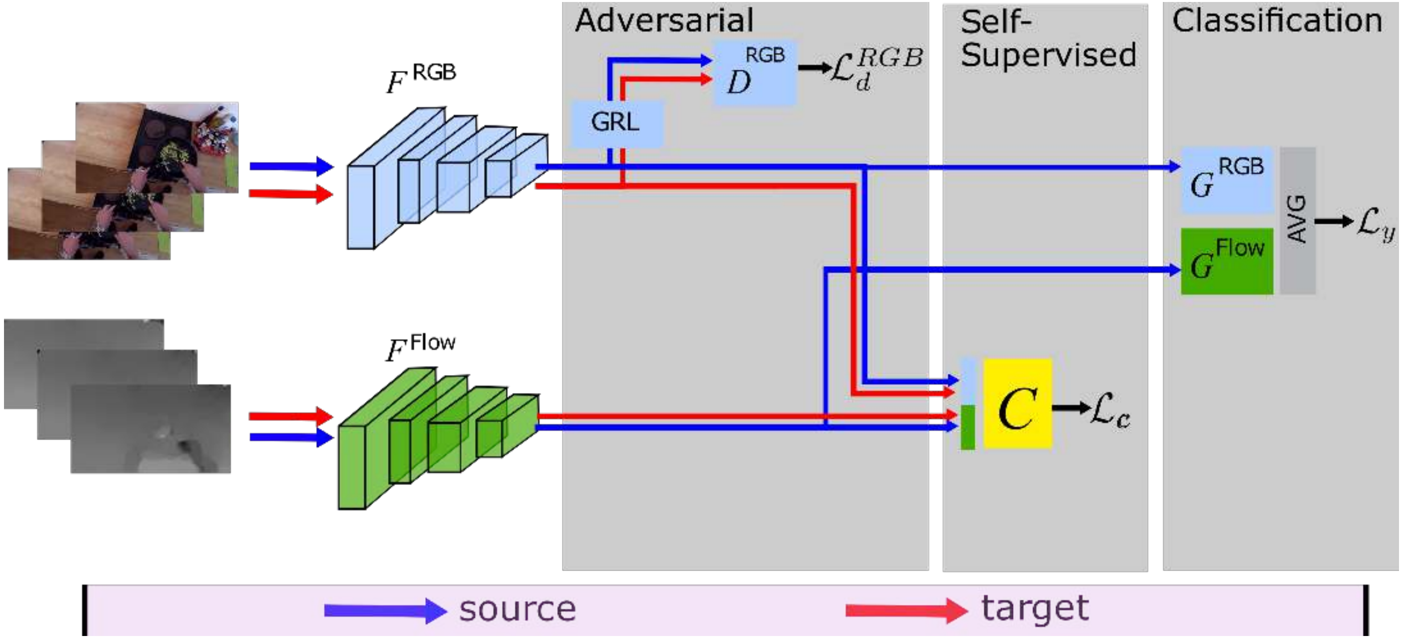
with: Jonathan Munro



J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

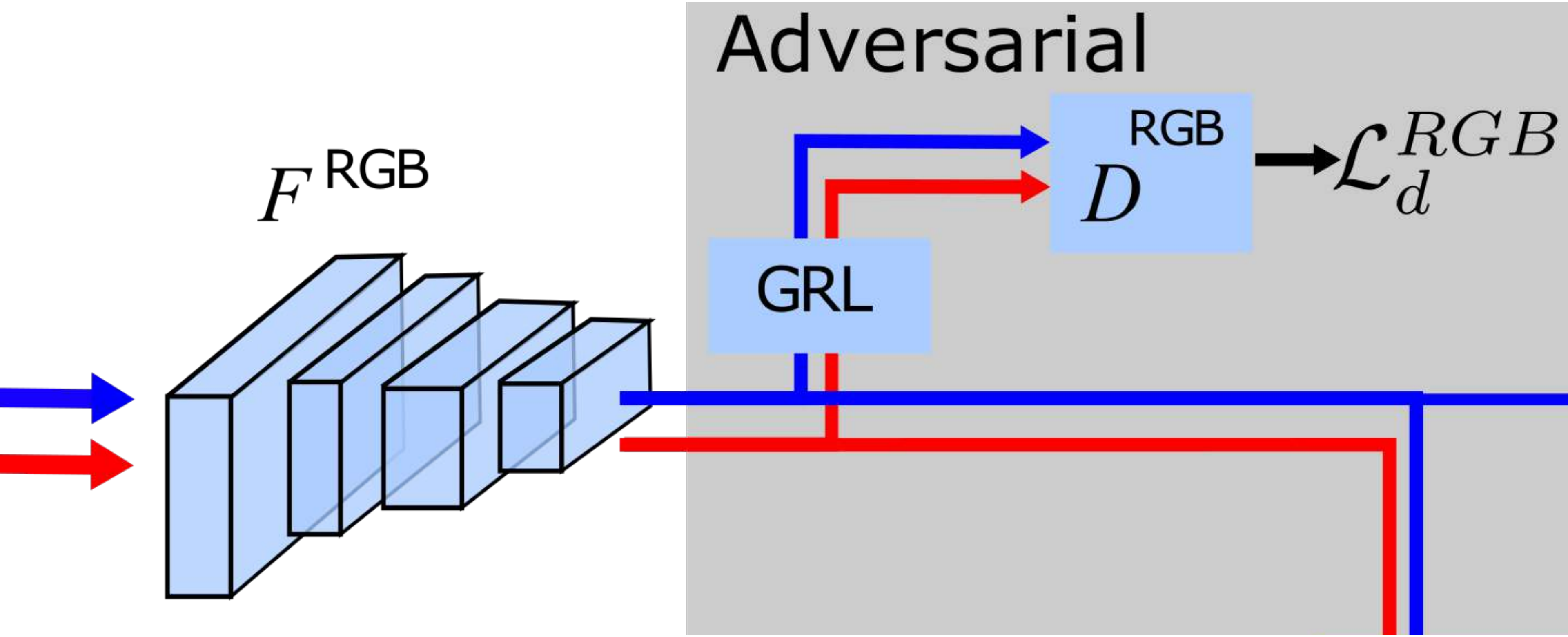
Multi-modal UDA

with: Jonathan Munro



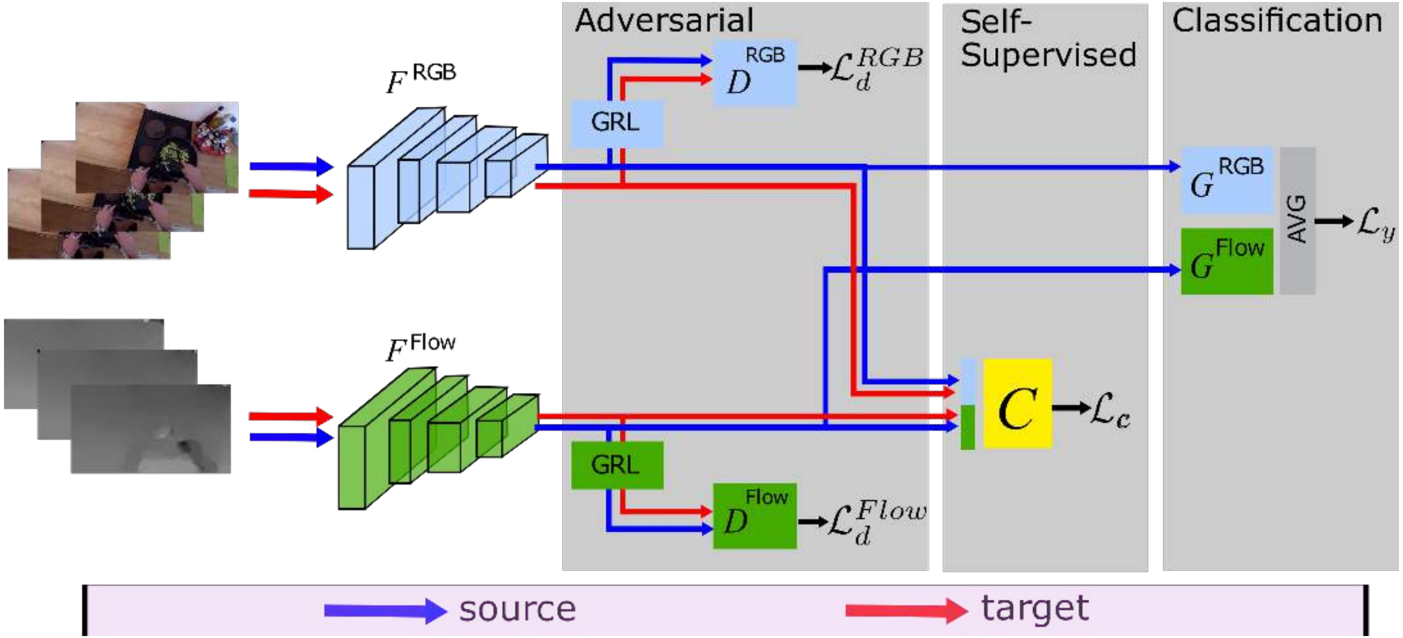
J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

with: Jonathan Munro



J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

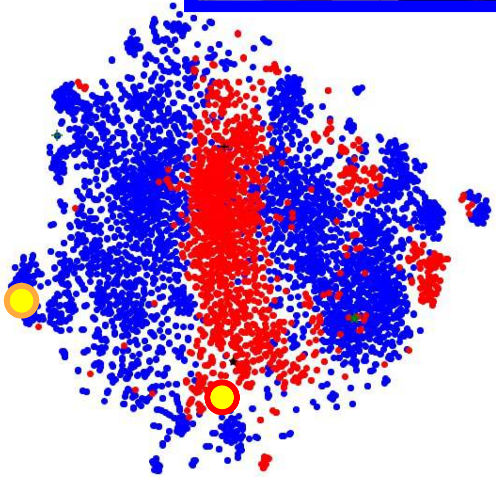
Multi-modal UDA



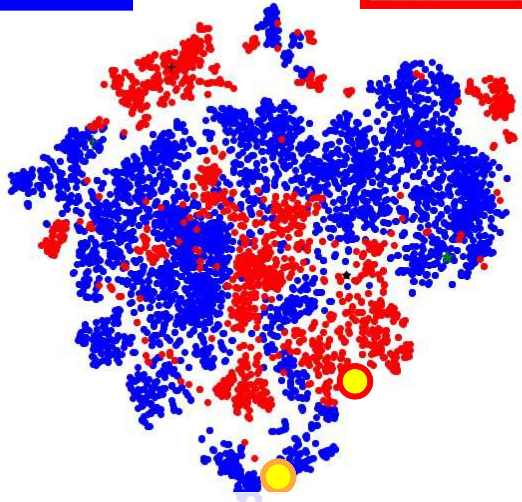
J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

Multi-modal UDA

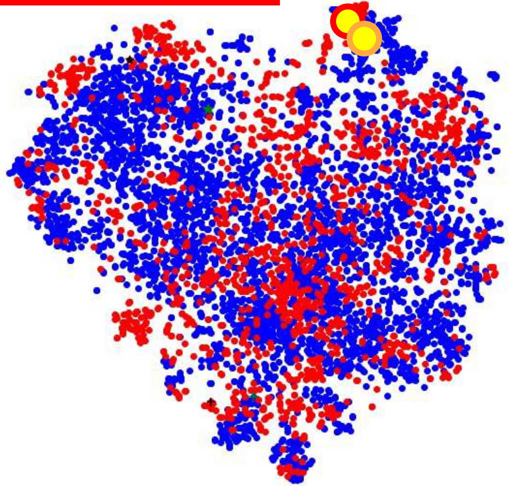
with: Jonathan Munro



Source-Only



Self-Supervision

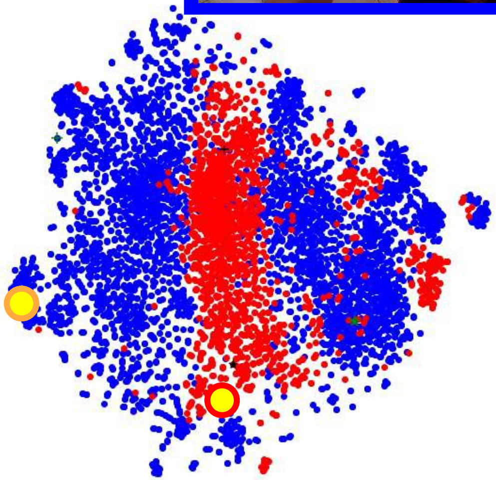


MM-SADA

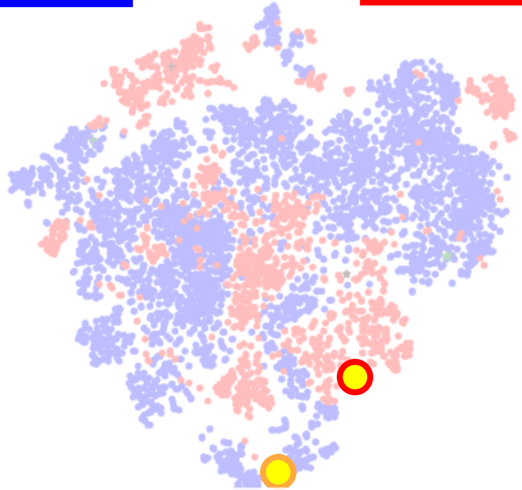
J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

Multi-modal UDA

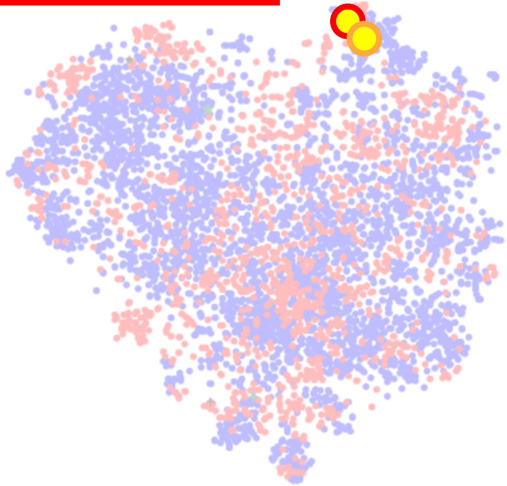
with: Jonathan Munro



Source-Only



Self-Supervision

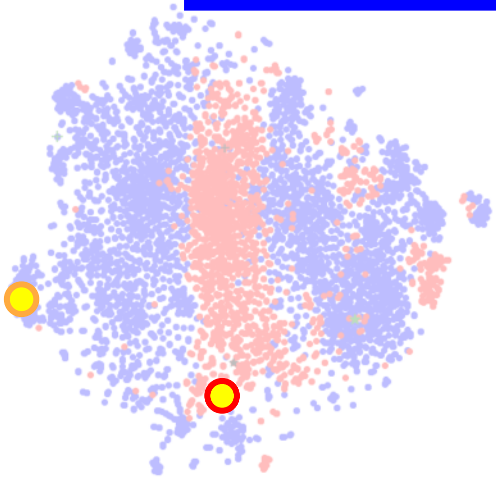


MM-SADA

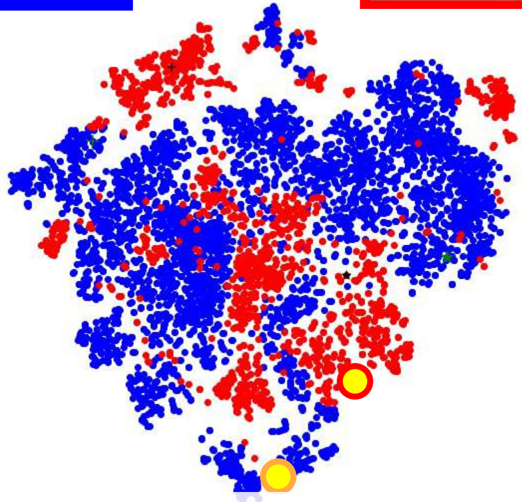
J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

Multi-modal UDA

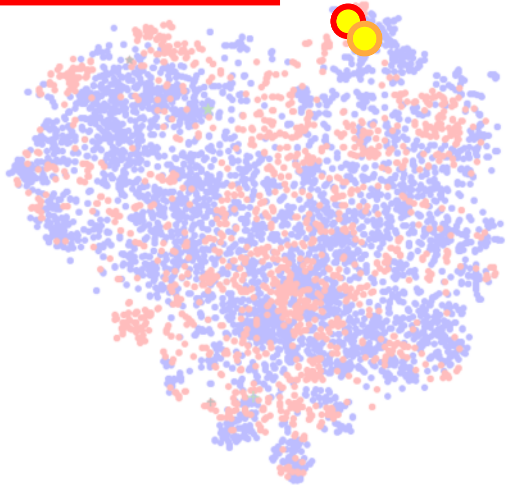
with: Jonathan Munro



Source-Only



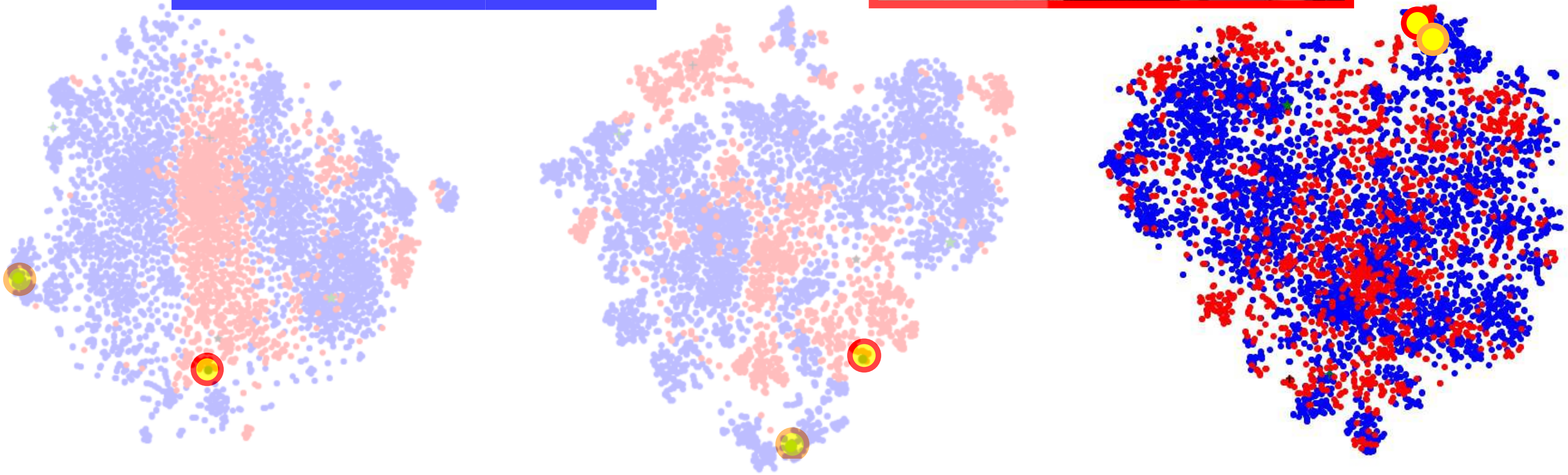
Self-Supervision



MM-SADA

J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

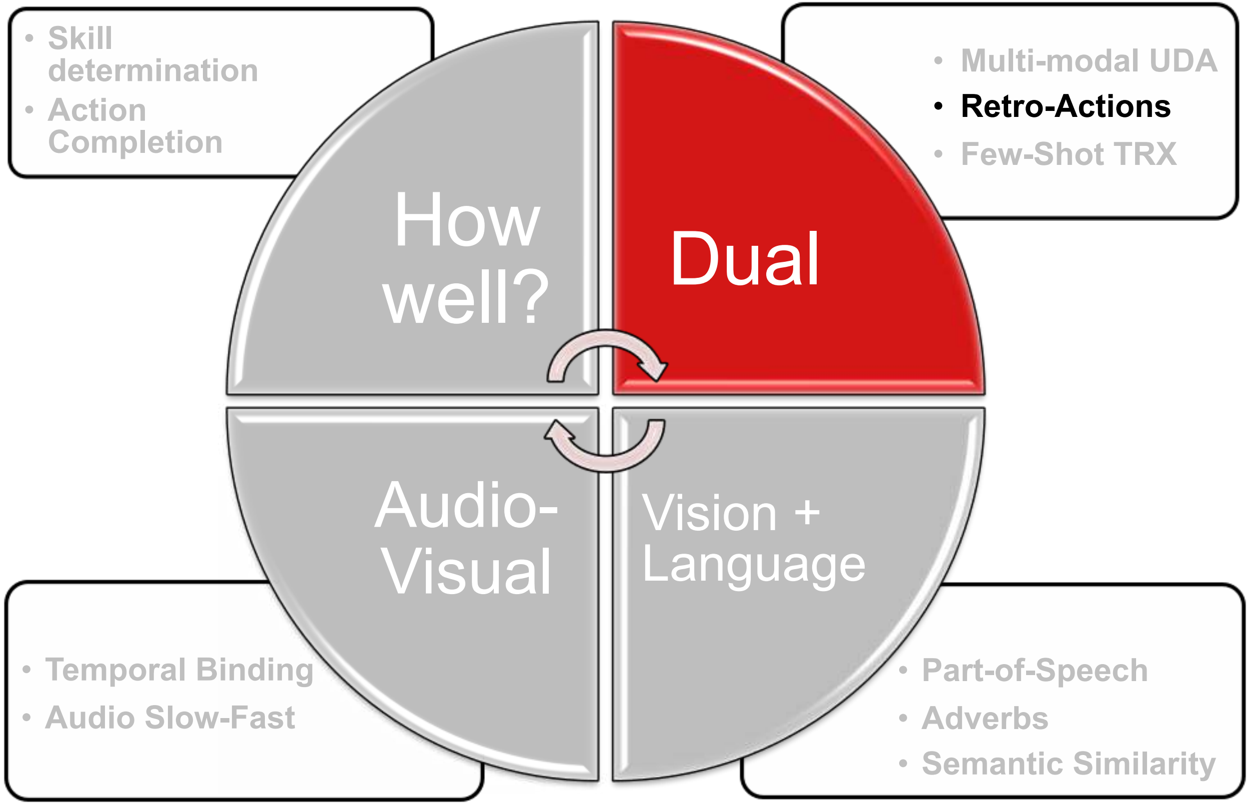
with: Jonathan Munro



Source-Only Self-Supervision **MM-SADA**

J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

VU - An Egocentric Perspective

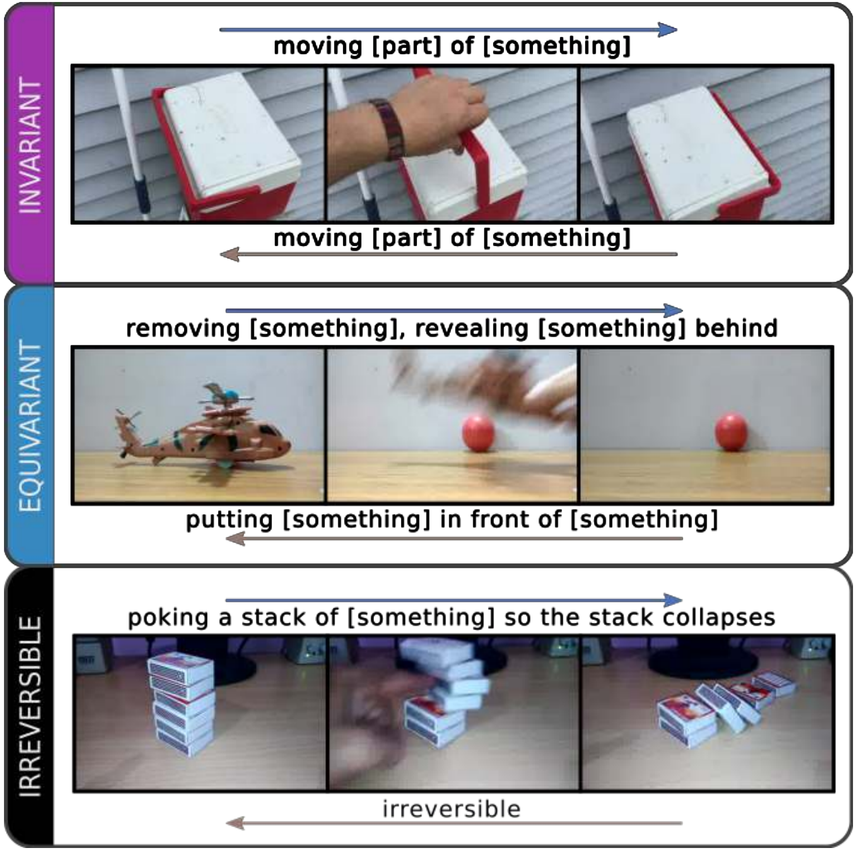


Retro-Actions

W Price, D Damen (2019). Retro-Actions: Learning 'Close' by Time-Reversing 'Open' Videos. ICCV MDALC Workshop

Retro-Actions

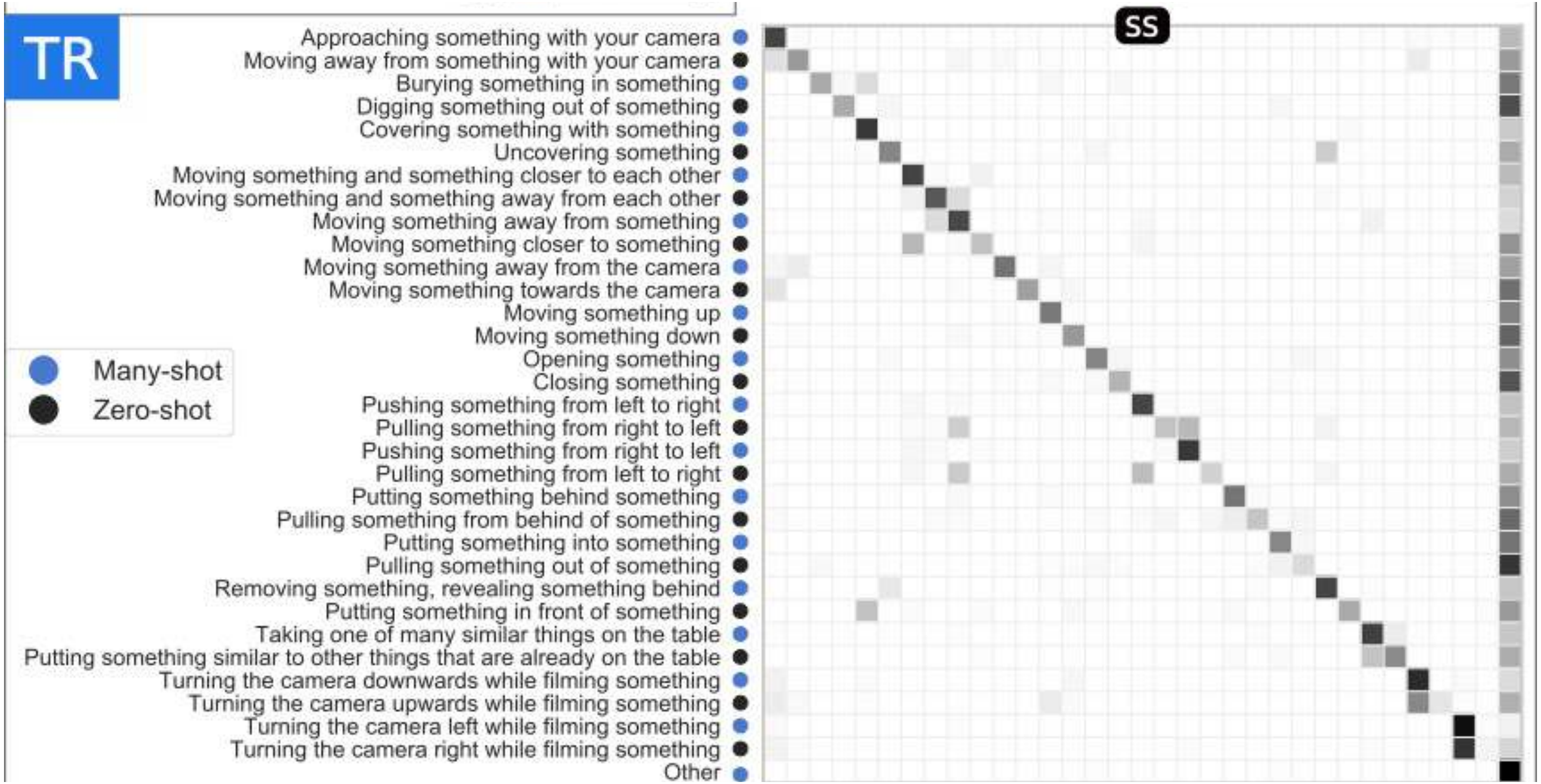
with: Will Price



W Price, D Damen (2019). Retro-Actions: Learning 'Close' by Time-Reversing 'Open' Videos. ICCV MDALC Workshop

Retro-Actions

with: Will Price



W Price, D Damen (2019). Retro-Actions: Learning ‘Close’ by Time-Reversing ‘Open’ Videos. ICCV MDALC Workshop

VU - An Egocentric Perspective

- Skill determination
- Action Completion

How well?

- Multi-modal UDA
- Retro-Actions
- **Few-Shot TRX**

Dual

Audio-Visual

- Temporal Binding
- Audio Slow-Fast

Vision + Language

- Part-of-Speech
- Adverbs
- Semantic Similarity

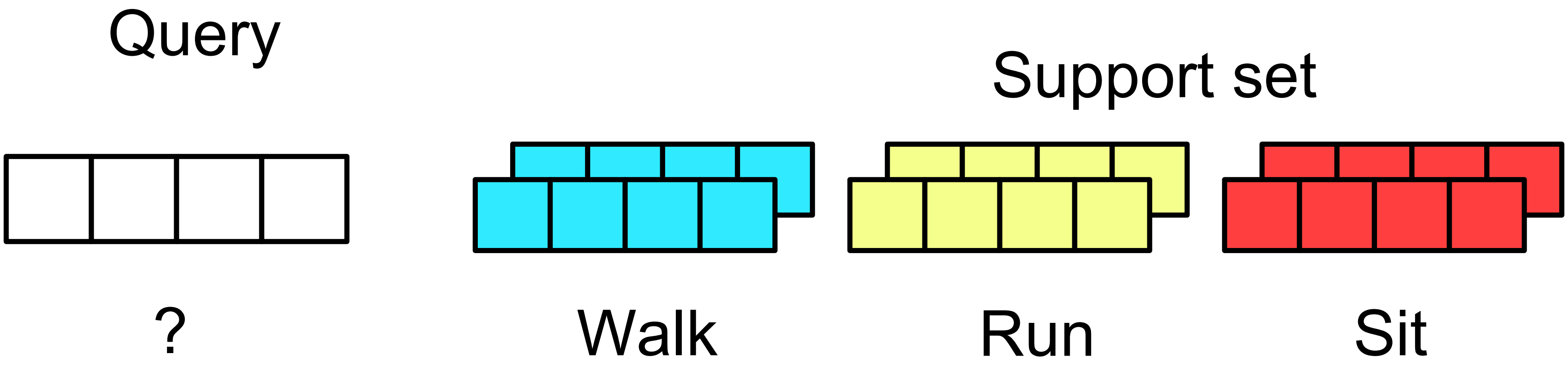
University of BRISTOL

- 2-Shot Recognition?

Perrett et al (2021). Temporal-Relational CrossTransformers for Few-Shot Action Recognition. CVPR

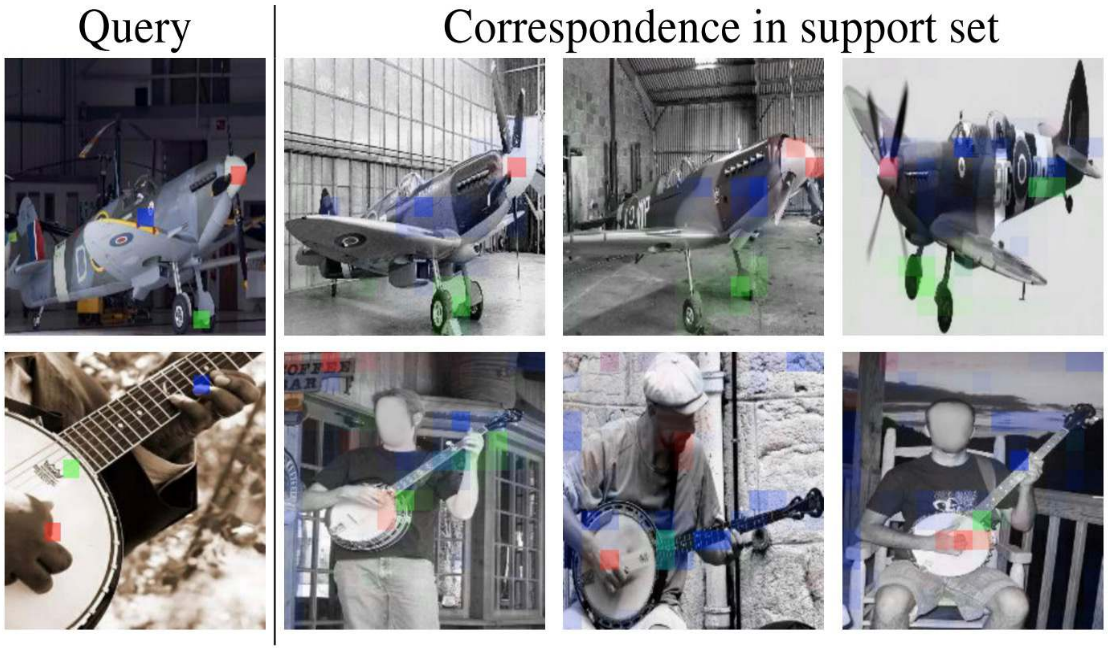
Few-Shot Action Recognition

- Videos in X-Shot are known as the *support set*
- Learn a classifier, using the support set, which can classify query videos.
- All prior works compare the query video to *each video in the support set* separately, including using temporal time warping - for best matching



Perrett et al (2021). Temporal-Relational CrossTransformers for Few-Shot Action Recognition. CVPR

Our Idea



Mechanism to construct class prototypes using relevant frames from the support set.

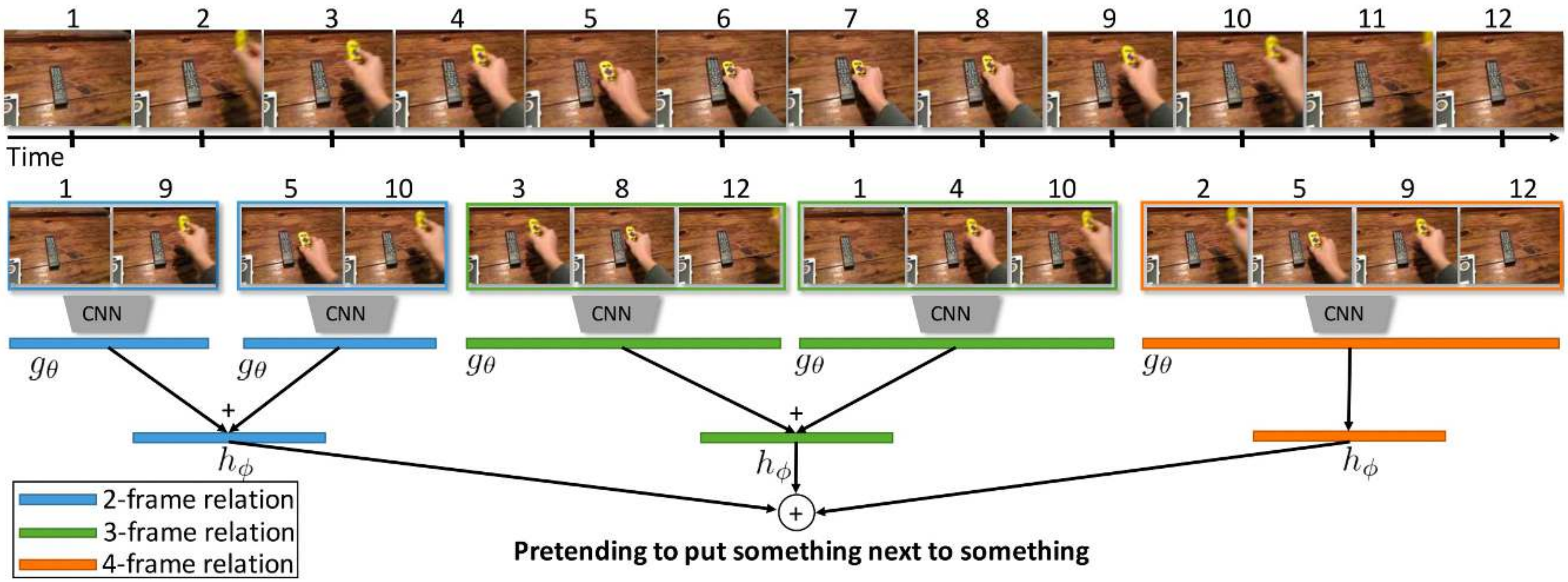
No positional relationships.

How to preserve temporal ordering?

Doersch et al (2020) CrossTransformers: spatially-aware few-shot transfer. ICLR

Our Idea

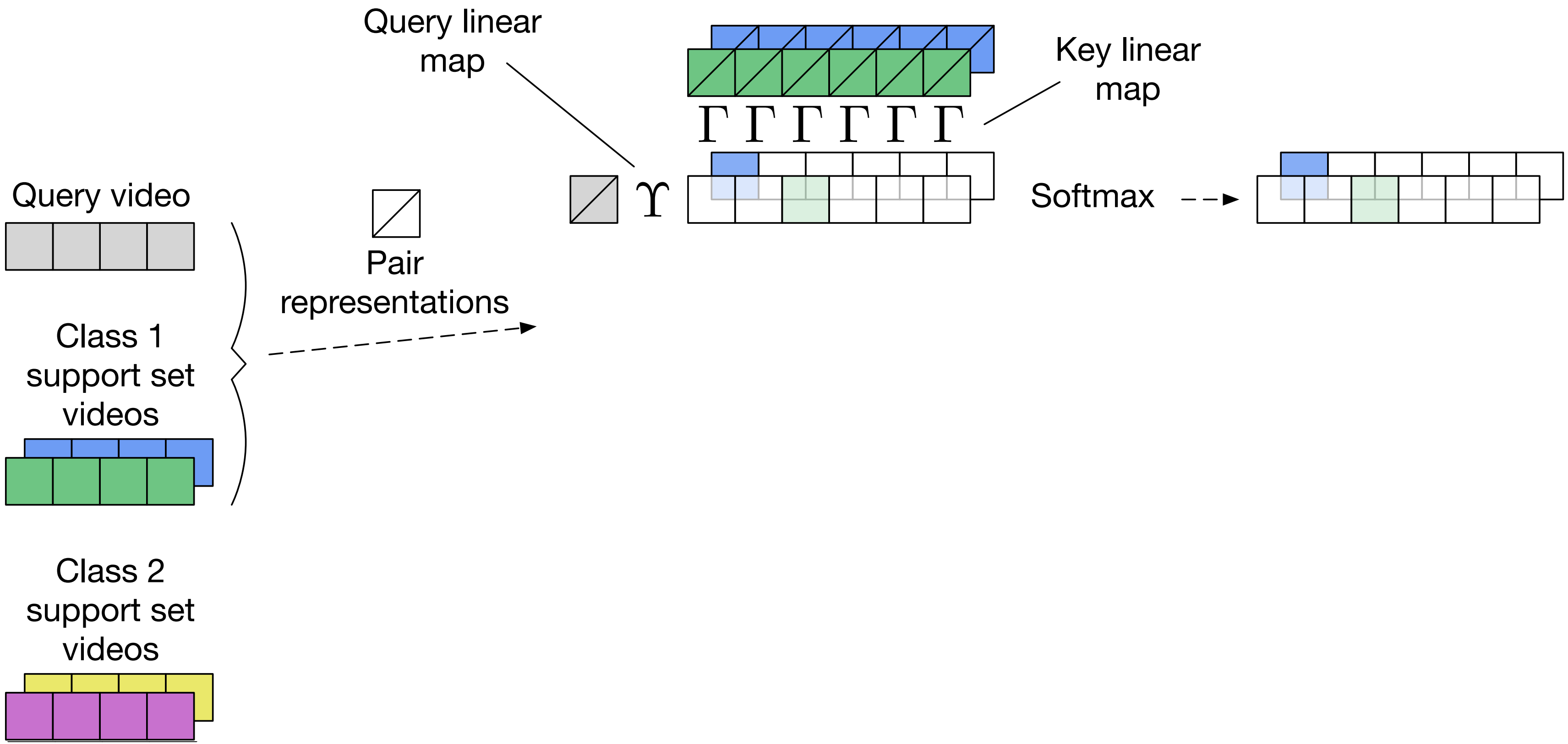
with: Toby Perrett



Zhou et al (2018) Temporal Relational Network. ECCV

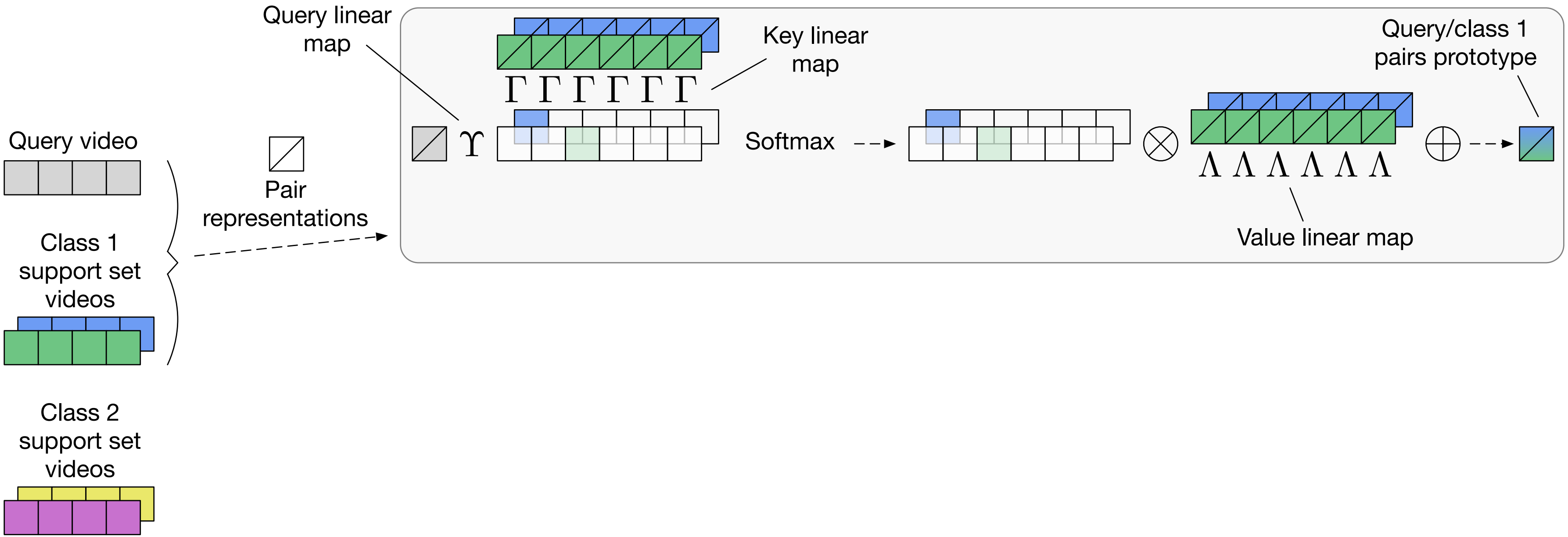
Temporal Relational CrossTransformers (TRX)

with: Toby Perrett



Temporal Relational CrossTransformers (TRX)

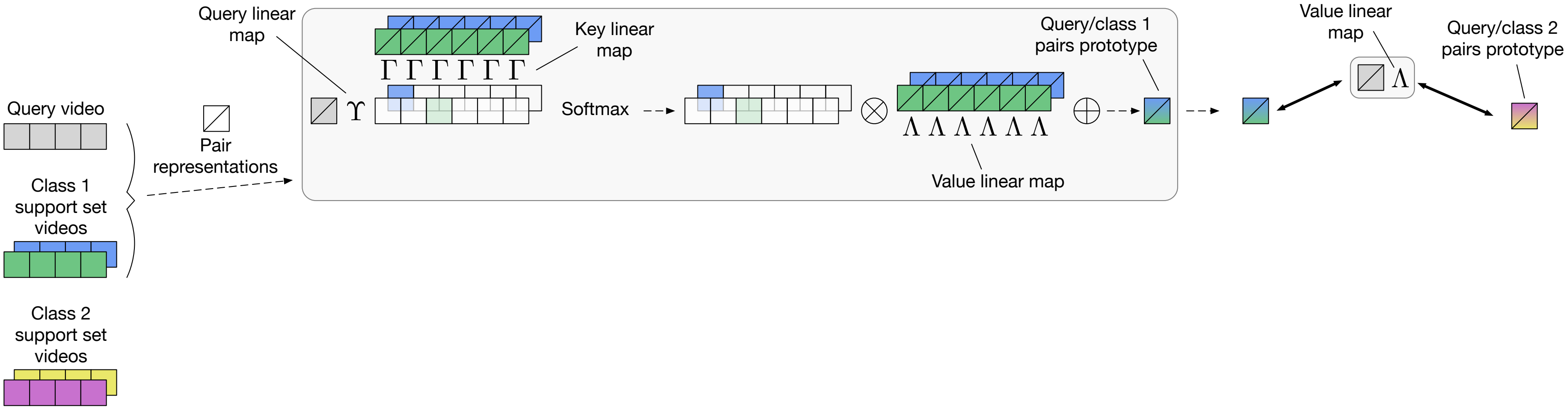
with: Toby Perrett



Perrett et al (2021). Temporal-Relational CrossTransformers for Few-Shot Action Recognition. CVPR

Temporal Relational CrossTransformers (TRX)

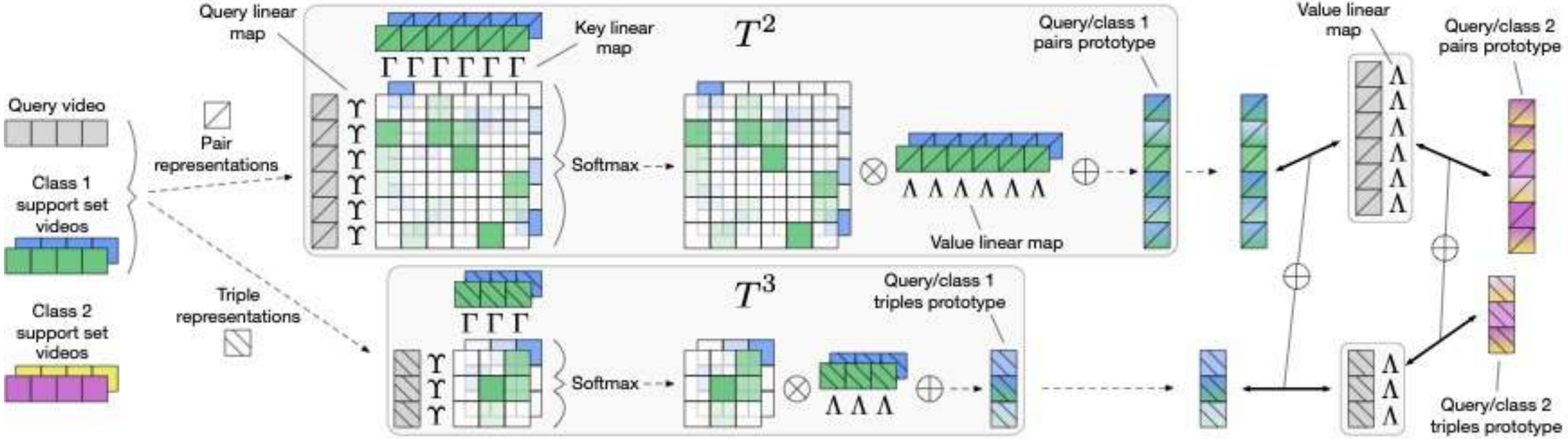
with: Toby Perrett



Perrett et al (2021). Temporal-Relational CrossTransformers for Few-Shot Action Recognition. CVPR

Temporal Relational CrossTransformers (TRX)

with: Toby Perrett



Perrett et al (2021). Temporal-Relational CrossTransformers for Few-Shot Action Recognition. CVPR

Results

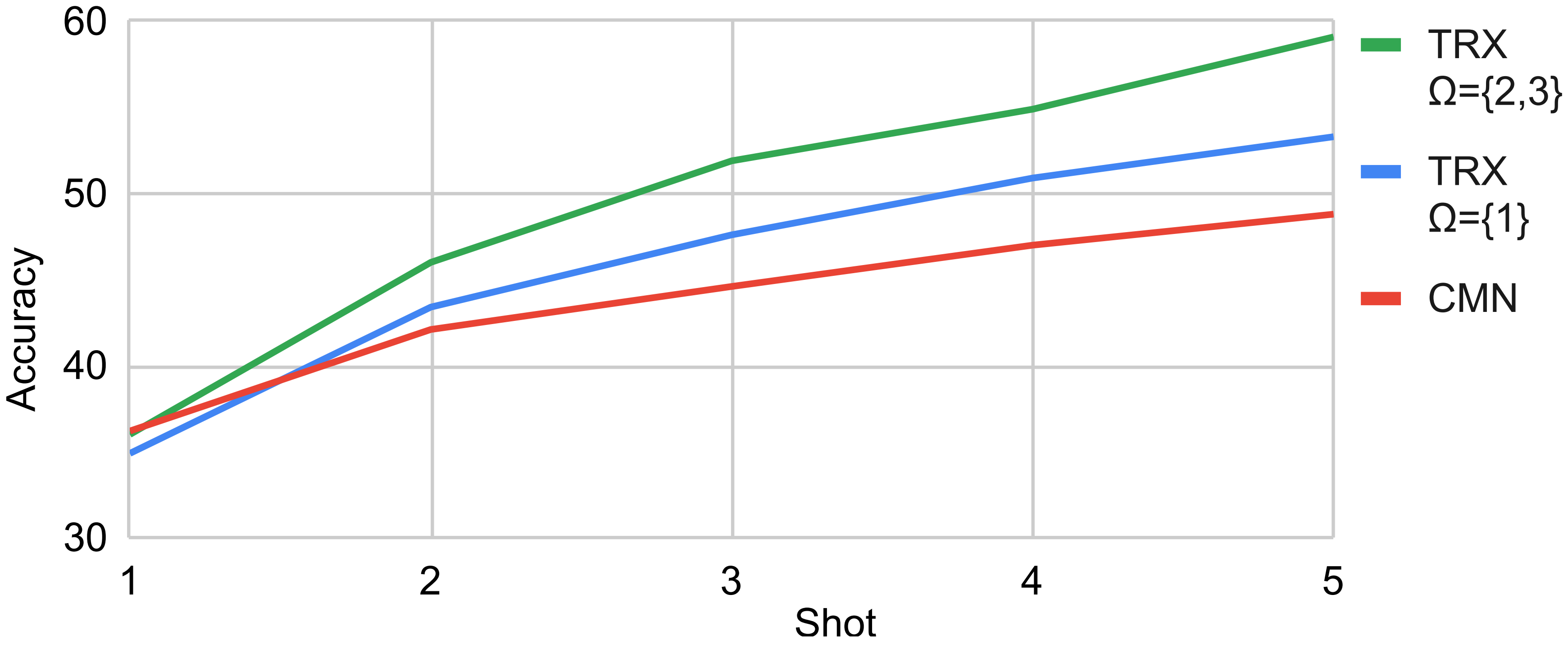
Method	Kinetics	SSv2†	SSv2*	HMDB	UCF
CMN [31]	78.9	-	-	-	-
CMN-J [32]	78.9	48.8	-	-	-
TARN [3]	78.5	-	-	-	-
ARN [27]	82.4	-	-	60.6	83.1
OTAM [4]	85.8	-	52.3	-	-
TRX (Ours)	**85.9**	**59.1**	**64.6**	**75.6**	**96.1**

Table 1: Results on 5-way 5-shot benchmarks of Kinetics (split from [32]), SSv2 (†: split from [32], *: split from [4]), HMDB51 and UCF101 (both splits from [27]).

Perrett et al (2021). Temporal-Relational CrossTransformers for Few-Shot Action Recognition. CVPR

Results

with: Toby Perrett



Perrett et al (2021). Temporal-Relational CrossTransformers for Few-Shot Action Recognition. CVPR

Temporal-Relational CrossTransformers for Few-Shot Action Recognition

Toby Perrett Alessandro Masullo Tilo Burghardt Majid Mirmehdi Dima Damen
<first>.<last>@bristol.ac.uk Department of Computer Science, University of Bristol, UK

Abstract

We propose a novel approach to few-shot action recognition, finding temporally-corresponding frame tuples between the query and videos in the support set. Distinct from previous few-shot works, we construct class prototypes using the CrossTransformer attention mechanism to observe relevant sub-sequences of all support videos, rather than using class averages or single best matches. Video representations are formed from ordered tuples of varying numbers of frames, which allows sub-sequences of actions at different speeds and temporal offsets to be compared.[1]

Our proposed Temporal-Relational CrossTransformers (TRX) achieve state-of-the-art results on few-shot splits of Kinetics, Something-Something V2 (SSv2), HMDB51 and UCF101. Importantly, our method outperforms prior work on SSv2 by a wide margin (12%) due to the its ability to model temporal relations. A detailed ablation showcases the importance of matching to multiple support set videos and learning higher-order relational CrossTransformers.

1. Introduction

Few-shot methods aim to learn new classes with only a handful of labelled examples. Success in few-shot approaches for image classification [11, 19, 8] and object recognition [26, 15] has triggered recent progress in few-shot video action recognition [31, 32, 3, 27, 4]. This is of particular interest for fine-grained actions where collecting enough labelled examples proves challenging [5, 12, 6].

Recent approaches that achieve state-of-the-art performance [3, 27, 4] acknowledge the additional challenges in few-shot video recognition, due to varying action lengths and temporal dependencies. However, these match the query video (*i.e.* the video to be recognised) to the single best video in the support set (*i.e.* the few labelled examples per class), e.g. [27], or to the average across all support set videos belonging to the same class [3, 4]. Inspired by part-based few-shot image classification [8], we consider that, within a few-shot regime, it is advantageous to compare sub-sequences of the query video to sub-sequences of

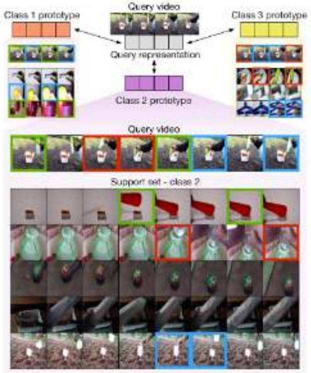
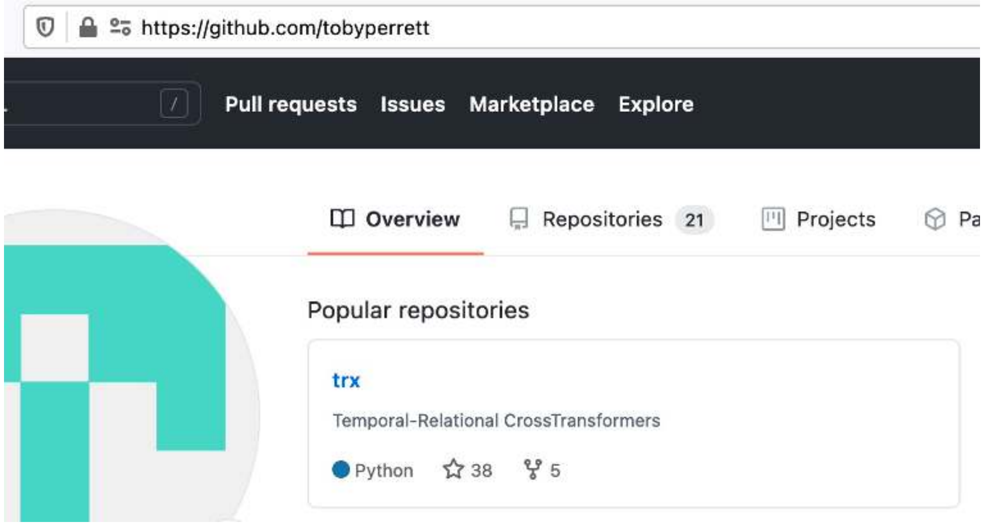


Figure 1: For a 3-way 5-shot example, pairs of temporally-ordered frames in the query (red, green, blue) are compared against all pairs in the support set (max attention with corresponding colour). Aggregated evidence is used to construct query-specific class prototypes. We show a correctly-recognised query using our method from SSv2 class "Failing to put something into something because it does not fit".

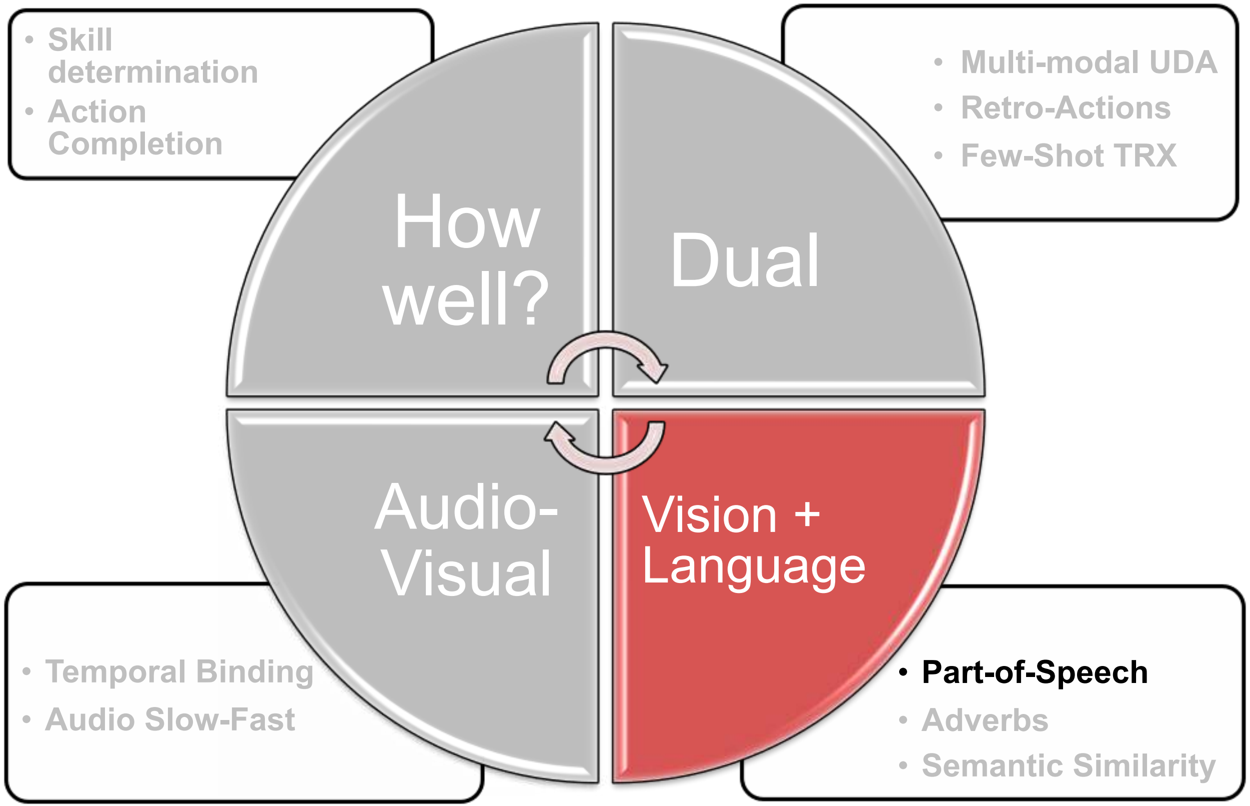
all support videos when constructing class prototypes. This better accumulates evidence, by matching sub-sequences at various temporal positions and shifts.

We propose a novel approach to few-shot action recognition, which we term Temporal-Relational CrossTransformers (TRX). A query-specific class prototype is constructed by using an attention mechanism to match each query sub-sequence against all sub-sequences in the support set, and aggregating this evidence. By performing the attention operation over temporally-ordered sub-sequences

[1] Code is available at https://github.com/tobyperrett/TRX



Fine-grained in Video?



What is a Cross-Modal Video Retrieval?

Video

put garlic down

Text

What is a Cross-Modal Video Retrieval?

Video-to-Text Retrieval Task

Q

Ranked Text – Gallery (or Retrieval Set)

put garlic down

Text-to-Video Retrieval Task

Ranked Video – Gallery (or Retrieval Set)

Q put garlic down

Fine-Grained Action Retrieval

with: Michael Wray
Gabriela Csurka
Diane Larlus

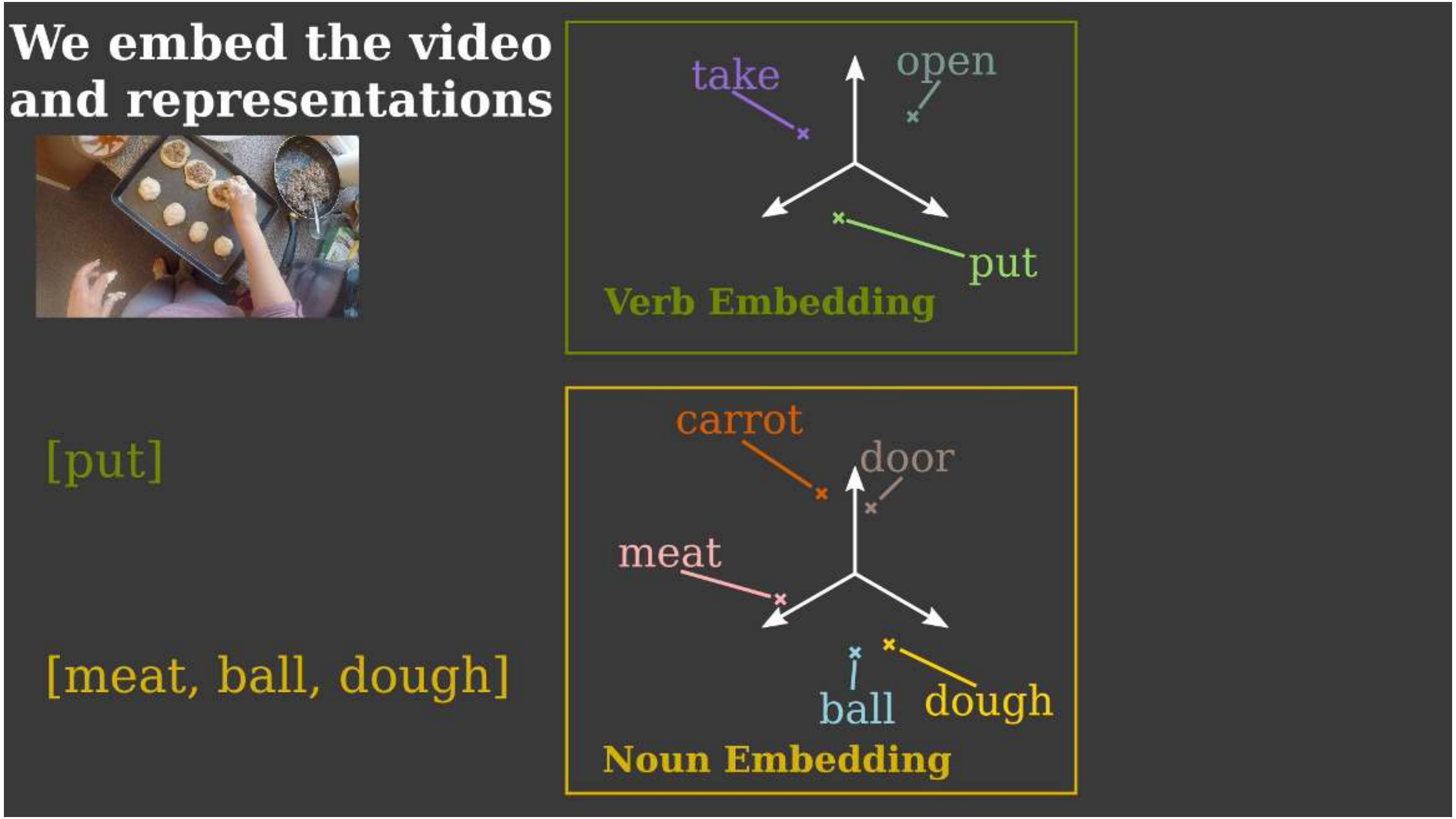
In this work we focus on
Fine-Grained Action Retrieval

I put meat on a ball of dough ⟷

M Wray, D Larlus, G Csurka, D Damen (2019). Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings. ICCV

Fine-Grained Action Retrieval

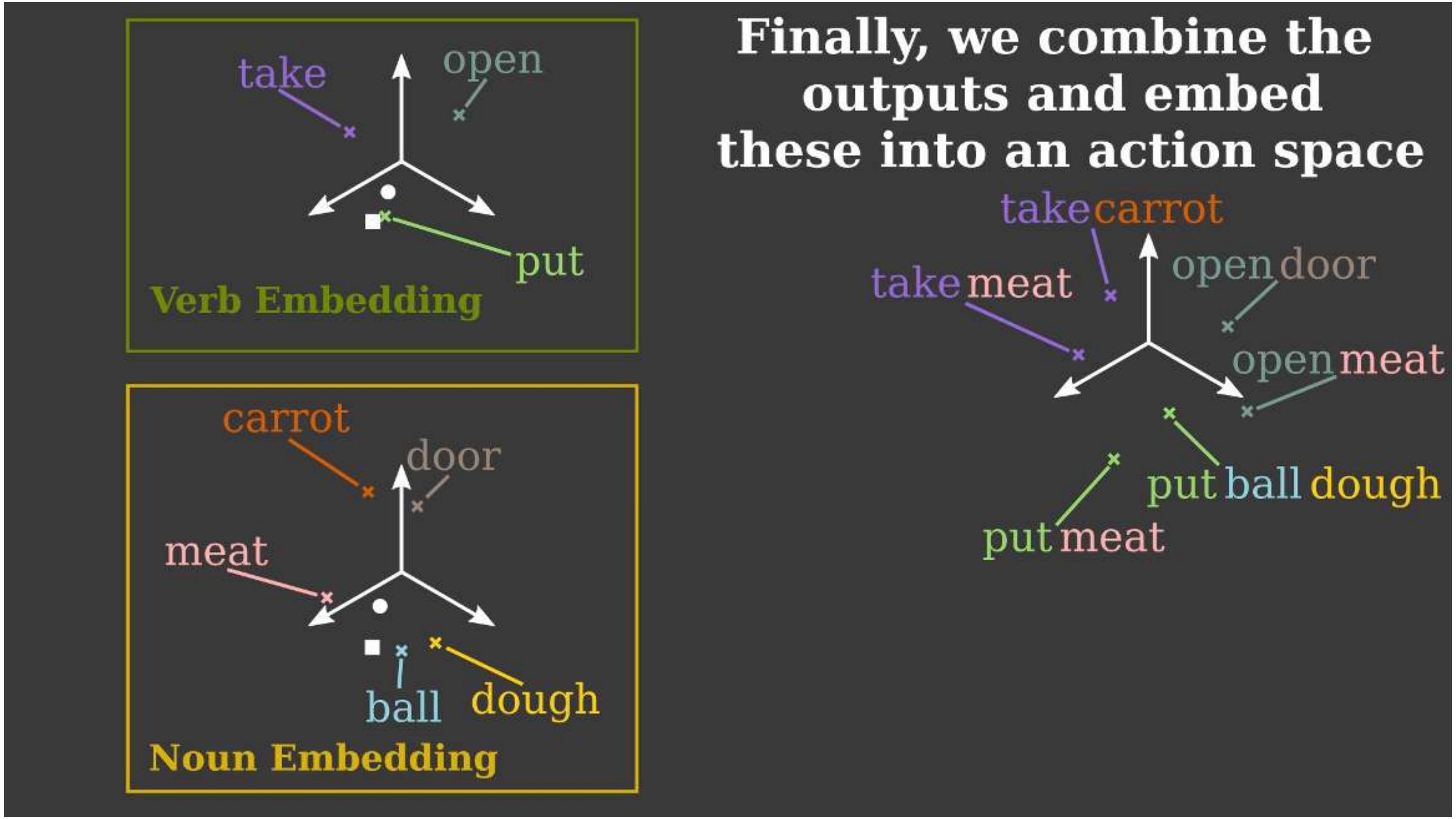
with: Michael Wray
Gabriela Csurka
Diane Larlus



M Wray, D Larlus, G Csurka, D Damen (2019). Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings. ICCV

Fine-Grained Action Retrieval

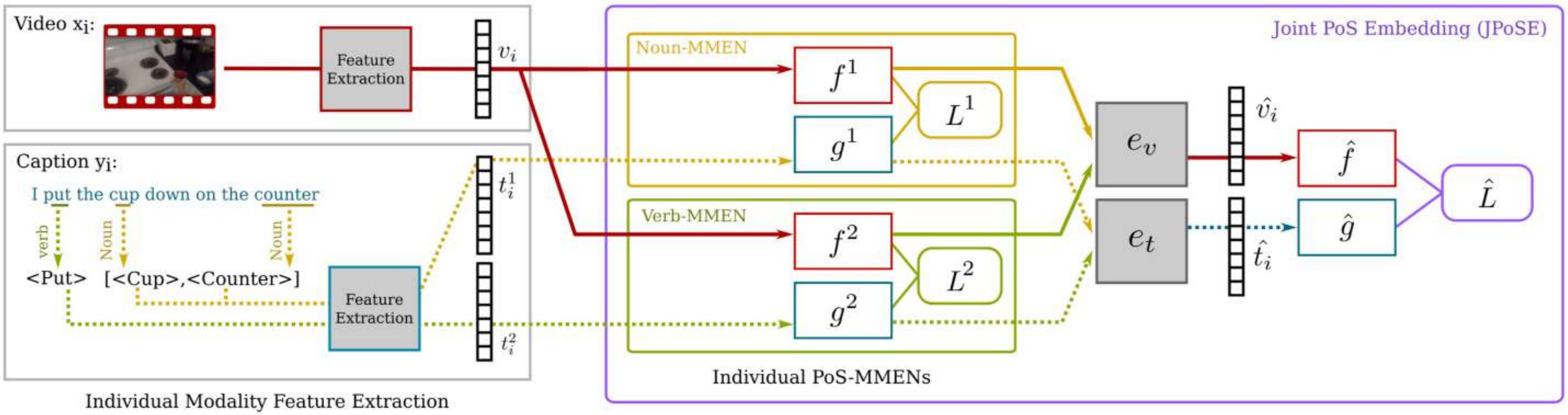
with: Michael Wray
Gabriela Csurka
Diane Larlus



M Wray, D Larlus, G Csurka, D Damen (2019). Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings. ICCV

Fine-Grained Action Retrieval

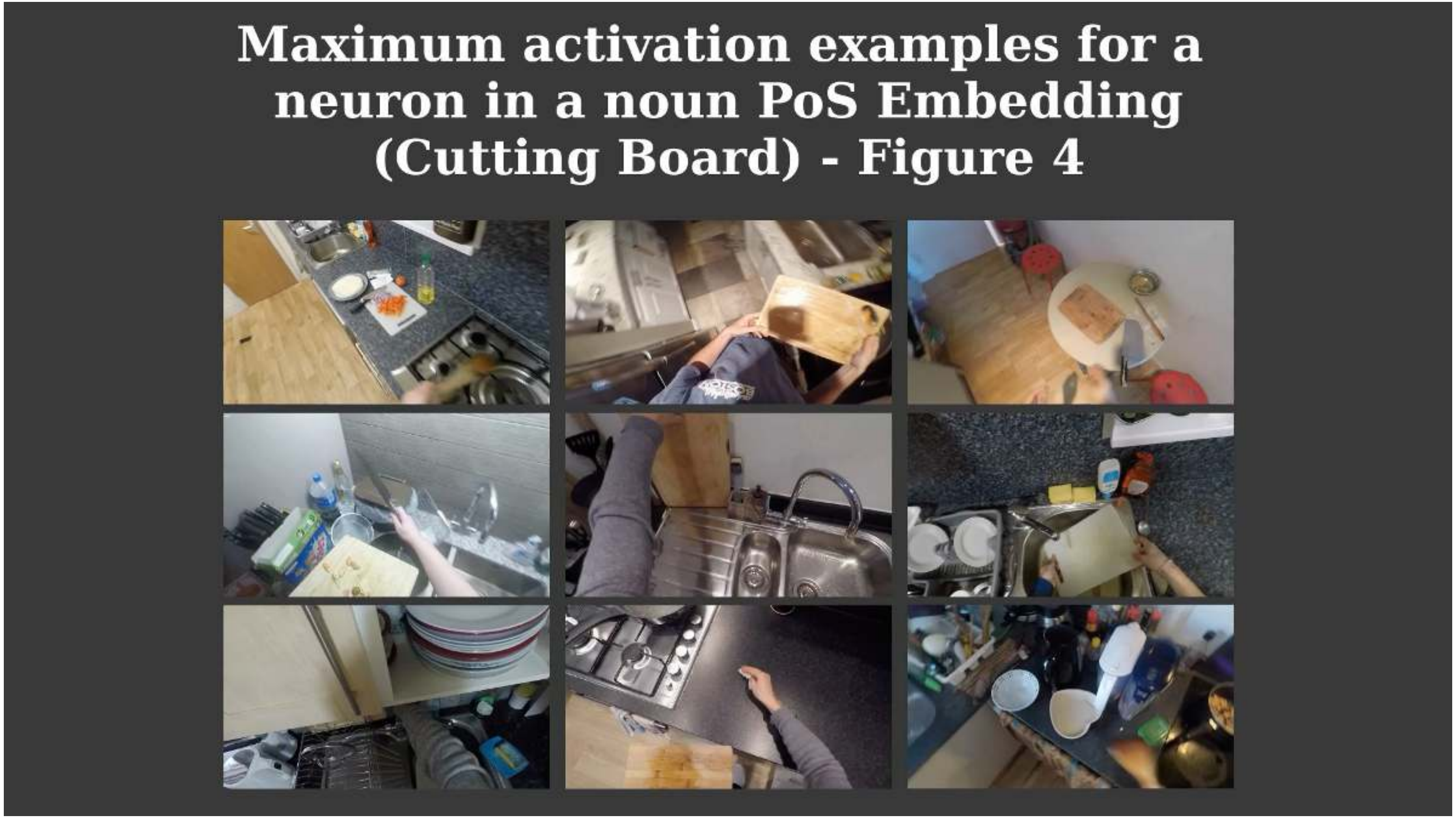
with: Michael Wray
Gabriela Csurka
Diane Larlus



M Wray, D Larlus, G Csurka, D Damen (2019). Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings. ICCV

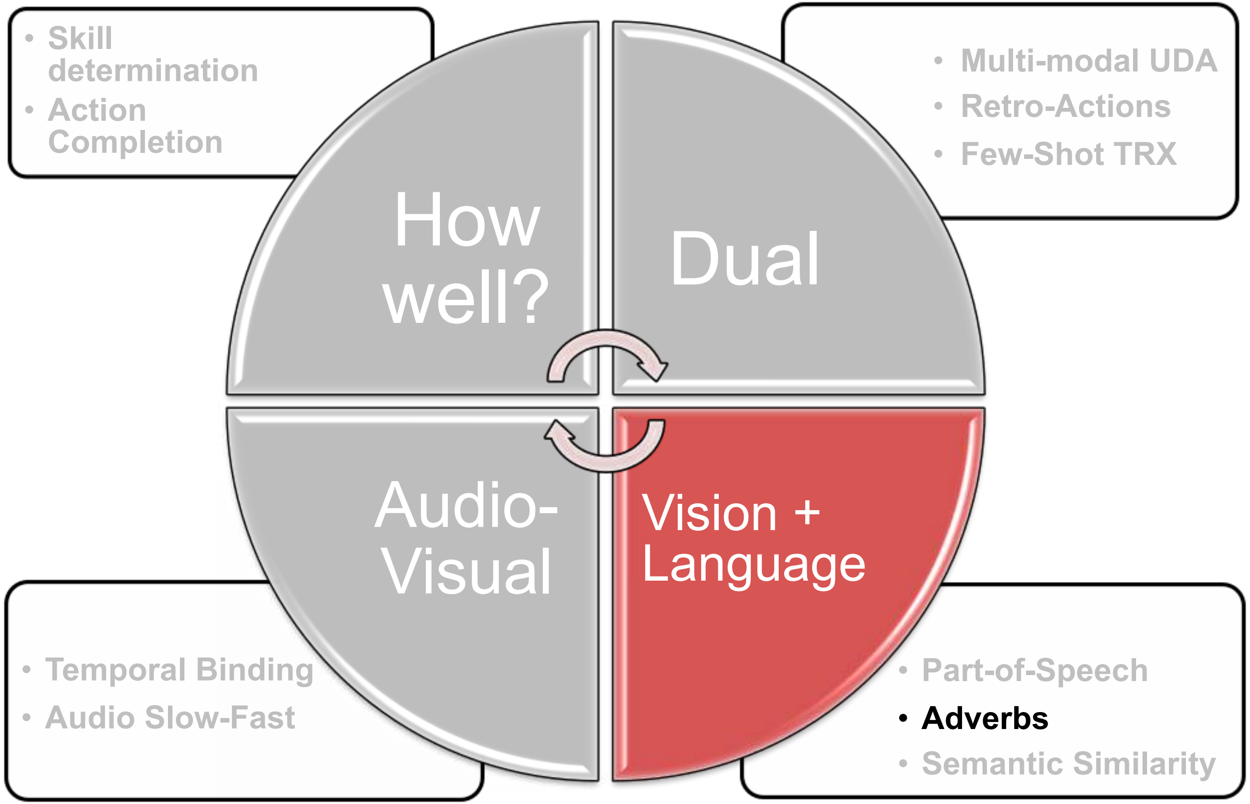
Fine-Grained Action Retrieval

with: Michael Wray
Gabriela Csurka
Diane Larlus



M Wray, D Larlus, G Csurka, D Damen (2019). Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings. ICCV

Fine-grained in Video?



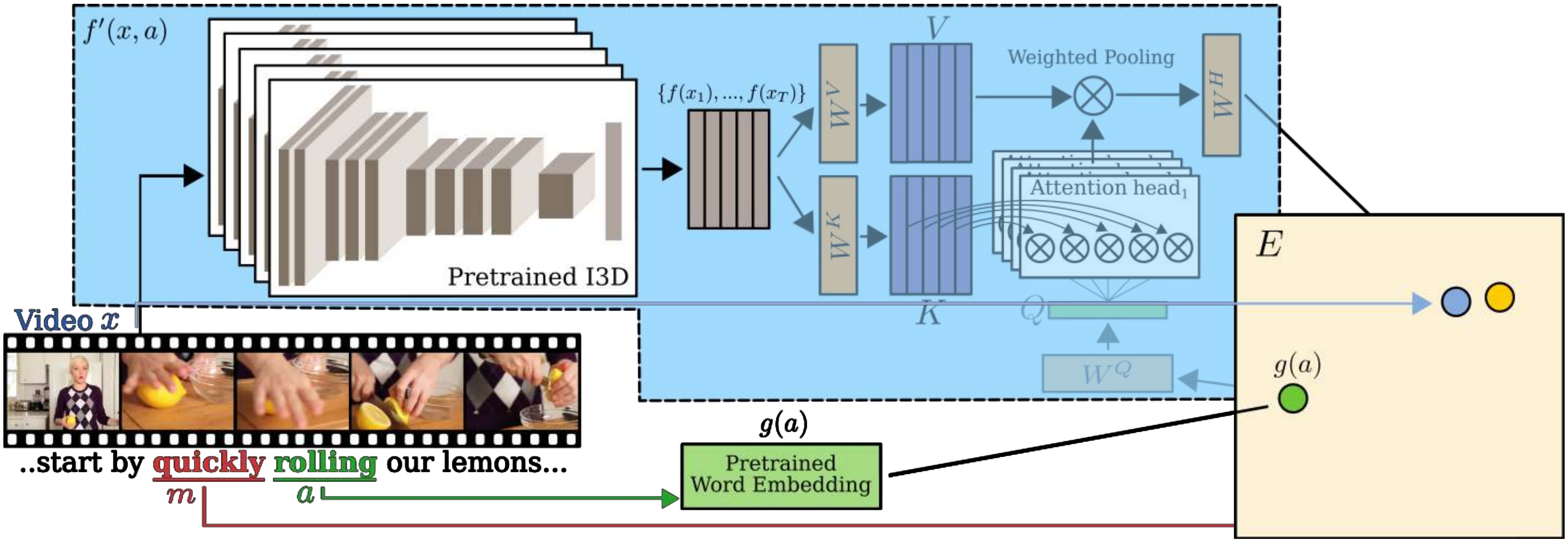
Action Modifiers: Learning from Adverbs

with: Hazel Doughty
Ivan Laptev
Walterio Mayol-Cuevas

H Doughty, I Laptev, W Mayol-Cuevas, D Damen (2020). Action Modifiers: Learning from Adverbs in Instructional Videos. Computer Vision and Pattern Recognition (CVPR)

Action Modifiers: Learning from Adverbs

with: Hazel Doughty
Ivan Laptev
Walterio Mayol-Cuevas



H Doughty, I Laptev, W Mayol-Cuevas, D Damen (2020). Action Modifiers: Learning from Adverbs in Instructional Videos. Computer Vision and Pattern Recognition (CVPR)

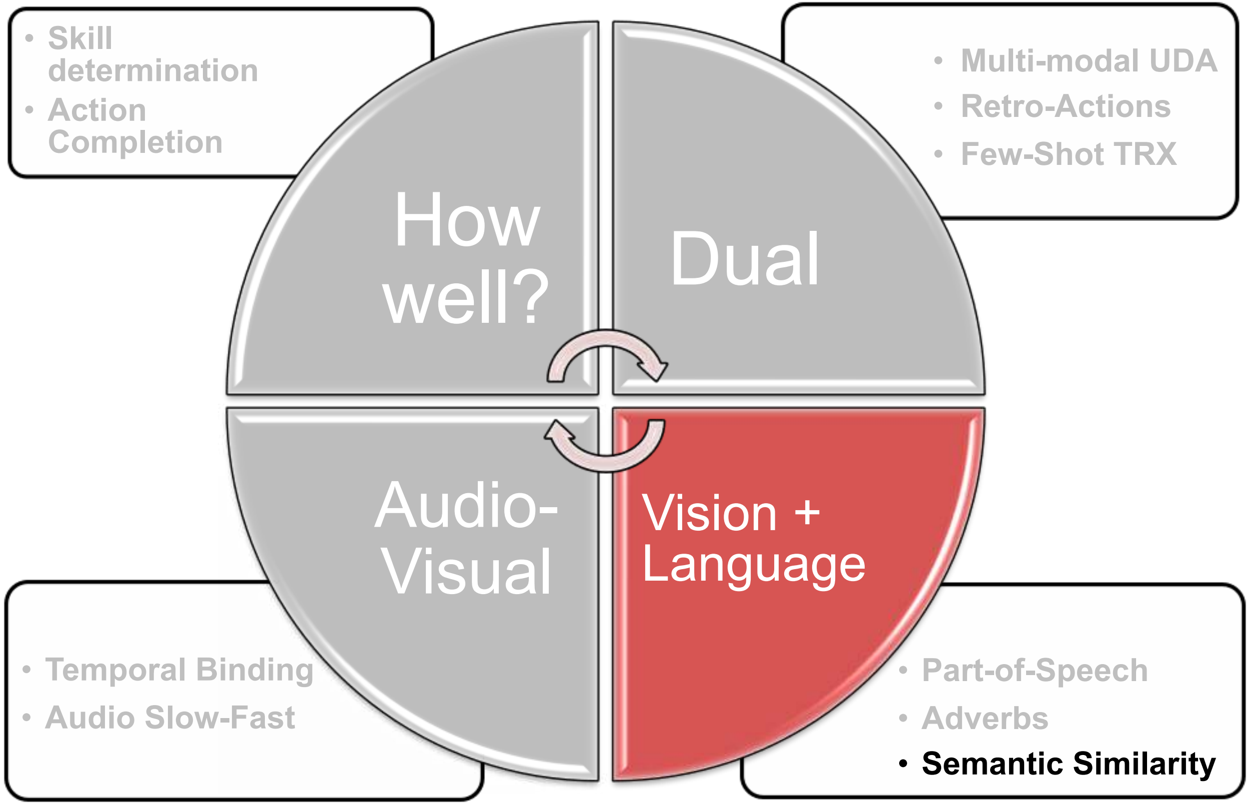
Action Modifiers: Learning from Adverbs

with: Hazel Doughty
Ivan Laptev
Walterio Mayol-Cuevas

... we're going to **mix** these up real **quick**...

H Doughty, I Laptev, W Mayol-Cuevas, D Damen (2020). Action Modifiers: Learning from Adverbs in Instructional Videos. Computer Vision and Pattern Recognition (CVPR)

Fine-grained in Video?



University of BRISTOL

On Semantic Similarity in Video Retrieval

with: Michael Wray
Hazel Doughty

- Which of these captions correspond to the following video?

A band is performing for the crowd

A man is peeling fruit carefully and neatly.

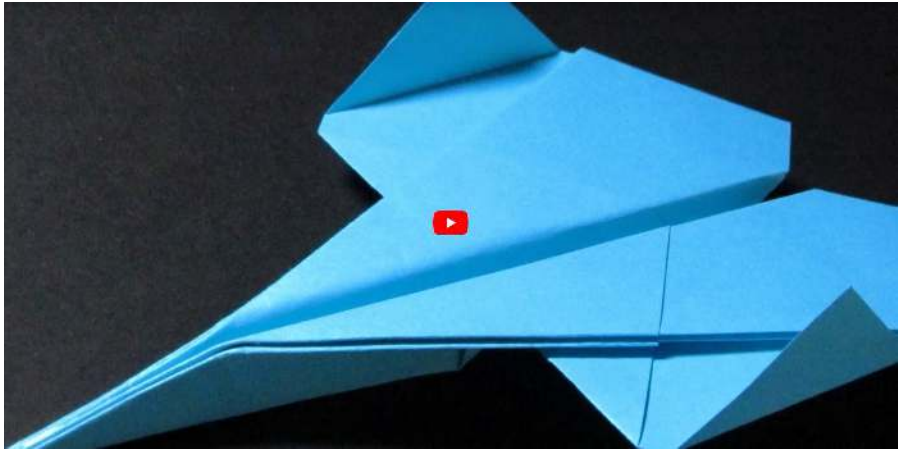
A girl is sitting in a chair

Add prawns to the pan and mix

On Semantic Similarity in Video Retrieval

with: Michael Wray
Hazel Doughty

- Which of these captions correspond to the following video?



A man performing an Origami tutorial

A demonstration in Origami

A guy explains the steps of folding paper

A man folding a piece of paper into a paper airplane

M Wray et al (2021). On Semantic Similarity in Video Retrieval. Computer Vision and Pattern Recognition (CVPR)

On Semantic Similarity in Video Retrieval

with: Michael Wray
Hazel Doughty

- Previous methods have made the following assumption
  - *"There exists only one corresponding caption for a given video and vice versa"*

Peel and chop the potatoes
Peel and cut up the potato
Peel the potatoes and cut them
Peel and cut the potatoes into chunks
Peel the potatoes and cut them into halves

YouCook2

Put fork and spoon in drying rack
Put spoons in drying rack
Put spoon in drying rack
Put bowl in drying rack
Put plate in drying rack

EPIC-KITCHENS

MSR-VTT

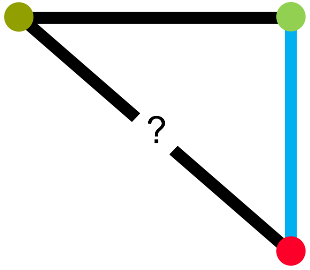
A band is performing for the crowd
A band is performing on a brightly lit stage
A band is playing a show
A band and singers perform
3 guys singing and playing instruments on a stage

M Wray et al (2021). On Semantic Similarity in Video Retrieval. Computer Vision and Pattern Recognition (CVPR)

On Semantic Similarity in Video Retrieval

with: Michael Wray
Hazel Doughty

- Want to relate two items semantically.



- Assume that a caption sufficiently describes a video.

- Define a proxy function that relates captions

$$S(x_i, y_j) = S'(y_i, y_j)$$

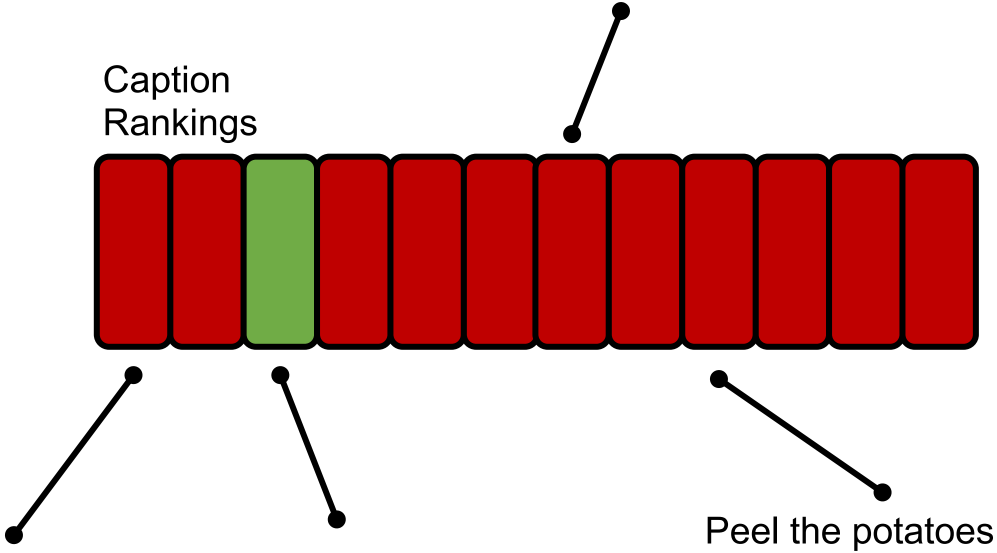
M Wray et al (2021). On Semantic Similarity in Video Retrieval. Computer Vision and Pattern Recognition (CVPR)

On Semantic Similarity in Video Retrieval

with: Michael Wray
Hazel Doughty

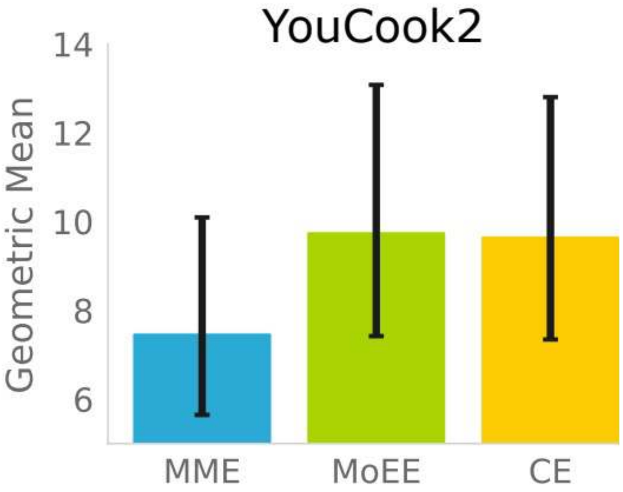
- When evaluating with a single caption, the correct caption can be arbitrary.

Query Video



On Semantic Similarity in Video Retrieval

with: Michael Wray
Hazel Doughty



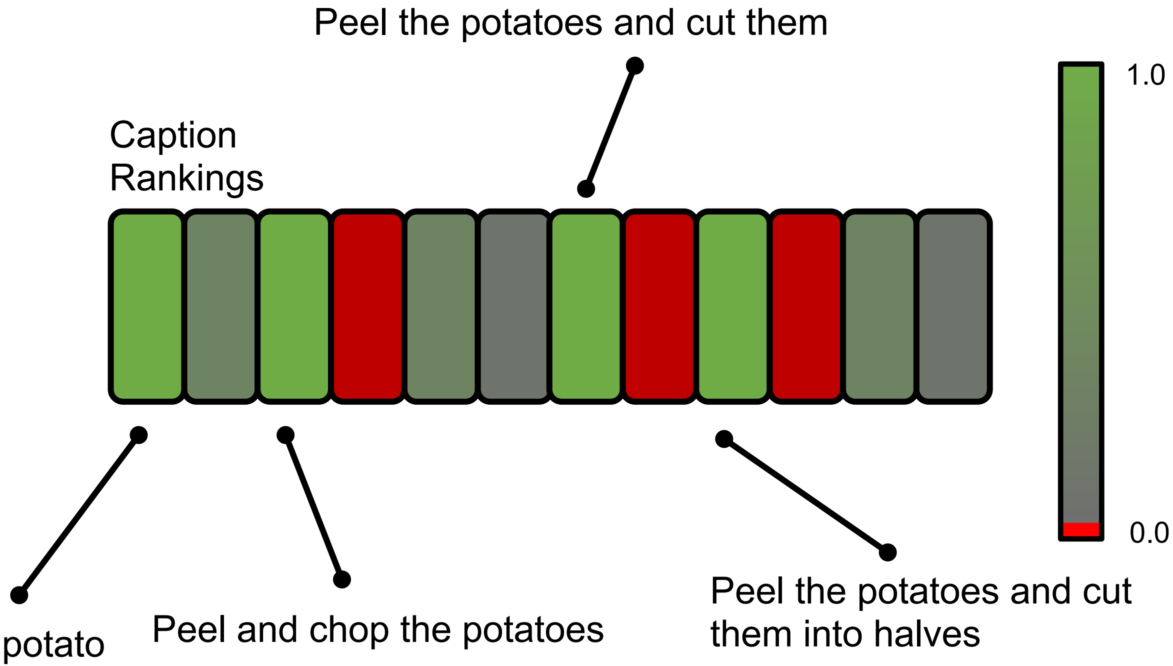
MoEE: Antoine Miech, Ivan Laptev, and Josef Sivic. Learning a text-video embedding from incomplete and heterogeneous data. CoRR, abs/1804.02516, 2018
CE: Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. Use what you have: Video retrieval using representations from collaborative experts. In BMVC, 2019
JPoSE: Michael Wray, Diane Larlus, Gabriela Csurka, and Dima Damen. Fine-grained action retrieval through multiple parts of-speech embeddings. In ICCV, 2019

On Semantic Similarity in Video Retrieval

with: Michael Wray
Hazel Doughty

- We instead, propose to use normalised Discounted Cumulative Gain to evaluate multiple items with differing relevance.

Query Video

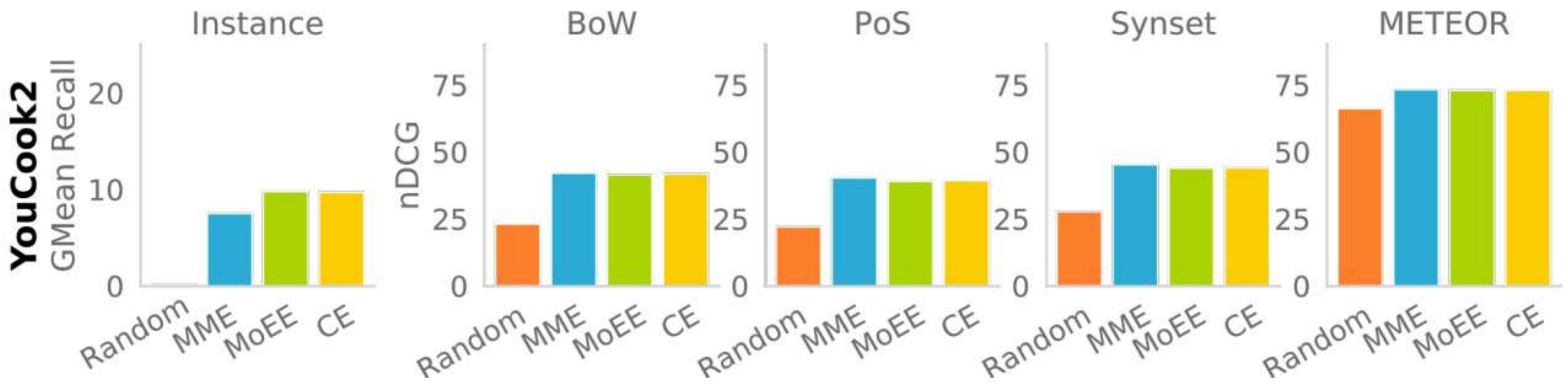


M Wray et al (2021). On Semantic Similarity in Video Retrieval. Computer Vision and Pattern Recognition (CVPR)

On Semantic Similarity in Video Retrieval

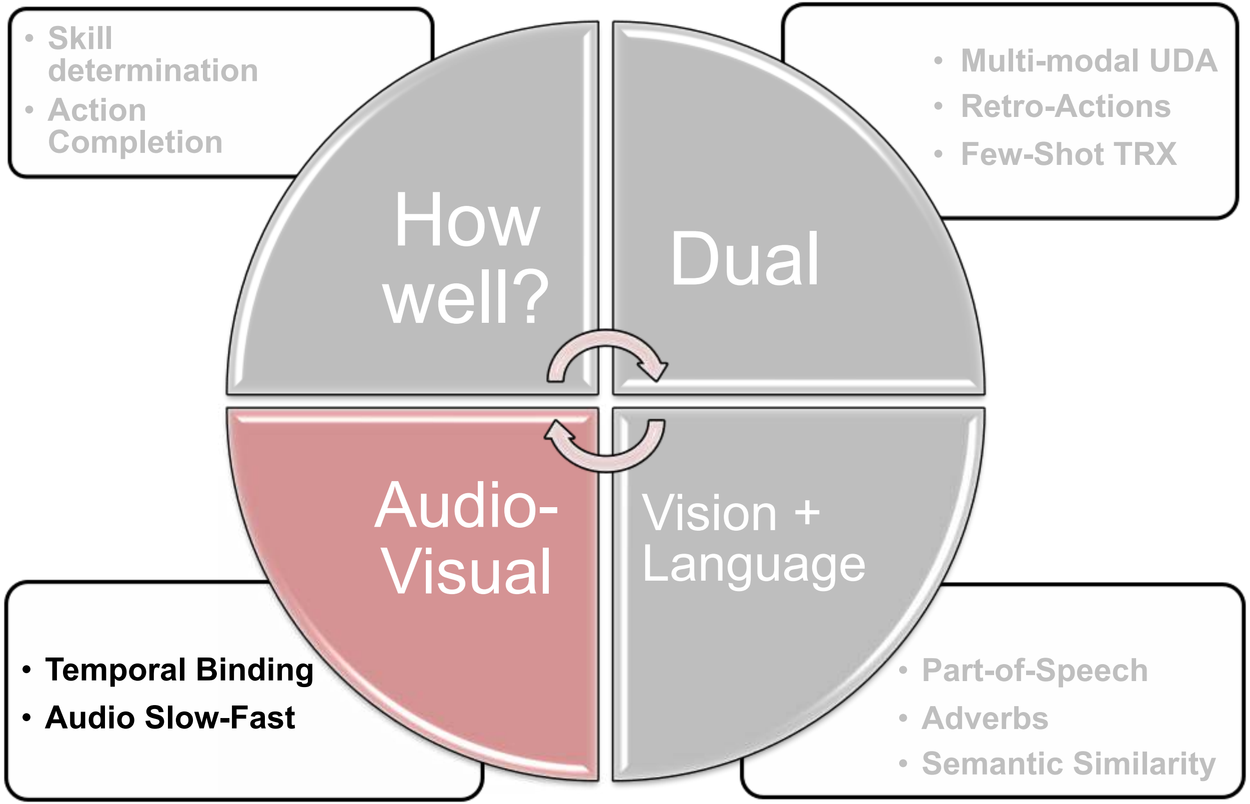
with: Michael Wray
Hazel Doughty

- Whilst models outperform the MLP baseline (MME) for Instance Video Retrieval, this isn't the case when Semantic Similarity is used.



M Wray et al (2021). On Semantic Similarity in Video Retrieval. Computer Vision and Pattern Recognition (CVPR)

VU - An Egocentric Perspective



Why do we need audio?

- The magic of audio-visual understanding…

Why do we need audio?

- The magic of audio-visual understanding…

Why do we need audio?

- The magic of audio-visual understanding…

Why do we need audio?

- The magic of audio-visual understanding…

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

Harmonic vs Percussive

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

VGG-Sound

Harmonic Sounds

Percussive Sounds

Auditory Slow-Fast

Outstanding Paper Award – ICASSP 2021

University of BRISTOL

Harmonic vs Percussive

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

- Strong evidence in neuroscience about ventral-dorsal streams in human auditory system
  - Some works suggest that ventral has high spectral resolution, while dorsal has high temporal resolution and operates at a higher sampling rate

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

Harmonic vs Percussive

SlowFast Networks for Video Recognition

Christoph Feichtenhofer   Haoqi Fan   Jitendra Malik   Kaiming He

Facebook AI Research (FAIR)

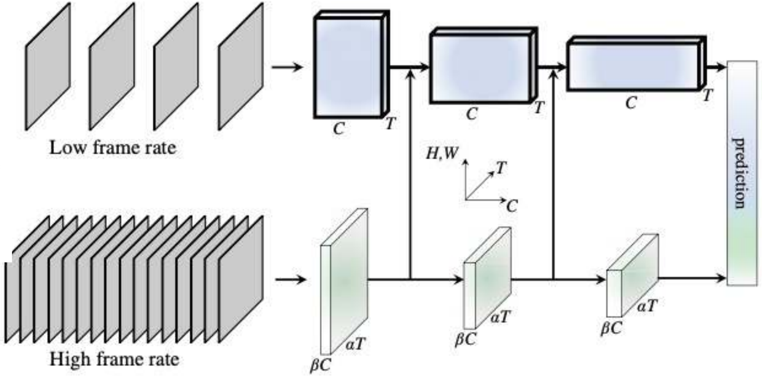
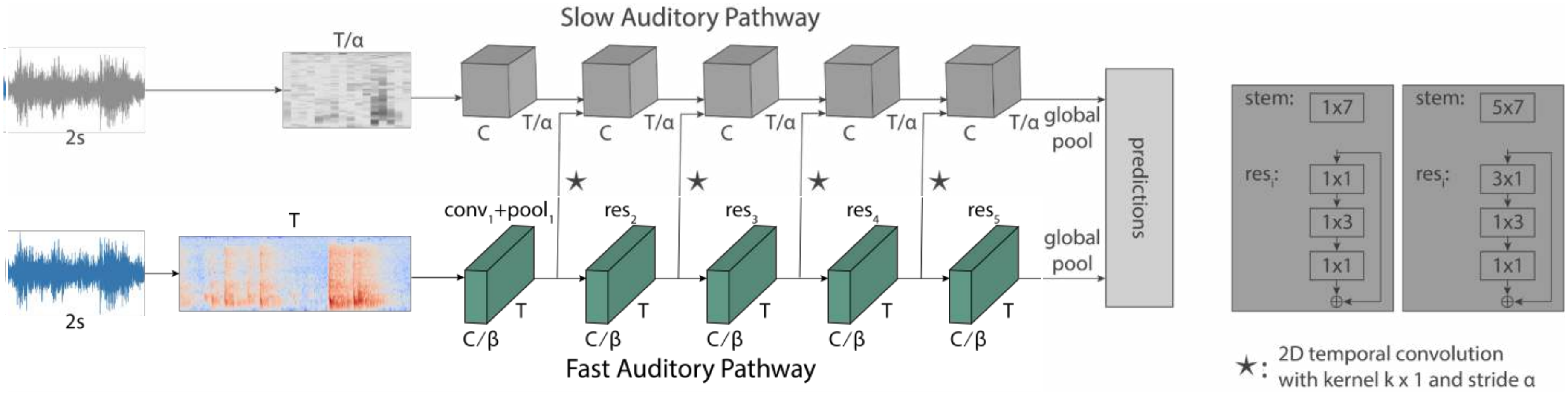


Figure 1. **A SlowFast network** has a low frame rate, low temporal resolution *Slow* pathway and a high frame rate, $\alpha\times$ higher temporal resolution *Fast* pathway. The Fast pathway is lightweight by using a fraction (β, *e.g.*, 1/8) of channels. Lateral connections fuse them.

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

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

VGG-Sound

Model	Top-1	Top-5
Chen et al. [2]	51.00	76.40
McDonnell & Gao [3]	39.74	71.65
Slow	45.20	72.53
Fast	41.44	70.68
Slow-Fast (Proposed)	**52.46**	**78.12**

EPIC-KITCHENS

		Top-1 Accuracy (%)			
Split	Model	Verb	Noun	Action	# Param.
Test	Damen et al. [1]	42.12	21.51	14.76	10.67M
Test	Slow-Fast (Proposed)	**46.47**	**22.77**	**15.44**	26.88M

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

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

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

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

111.12

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

Audio Slow-Fast

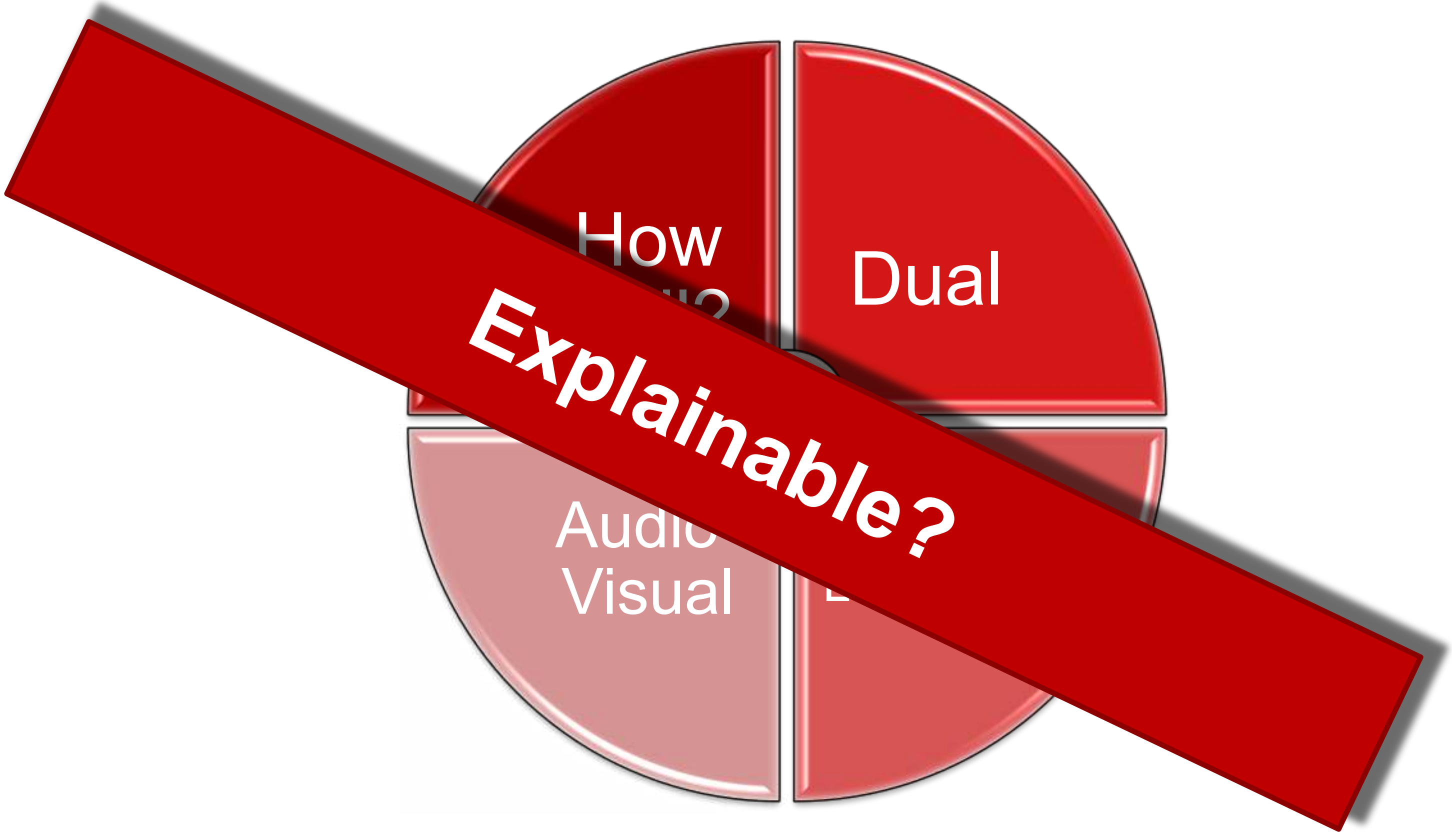
with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

- Project webpage: https://ekazakos.github.io/auditoryslowfast/

- Code & models: https://github.com/ekazakos/auditory-slow-fast

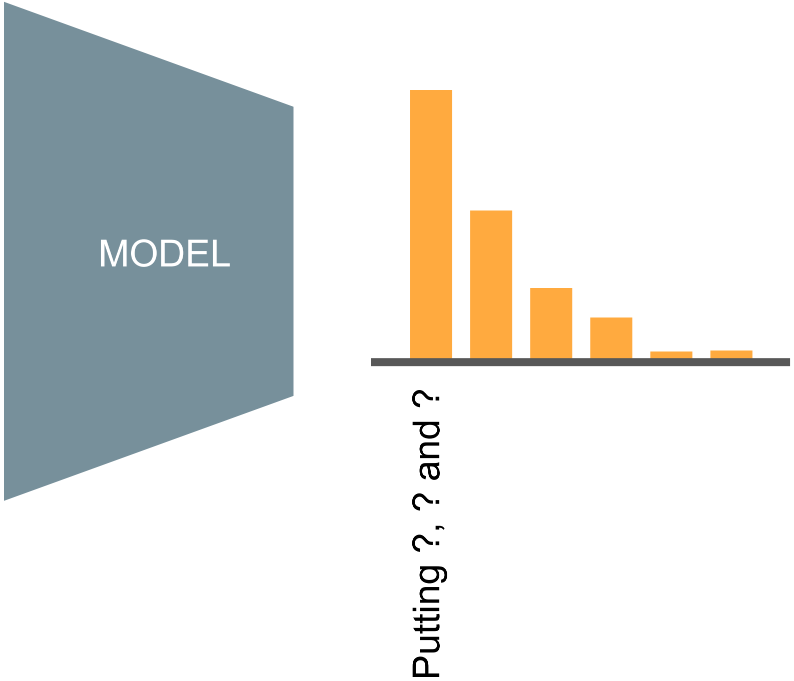
E Kazakos, A Nagrani, A Zisserman, D Damen (2021). Slow-Fast Auditory Streams For Audio Recognition. ICASSP

VU - An Egocentric Perspective



Frame Attributions in Video Models

with: Will Price

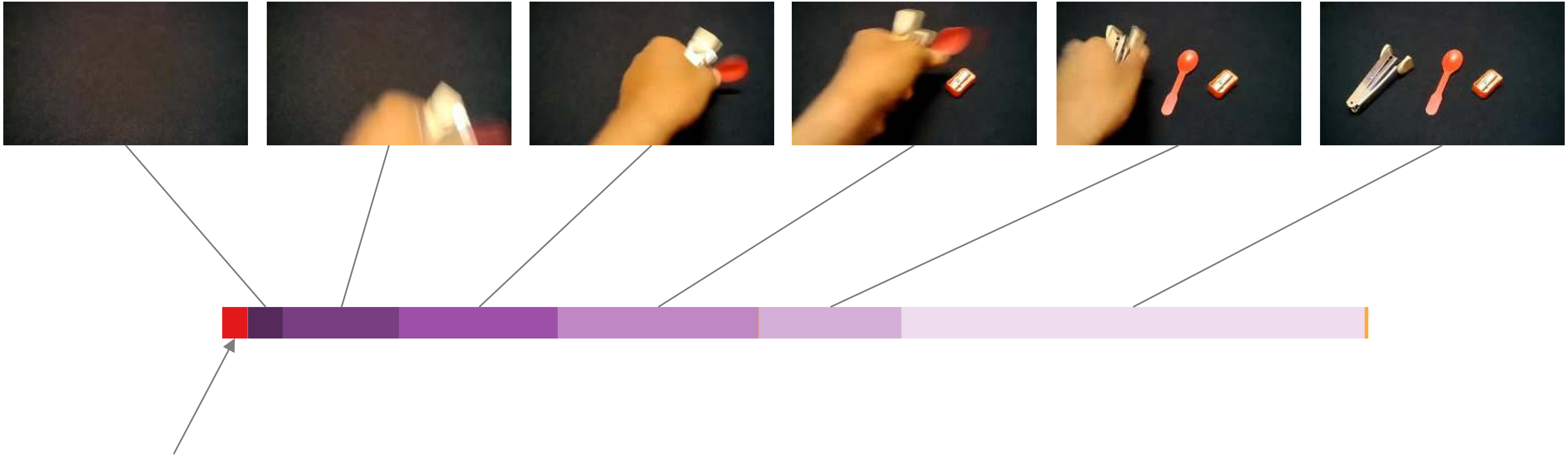


University of BRISTOL

W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. Asian Conference on Computer Vision (ACCV)

Frame Attributions in Video Models

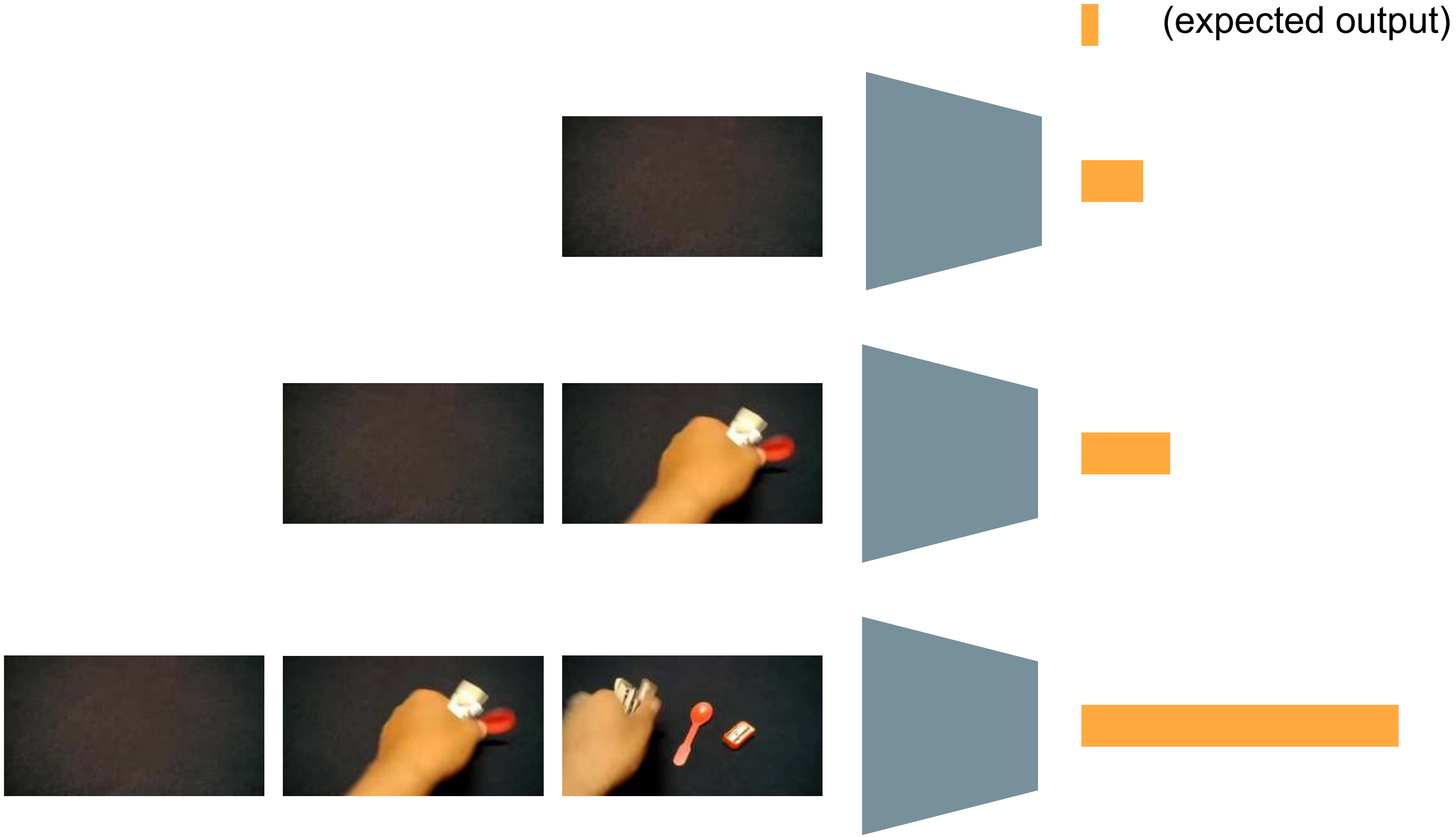
with: Will Price



Expected output
(Prior probability for
classification model)

Frame Attributions in Video Models

with: Will Price

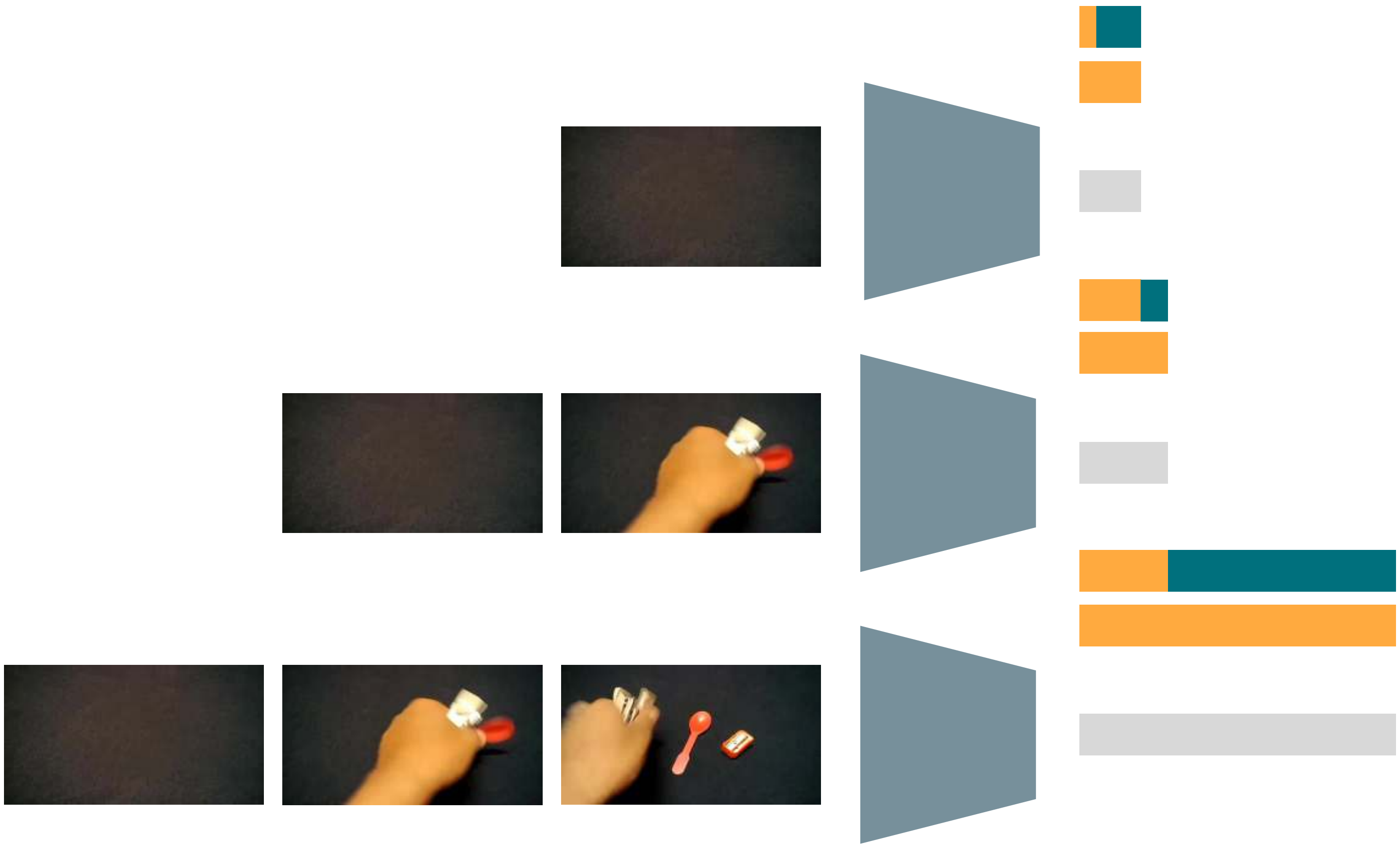


University of BRISTOL

W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. Asian Conference on Computer Vision (ACCV)

Frame Attributions in Video Models

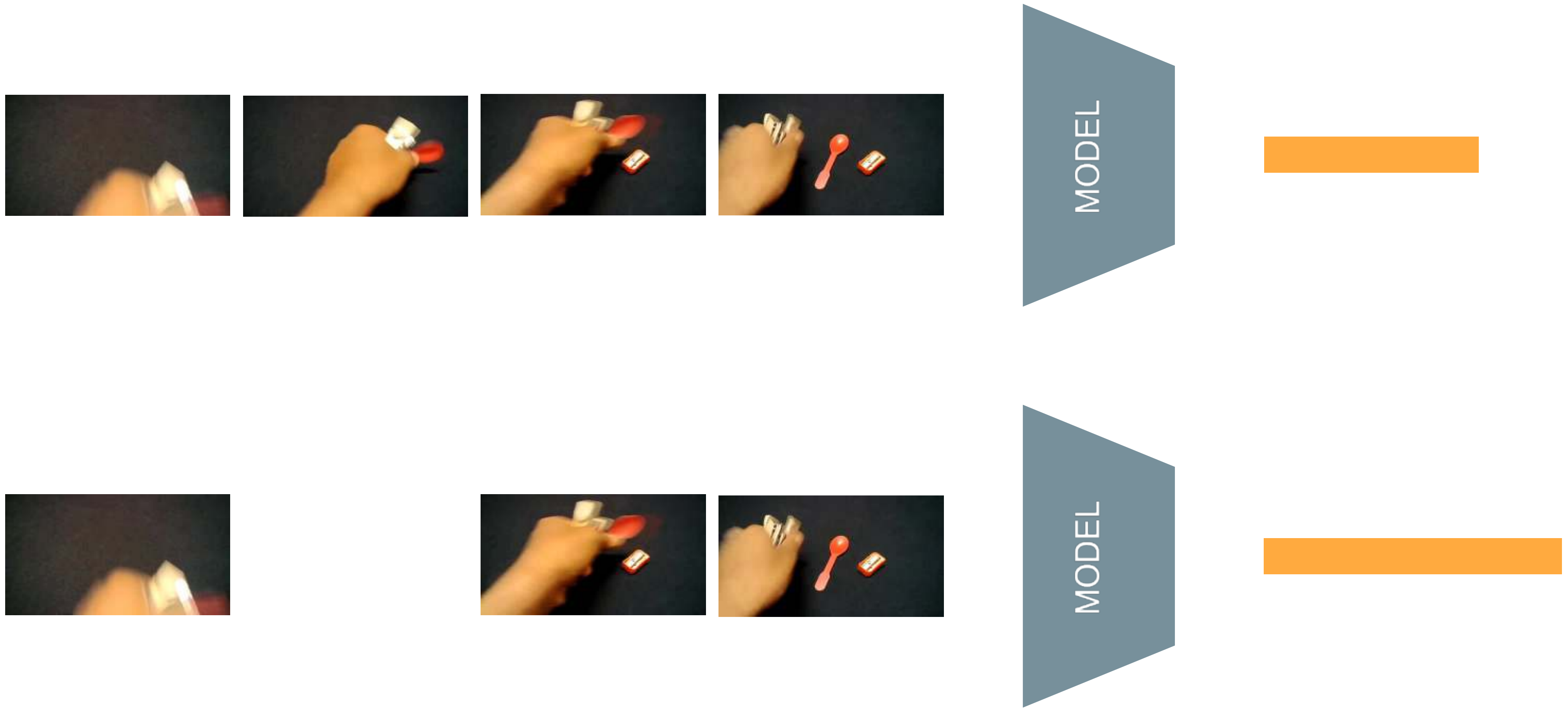
with: Will Price



W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. Asian Conference on Computer Vision (ACCV)

Frame Attributions in Video Models

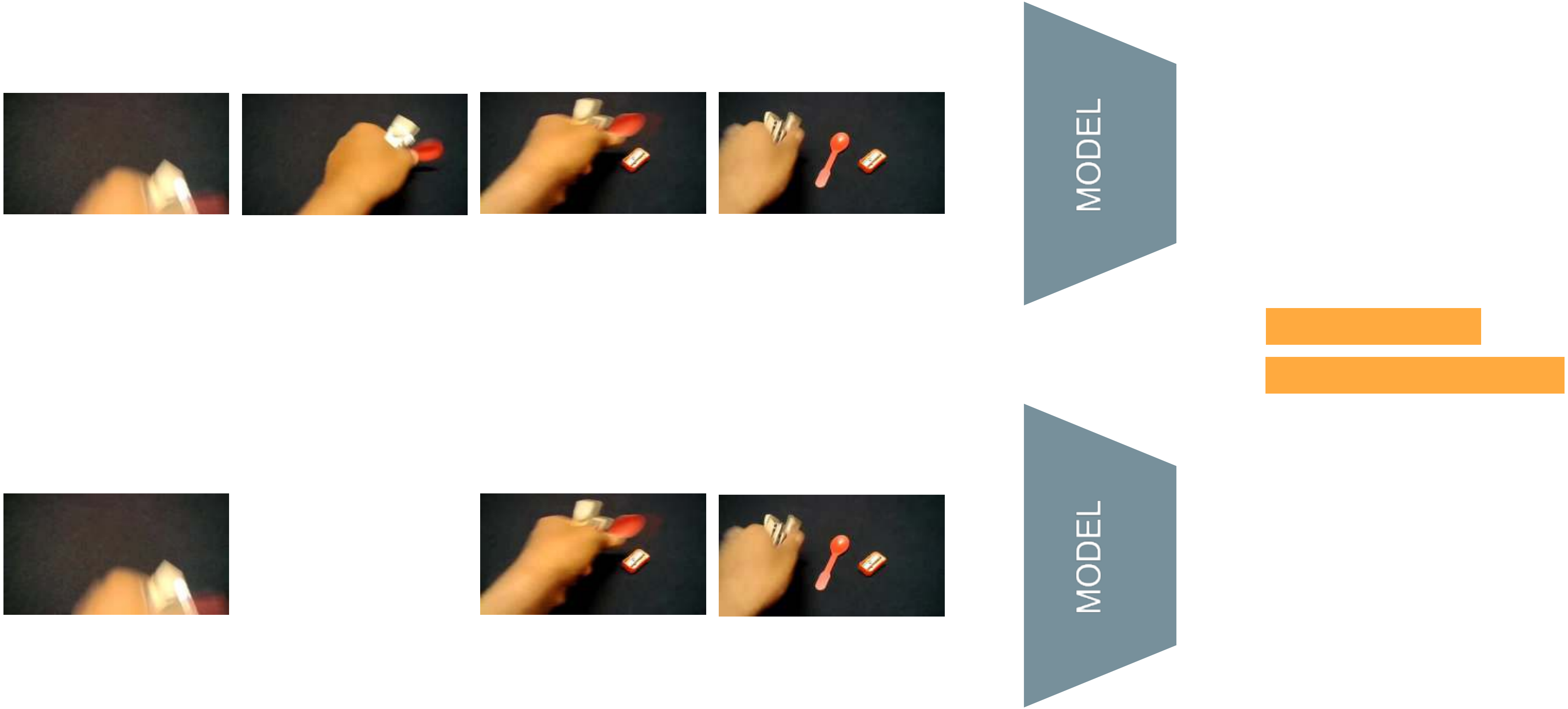
with: Will Price



W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. Asian Conference on Computer Vision (ACCV)

Frame Attributions in Video Models

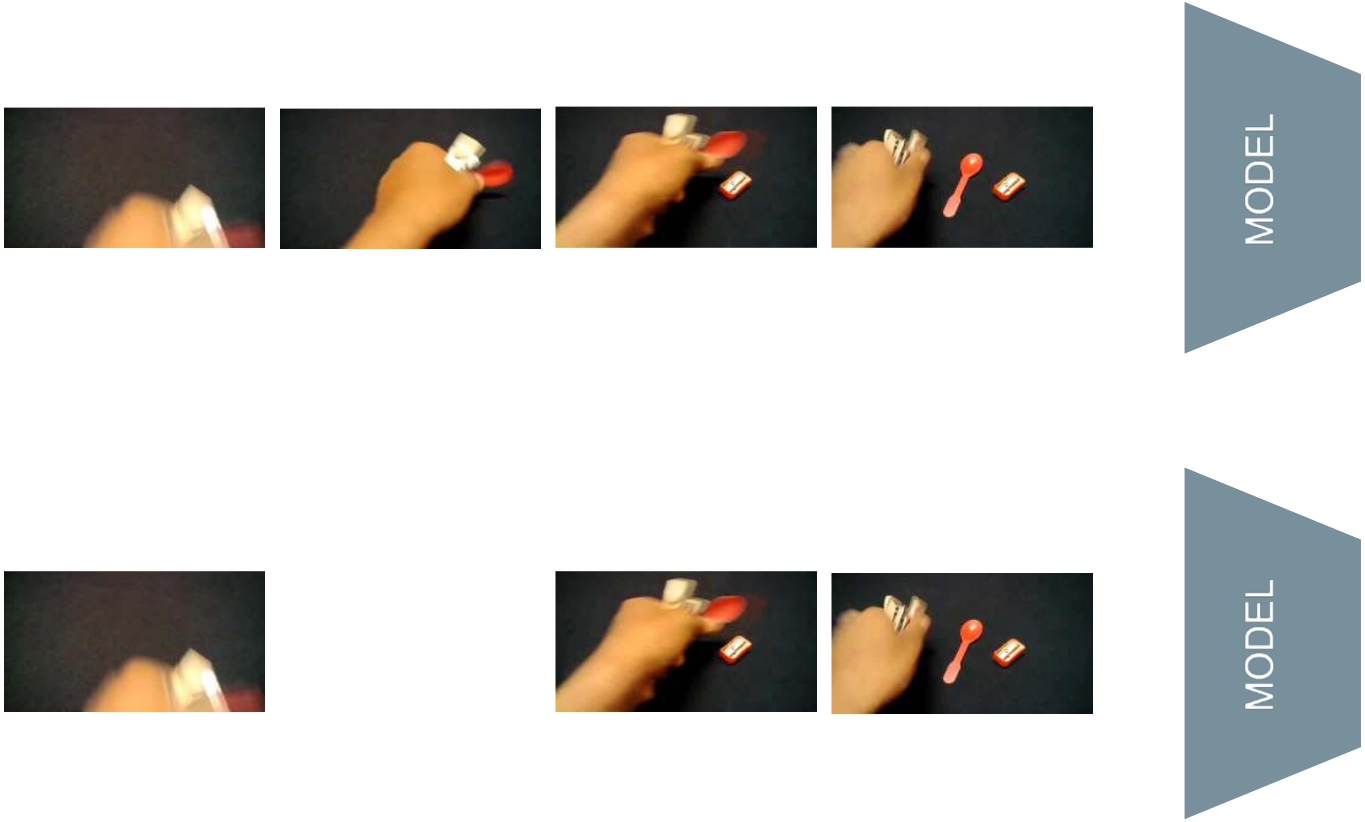
with: Will Price



W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. Asian Conference on Computer Vision (ACCV)

Frame Attributions in Video Models

with: Will Price

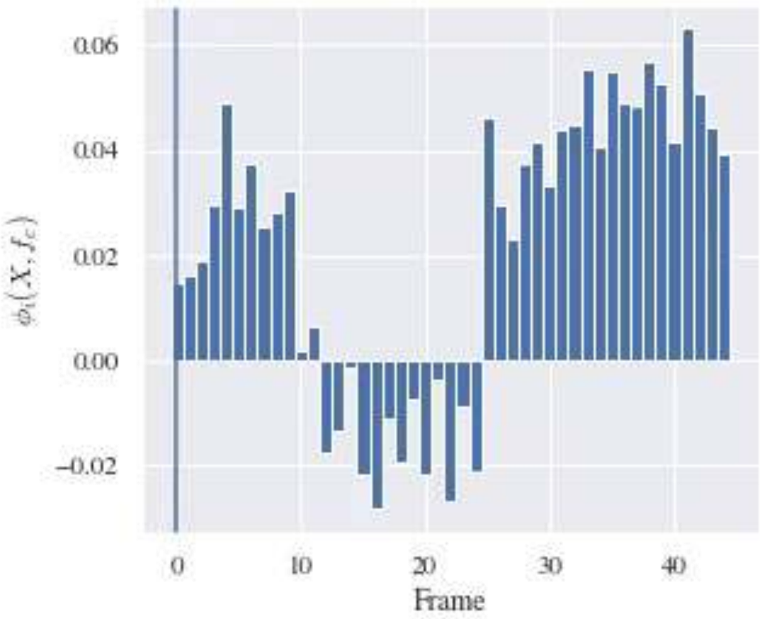


$\Delta_3(\{1,2,4,5\}) = -.2$

W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. Asian Conference on Computer Vision (ACCV)

Frame Attributions in Video Models

with: Will Price



W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. Asian Conference on Computer Vision (ACCV)

Dashboard

Frame Attributions in Video Models

with: Will Price

ESV Dashboard - Something Something v2 - Multiscale TRN

Class:

Pouring something into something

Example:

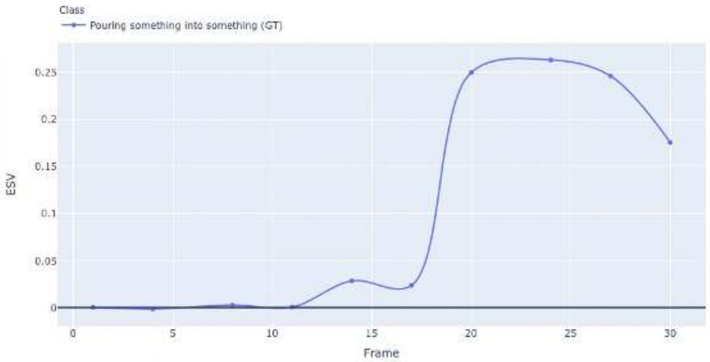
0 1 2 3 4 5 6 7 8 9

Frames fed to model:

1 2 3 4 5 6 7 8 9 10

Model Predictions

ESV Values



Hovered Frame:

Orignal Video:

W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. Asian Conference on Computer Vision (ACCV)

with: Will Price
Tom Stark

ESVs Dashboard for Epic

Select a verb: open

Select a noun: drawer

Select a video: P01_103_84

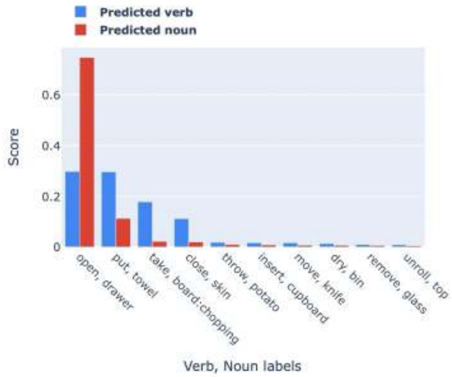
Select number of frames: 1 2 3 4 5 6 7 8

Original Video:

Selected frame: 2

Selected Verb: 3, Selected Noun: 8, Video P01_103_84

Model Predictions



ESV Predictions

W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. Asian Conference on Computer Vision (ACCV)

Frame Attributions in Video Models

with: Will Price
Tom Stark

ESVs Dashboard for Epic

Select a verb: cut
Select a noun: tomato
Select a video: P01_17_126

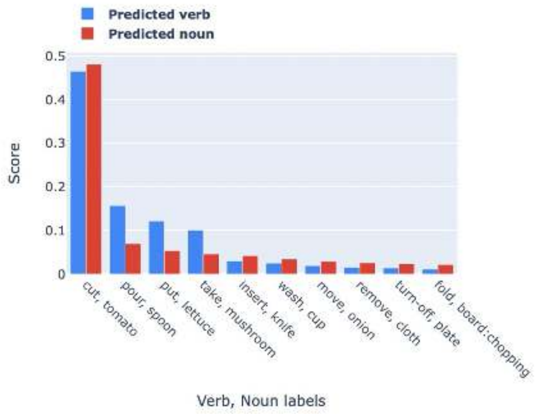
Select number of frames: 1 2 3 4 5 6 7 8

Original Video:

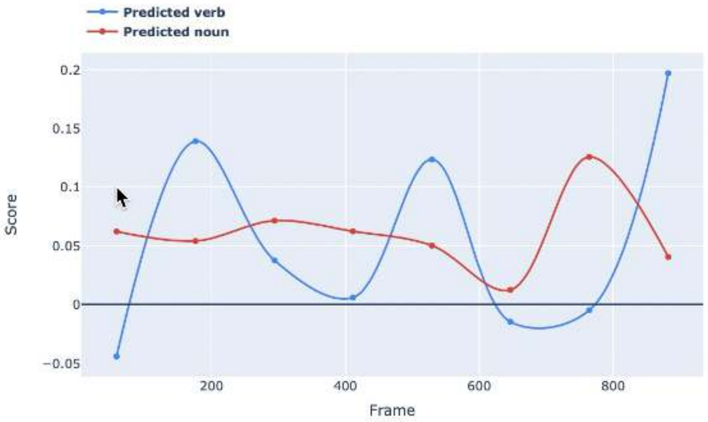
Selected frame: 529

Selected Verb: 7, Selected Noun: 43, Video P01_17_126

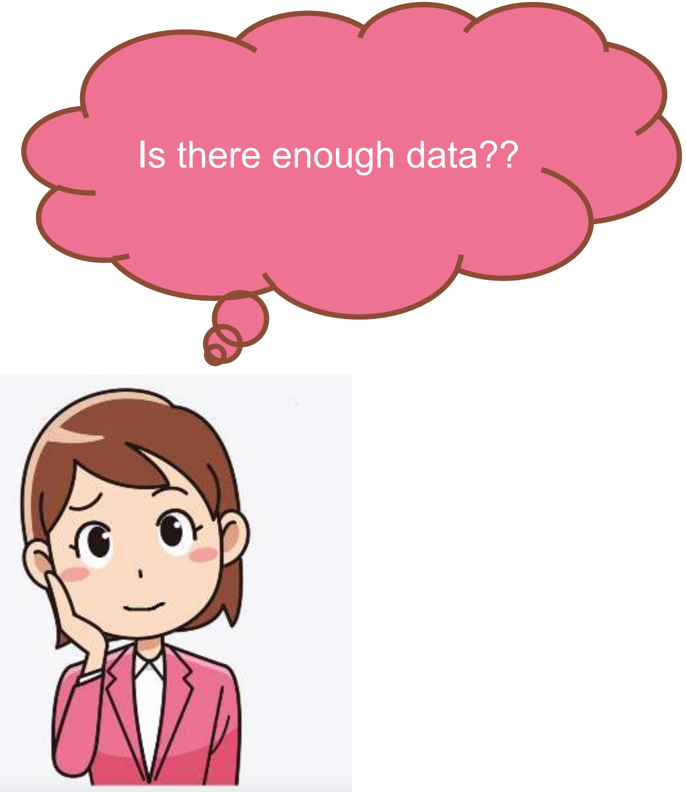
Model Predictions



ESV Predictions



W Price, D Damen (2020). Play Fair: Frame Attribution in Video Models. Asian Conference on Computer Vision (ACCV)

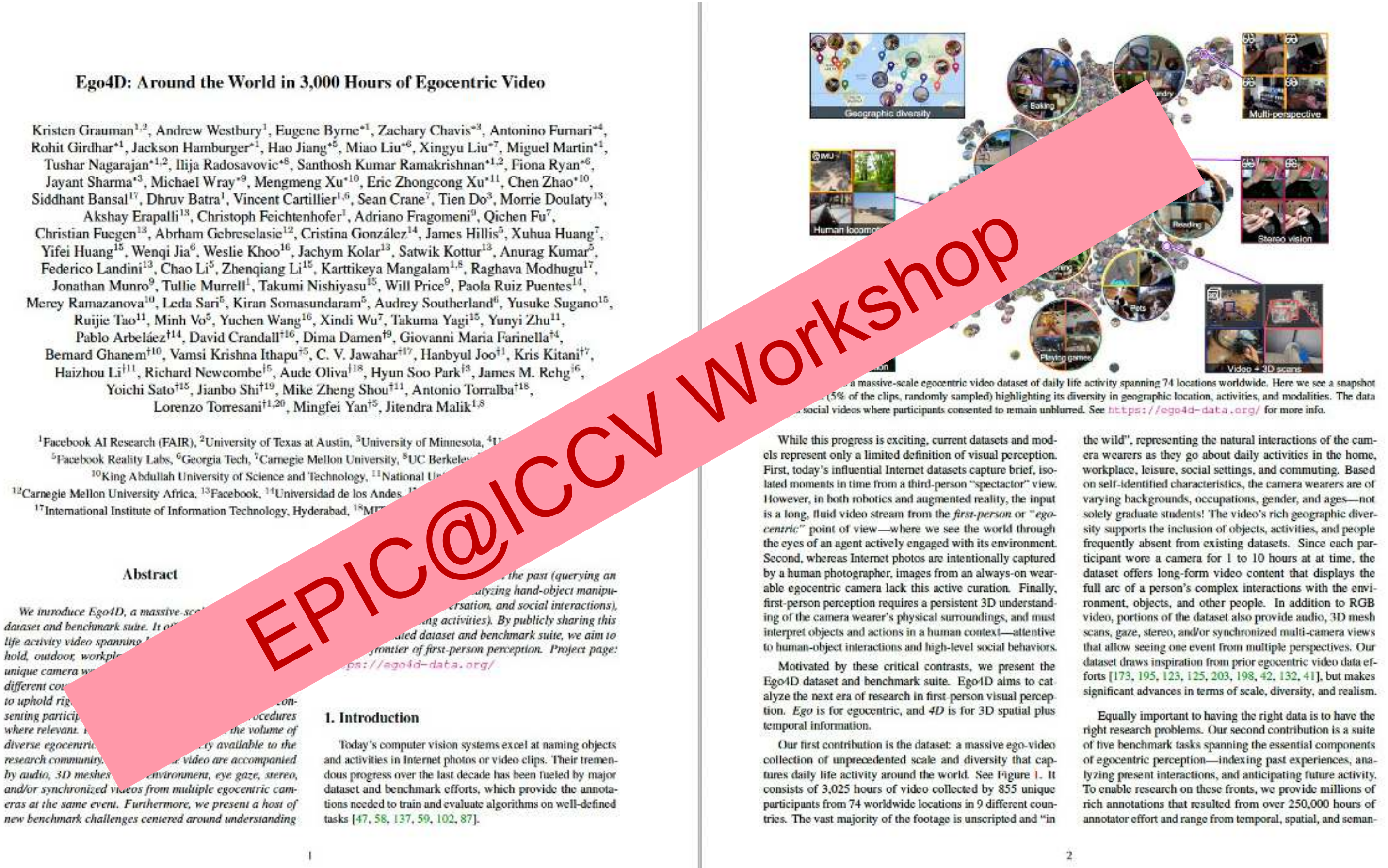

Is there enough data??

Coming soon… Ego4D Dataset

> 3400 hours of Egocentric Data

Ego4D

with: Kristen Grauman
+83 authors



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

Kristen Grauman[1,2], Andrew Westbury[1], Eugene Byrne*[1], Zachary Chavis*[3], Antonino Furnari*[4], Rohit Girdhar*[1], Jackson Hamburger*[1], Hao Jiang*[5], Miao Liu*[6], Xingyu Liu*[7], Miguel Martin*[1], Tushar Nagarajan*[1,2], Ilija Radosavovic*[8], Santhosh Kumar Ramakrishnan*[1,2], Fiona Ryan*[6], Jayant Sharma*[3], Michael Wray*[9], Mengmeng Xu*[10], Eric Zhongcong Xu*[11], Chen Zhao*[10], Siddhant Bansal[17], Dhruv Batra[1], Vincent Cartillier[1,6], Sean Crane[7], Tien Do[3], Morrie Doulaty[13], Akshay Erapalli[13], Christoph Feichtenhofer[1], Adriano Fragomeni[9], Qichen Fu[7], Christian Fuegen[13], Abrham Gebreselasie[12], Cristina González[14], James Hillis[5], Xuhua Huang[7], Yifei Huang[15], Wenqi Jia[6], Weslie Khoo[16], Jachym Kolar[13], Satwik Kottur[13], Anurag Kumar[5], Federico Landini[13], Chao Li[5], Zhenqiang Li[15], Karttikeya Mangalam[1,8], Raghava Modhugu[17], Jonathan Munro[9], Tullie Murrell[1], Takumi Nishiyasu[15], Will Price[9], Paola Ruiz Puentes[14], Merey Ramazanova[10], Leda Sari[5], Kiran Somasundaram[5], Audrey Southerland[6], Yusuke Sugano[15], Ruijie Tao[11], Minh Vo[5], Yuchen Wang[16], Xindi Wu[7], Takuma Yagi[15], Yunyi Zhu[11], Pablo Arbeláez†[14], David Crandall†[16], Dima Damen†[9], Giovanni Maria Farinella†[4], Bernard Ghanem†[10], Vamsi Krishna Ithapu†[5], C. V. Jawahar†[17], Hanbyul Joo†[1], Kris Kitani†[7], Haizhou Li†[11], Richard Newcombe†[5], Aude Oliva†[18], Hyun Soo Park†[3], James M. Rehg†[6], Yoichi Sato†[15], Jianbo Shi†[19], Mike Zheng Shou†[11], Antonio Torralba†[18], Lorenzo Torresani†[1,20], Mingfei Yan†[5], Jitendra Malik[1,8]

[1]Facebook AI Research (FAIR), [2]University of Texas at Austin, [3]University of Minnesota, [4]U
[5]Facebook Reality Labs, [6]Georgia Tech, [7]Carnegie Mellon University, [8]UC Berkele
[10]King Abdullah University of Science and Technology, [11]National U
[12]Carnegie Mellon University Africa, [13]Facebook, [14]Universidad de los Andes,
[17]International Institute of Information Technology, Hyderabad, [18]MI

Abstract

We introduce Ego4D, a massive-sc… dataset and benchmark suite. It o… life activity video spanning… hold, outdoor, workpl… unique camera w… different co… to uphold rig… senting particip… where relevant… diverse egocentri… research community… video are accompanied by audio, 3D meshes… 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 past (querying an … alyzing hand-object manipu… ersation, and social interactions), … ng activities). By publicly sharing this … ed dataset and benchmark suite, we aim to … frontier of first-person perception. Project page: …ps://ego4d-data.org/

1. Introduction

Today's computer vision systems excel at naming objects and activities in Internet photos or video clips. Their tremendous progress over the last decade has been fueled by major dataset and benchmark efforts, which provide the annotations needed to train and evaluate algorithms on well-defined tasks [47, 58, 137, 59, 102, 87].

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

While this progress is exciting, current datasets and models represent only a limited definition of visual perception. First, today's influential Internet datasets capture brief, isolated moments in time from a third-person "spectactor" view. However, in both robotics and augmented reality, the input is a long, fluid video stream from the *first-person* or "*egocentric*" point of view—where we see the world through the eyes of an agent actively engaged with its environment. Second, whereas Internet photos are intentionally captured by a human photographer, images from an always-on wearable egocentric camera lack this active curation. Finally, first-person perception requires a persistent 3D understanding of the camera wearer's physical surroundings, and must interpret objects and actions in a human context—attentive to human-object interactions and high-level social behaviors.

Motivated by these critical contrasts, we present the Ego4D dataset and benchmark suite. Ego4D aims to catalyze the next era of research in first-person visual perception. *Ego* is for egocentric, and *4D* is for 3D spatial plus temporal information.

Our first contribution is the dataset: a massive ego-video collection of unprecedented scale and diversity that captures daily life activity around the world. See Figure 1. It consists of 3,025 hours of video collected by 855 unique participants from 74 worldwide locations in 9 different countries. The vast majority of the footage is unscripted and "in the wild", representing the natural interactions of the camera wearers as they go about daily activities in the home, workplace, leisure, social settings, and commuting. Based on self-identified characteristics, the camera wearers are of varying backgrounds, occupations, gender, and ages—not solely graduate students! The video's rich geographic diversity supports the inclusion of objects, activities, and people frequently absent from existing datasets. Since each participant wore a camera for 1 to 10 hours at at time, the dataset offers long-form video content that displays the full arc of a person's complex interactions with the environment, objects, and other people. In addition to RGB video, portions of the dataset also provide audio, 3D mesh scans, gaze, stereo, and/or synchronized multi-camera views that allow seeing one event from multiple perspectives. Our dataset draws inspiration from prior egocentric video data efforts [173, 195, 123, 125, 203, 198, 42, 132, 41], but makes significant advances in terms of scale, diversity, and realism.

Equally important to having the right data is to have the right research problems. Our second contribution is a suite of five benchmark tasks spanning the essential components of egocentric perception—indexing past experiences, analyzing present interactions, and anticipating future activity. To enable research on these fronts, we provide millions of rich annotations that resulted from over 250,000 hours of annotator effort and range from temporal, spatial, and seman-

EPIC@ICCV Workshop

K Grauman et al (2021). Around the World in 3,000 Hours of Egocentric Video. ArXiv

Ego4D Narrations

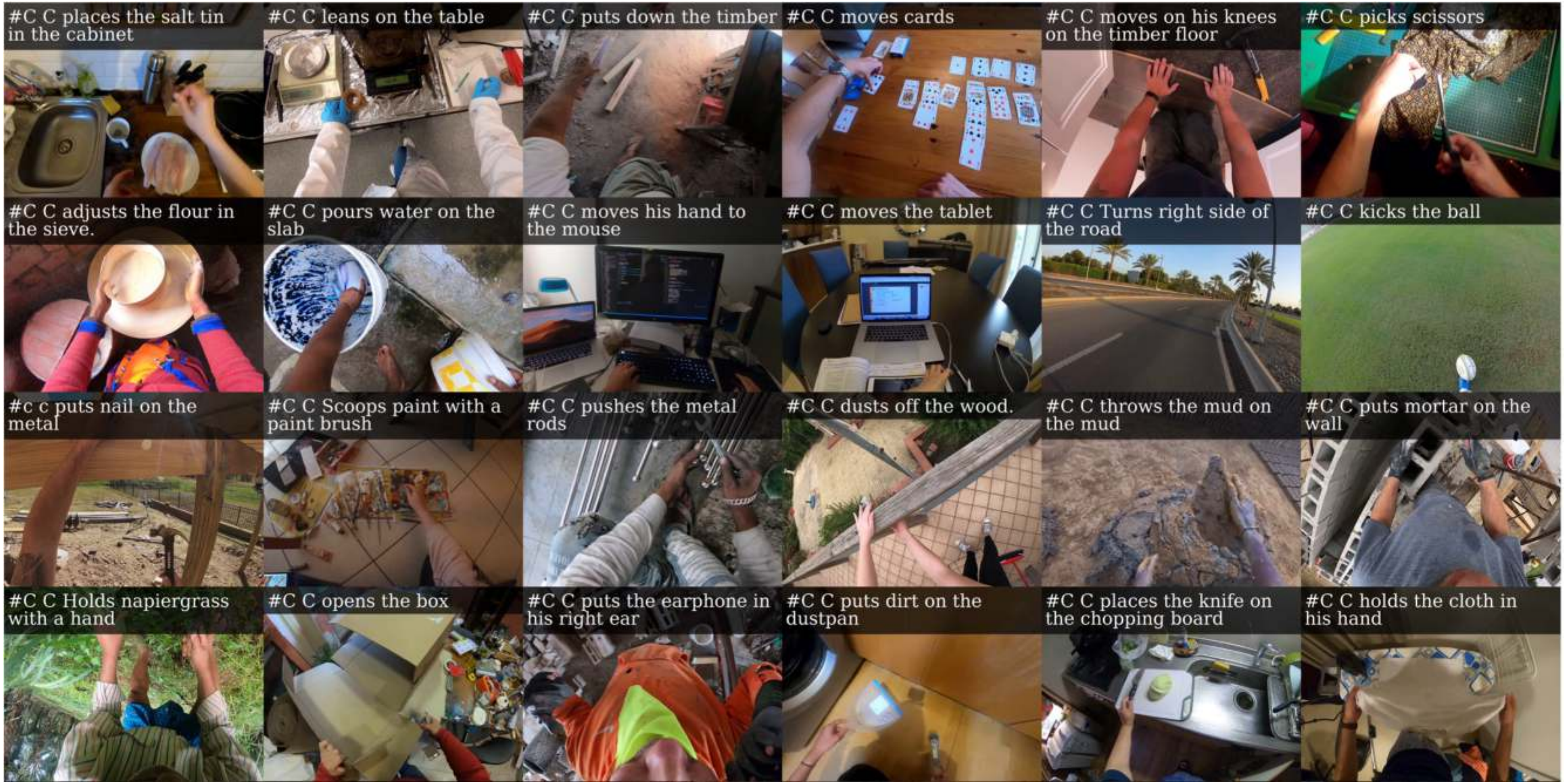
with: Kristen Grauman
+83 authors

- Inspired by the EPIC-KITCHENS narrations
- Narrators are provided the following prompt: *"Pretend as you watch this video that you are also talking to a friend on the phone, and you need to describe to your friend everything that is happening in the video. Your friend cannot see the video."*

K Grauman et al (2021). Around the World in 3,000 Hours of Egocentric Video. ArXiv

Ego4D Narrations

with: Kristen Grauman
+83 authors



K Grauman et al (2021). Around the World in 3,000 Hours of Egocentric Video. ArXiv

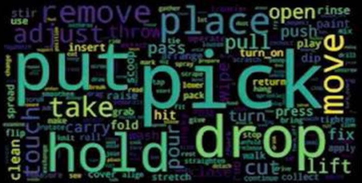
Narration

C: camera wearer

#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

Ego4D Team

Carnegie Mellon University
Ye Yuan
Abrham Gebreselasie
Xingyu Liu
Xuhua Huang
Sean Crane
Xindi Wu
Kris Kitani

Indiana University Bloomington
Yuchen Wang
Weslie Khoo
David Crandall

University of Catania
Giovanni Maria Farinella
Antonino Furnari

MIT
SouYoung Jin
Alex Lascelles
Mathew Monfort
Aude Oliva
Antonio Torralba

Georgia Tech
Fiona Ryan
Miao Liu
James M. Rehg

University of Minnesota
Jayant Sharma
Tien Do
Zachary Chavis
Hyun Soo Park

University of Bristol
Michael Wray
Will Price
Jonathan Munro
Adriano Fragomeni
Dima Damen

University of Tokyo
Takuma Yagi
Takumi Nishiyasu
Yifei Huang
Yusuke Sugano
Zhenqiang Li
Yoichi Sato

IIIT- Hyderabad
Raghava Modhugu
Siddhant Bansal
C. V. Jawahar

KAUST
Chen Zhao
Mengmeng Xu
Merey Ramazanova
Bernard Ghanem

National Univ. of Singapore
Haizhou Li
Eric Z. Xu
Ruijie Tao
Yunyi Zhu
Mike Zheng Shou

University of Los Andes
Juan C. Perez
Guillaume Jeanneret
Pablo Arbelaez

University of Pennsylvania
Jianbo Shi

Facebook AI
Aaron Adcock
Ruohan Gao
Rohit Girdhar
Kristen Grauman
Jackson Hamburger
James Hillis
Vamsi Krishna Ithapu
Hao Jiang
Hanbyul Joo
Satwik Kottur
Sanjana Krishnan
Yanghao Li
Jitendra Malik
Miguel Martin
Tullie Murrell
Tushar Nagarajan
Maja Pantic
Ilija Radosavovic
Santhosh Ramakrishnan
Lorenzo Torresani
Andrew Westbury

More?

with: Kristen Grauman
+83 authors

https://ego4d-data.org/

K Grauman et al (2021). Around the World in 3,000 Hours of Egocentric Video. ArXiv

Conclusion

- Video Understanding goals depend on the video source.
- Egocentric videos offer unscripted footage with plenty of potential
- The value of audio and language cannot be underestimated.
- Frame attribution methods can offer new insights into next steps…
- Large-scale Egocentric Vision is here (EPIC-KITCHENS)
- Massive-scale Egocentric Vision is on the way (Ego4D)

The Team


2017


2018


2019


Michael Wray
Dima Damen (University of Bristol)
Jian Ma
Toby Perrett
Hazel Doughty
Davide Moltisanti
Will Price
Jonny Munro
Adriano Fragomeni
Dan Whettam
2020


2021

Thank you

For further info, datasets, code, publications…

http://dimadamen.github.io

@dimadamen

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

Q&A