Opportunities in Egocentric Video Understanding

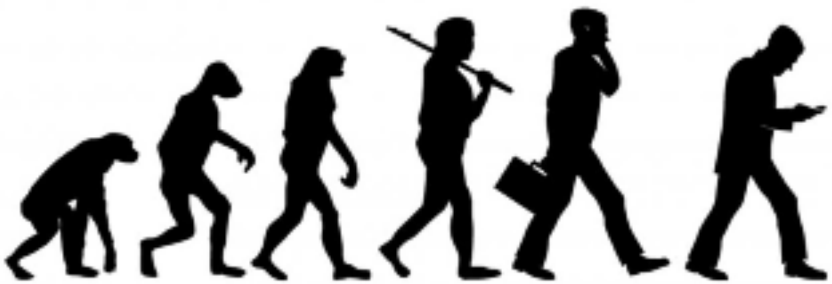
University of BRISTOL

Dima Damen
24 Oct 2022

The present…

The present…

Photo *Illustration* by Pelle Cass



The future…


HoloLens 2
A new vision for computing
FACEBOOK
PROJECT ARIA GLASSES
Samsung patent application reveals augmented reality headset design

Visual Sensors

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 6:27PM

Egocentric cameras are coming

What can we do with such footage?

Egocentric Videos?

Egocentric Videos?

Data Collection Exercise

EPIC KITCHENS

2017 - now

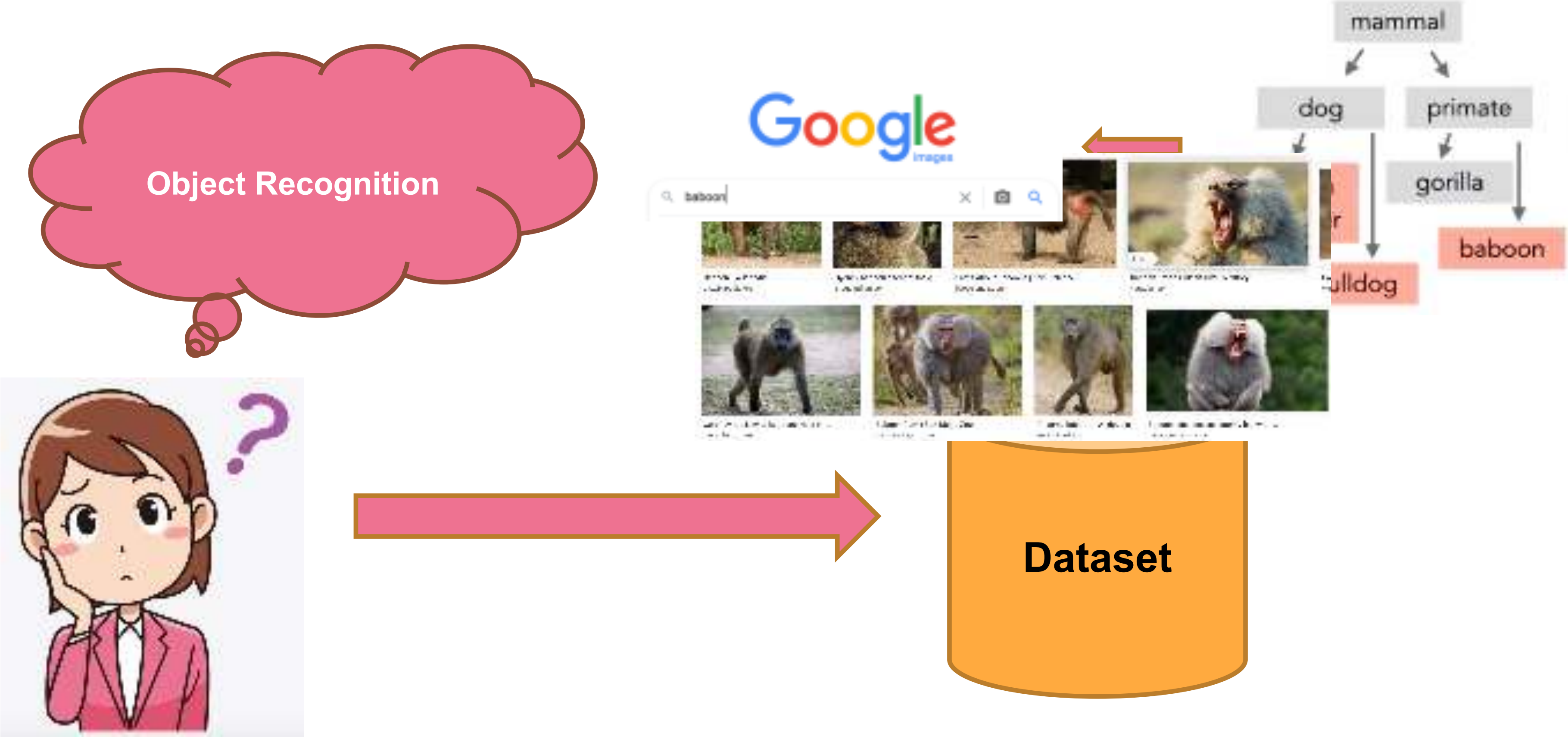
100 hours
45 kitchens
4 countries
Long-term recording
Kitchen-based activities

EGO 4D

2020 - now

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

ImageNet Dataset



Kinetics Dataset

Action Recognition

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)
7. arm wrestling (1123)
8. arranging flowers (583)
9. assembling computer (542)
10. auctioning (478)
11. baby waking up (611)
12. baking cookies (927)
13. balloon blowing (826)
14. bandaging (569)
15. barbequing (1070)

Statistics: The dataset has 400 human action classes, with 400–1150 clips for each action, each from a unique video. Each clip lasts around 10s. The current version has 306,245 videos, and is divided into three splits, one for training having 250–1000 videos per class, one for validation with 50 videos per class and one for testing with 100 videos per class. The statistics are given in table 2.

Machine Learning in Practice

- Applies to Most ML research at the moment
  - Object Recognition (Pascal, ImageNet, Places, …)
  - Action Recognition (Kinetics-400, -600, -700, AVA, SS, …)
  - …
- Datasets are
  - Overfit to the dataset
  - useful for ONE task
  - biased by choice of researchers
  - unnaturally balanced (or nearly balanced) – unrelated to priors outside the dataset itself

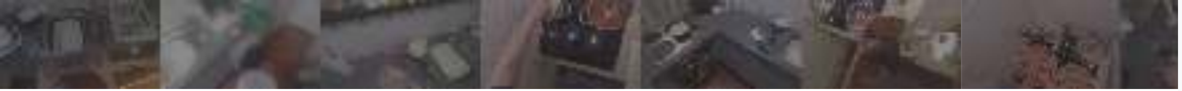
Machine Learning in Practice

~~Object Recognition~~

Let's collect Data!

EPIC KITCHENS-100

University of BRISTOL



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

EPIC-KITCHENS

Dima Damen
24 Oct 2022

Scaling and Rescaling Egocentric Vision

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

Data Collection Exercise

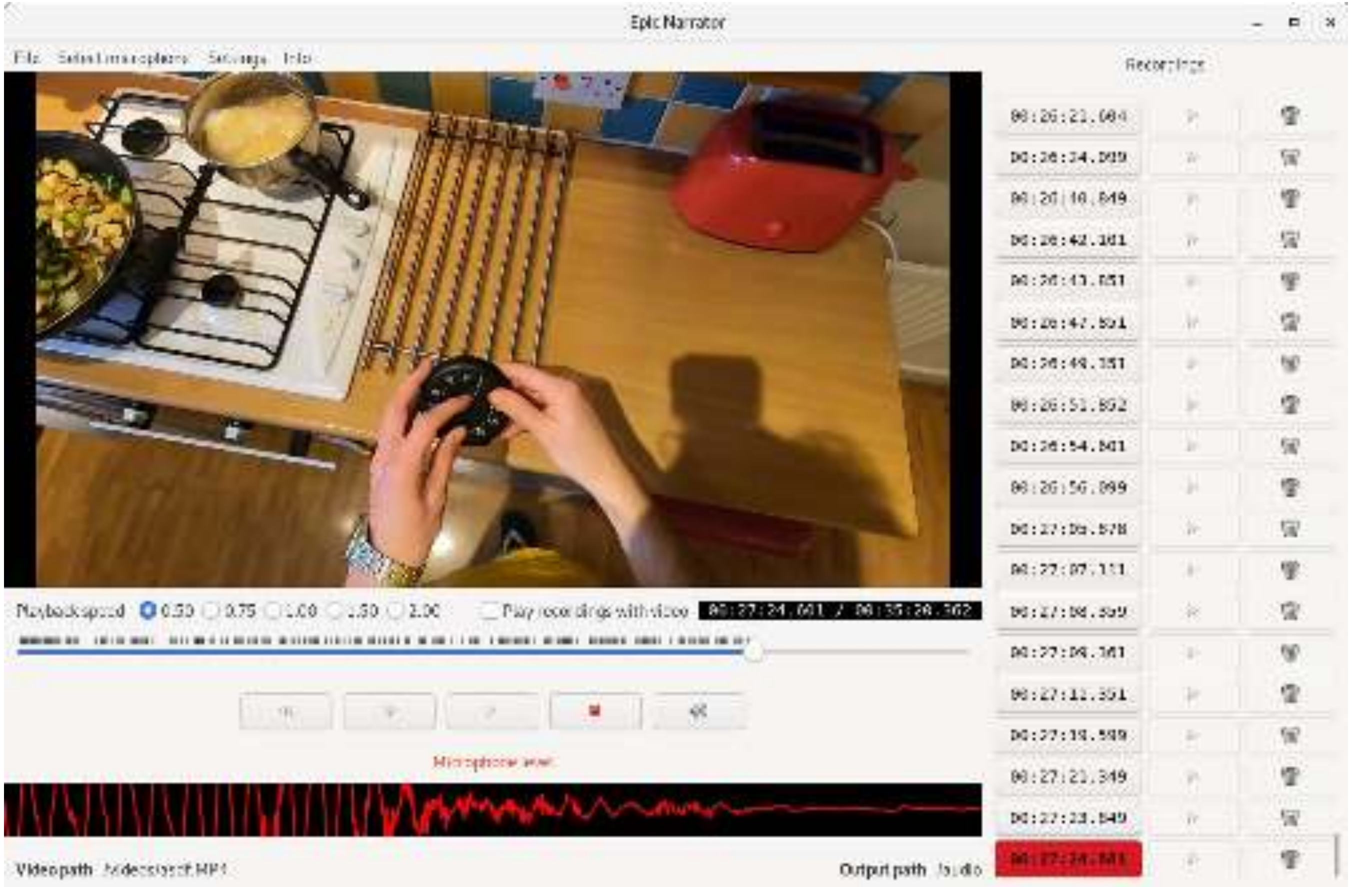
Pascal VOC
ImageNet
Kinetics
Something-Something

Labels

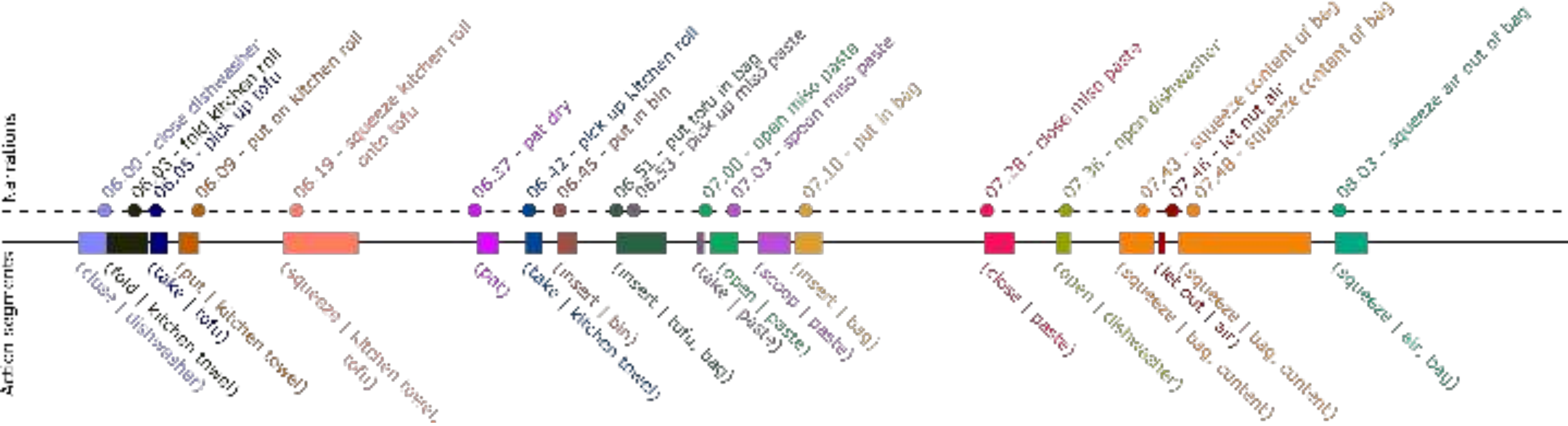
Data

EPIC-KITCHENS
Ego4D
…
KITTI

Scaling and Rescaling Egocentric Vision



Scaling and Rescaling Egocentric Vision



Dima Damen
24 Oct 2022

Ego4D

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

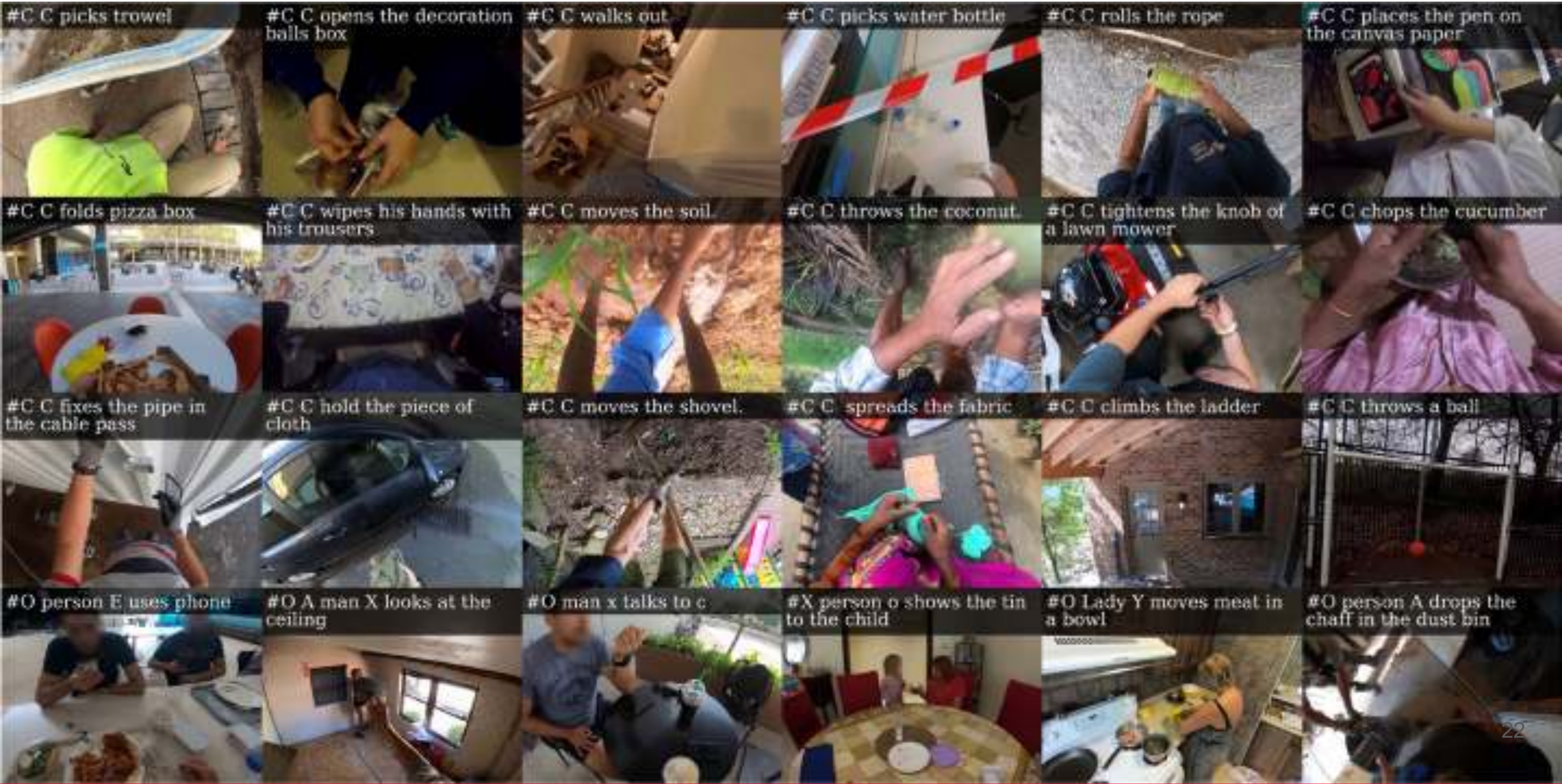
13.2 sentences/min
3.8 M sentences

1,772 verbs

4,336 nouns

Ego4D

with: Kristen Grauman
+83 authors



Scaling and Rescaling Egocentric Vision

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

stove burner
hob
stovetop gas

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

The chicken or the egg...

Data

Naturally unbalanced

Harder to label (exposes ambiguity)

Closer to application

Zero-shot, few-shot and multiple tasks

Labels

Unnaturally balanced (or nearly)

Easier to label (hides ambiguity)

Can be expanded

Single task

Tasks are harder

Detection, Recognition, 3D Mapping, Tracking, VOS …

Solutions prove more rewarding

Weak supervision, Domain Adaptation, Audio-Visual, long-term understanding

Opportunities in Egocentric Vision

Tasks are harder

Detection, Recognition, 3D Mapping, Tracking, VOS …

Solutions prove more rewarding

Weak supervision, Domain Adaptation, Audio-Visual, long-term understanding

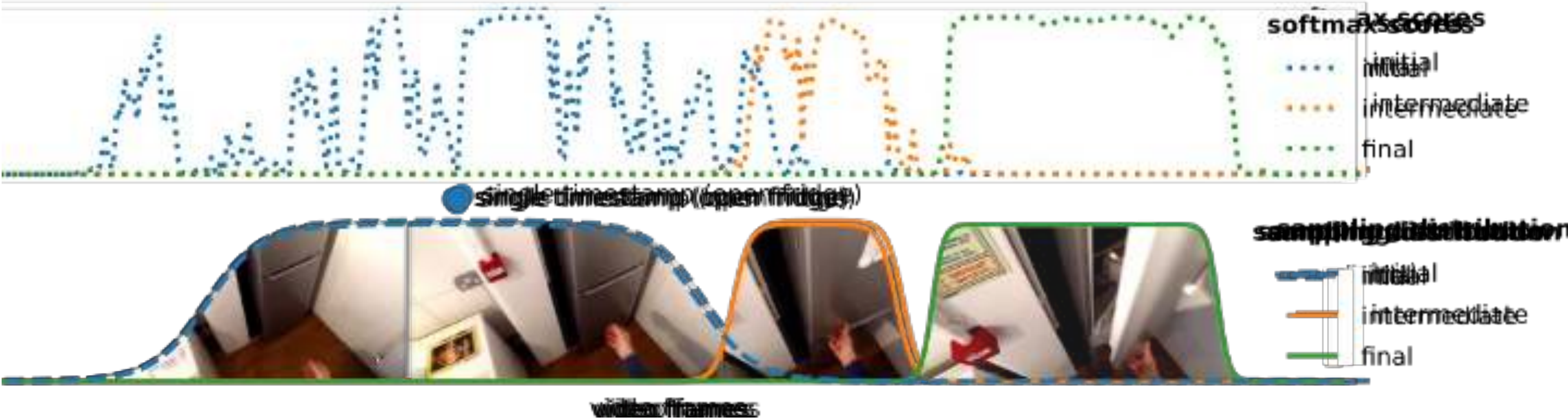
Weak Supervision from Single Timestamps

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

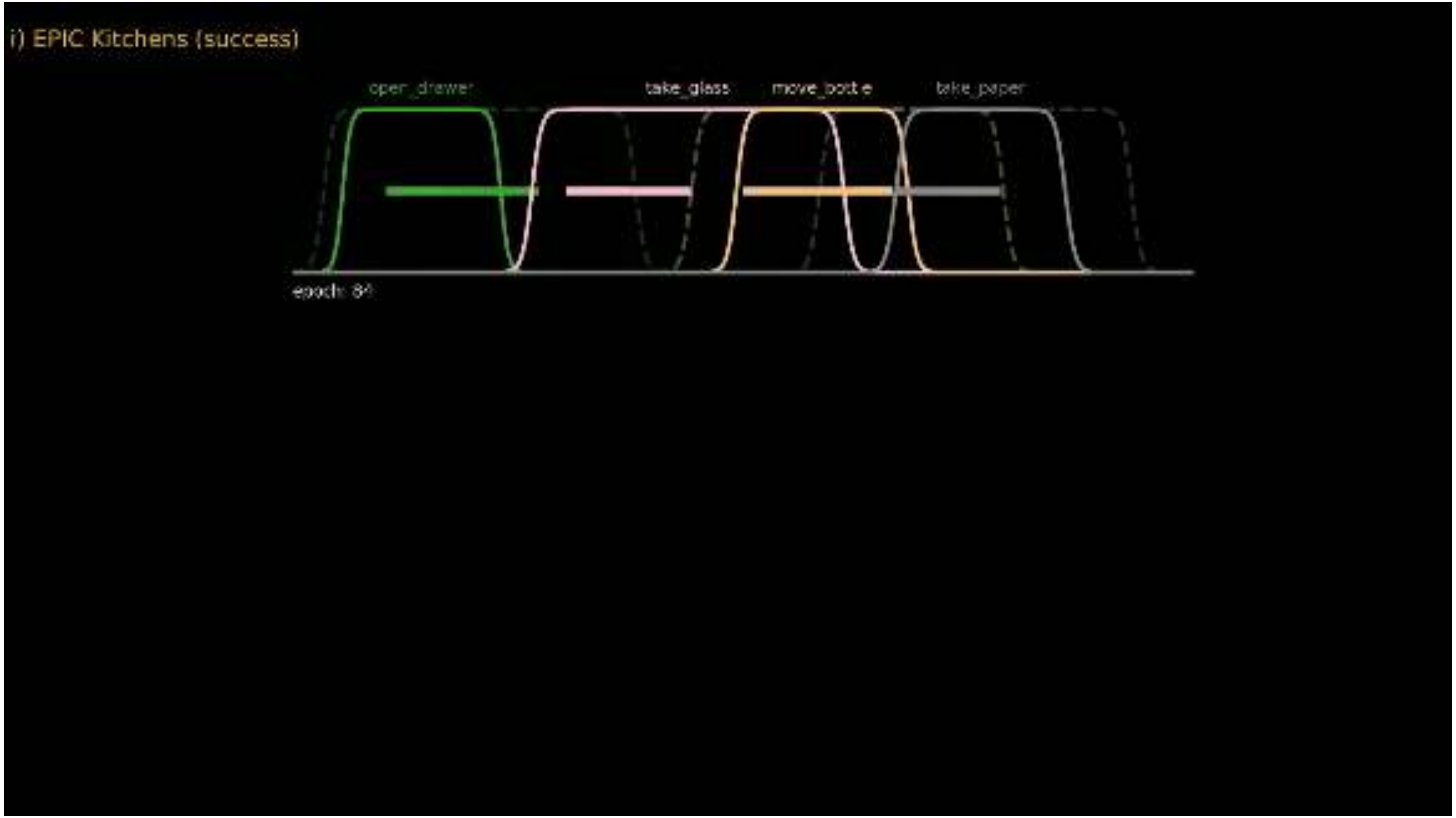
with: Davide Moltisanti
Sanja Fidler



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

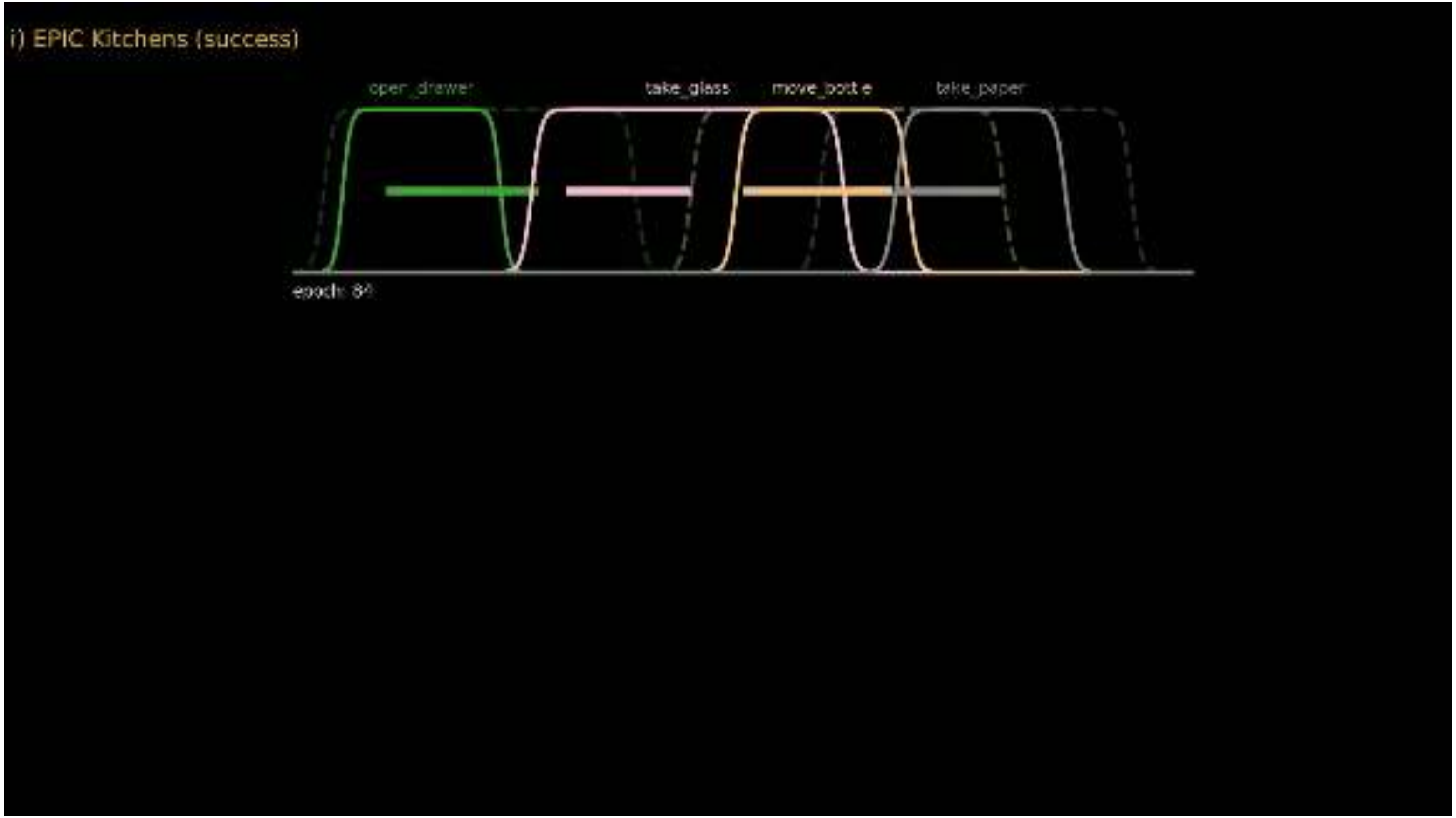
with: Davide Moltisanti
Sanja Fidler



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

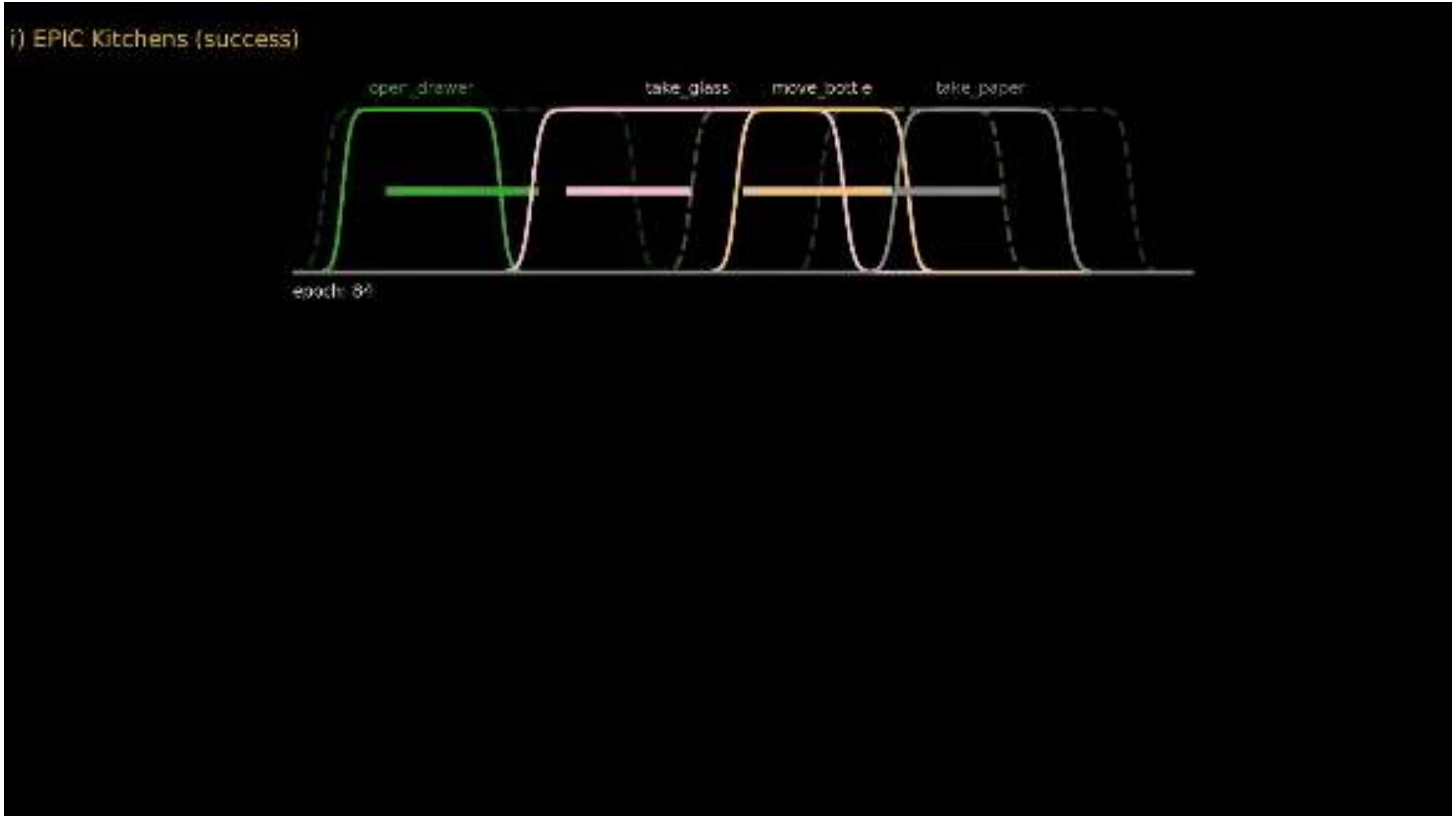
with: Davide Moltisanti
Sanja Fidler



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

Learning from a Single Timestamp

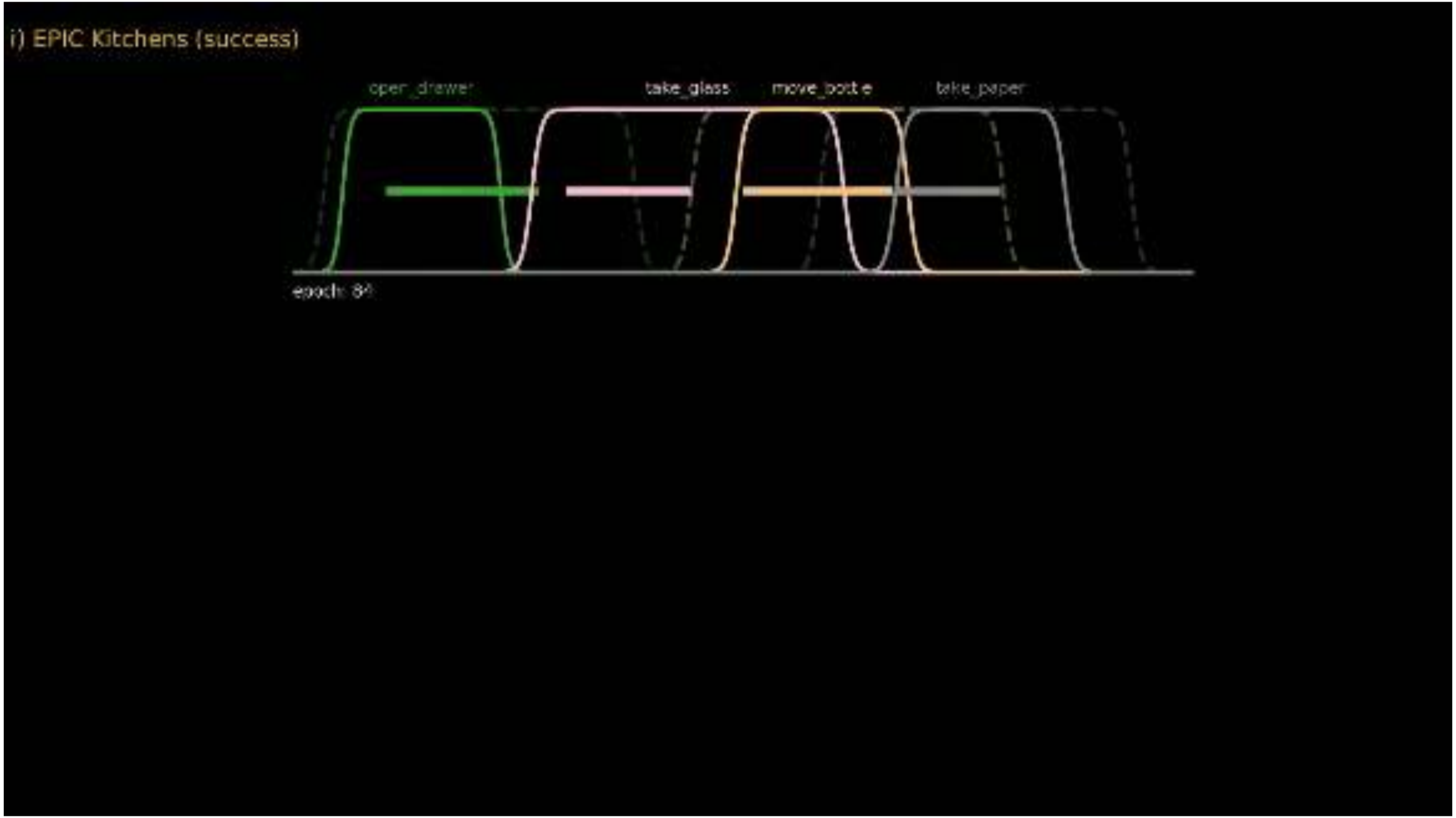
with: Davide Moltisanti
Sanja Fidler



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)*

Opportunities in Egocentric Vision

Tasks are harder

Detection, Recognition, 3D Mapping, Tracking, VOS …

Solutions prove more rewarding

Weak supervision, Domain Adaptation, 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

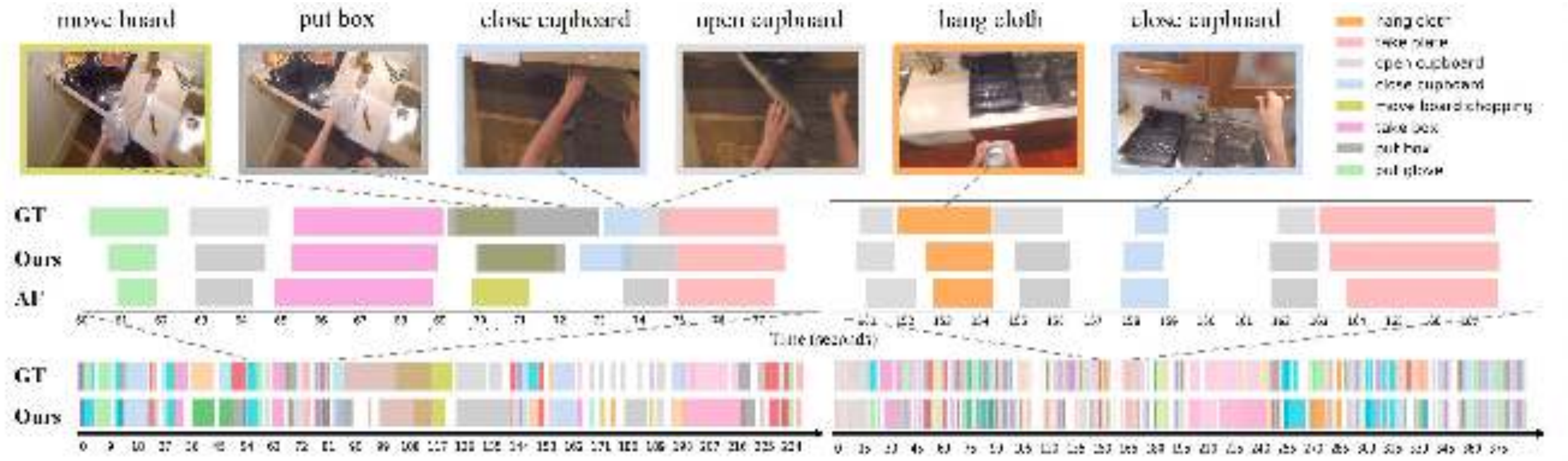


Figure 3. Qualitative results on the EPIC KITCHENS 100 validation set. Ground truth and predictions are shown with colour coded class

H Wang et al (2022). Refining Action Boundaries for One-stage Detection. *AVSS*

Opportunities in Egocentric Vision

Tasks are harder

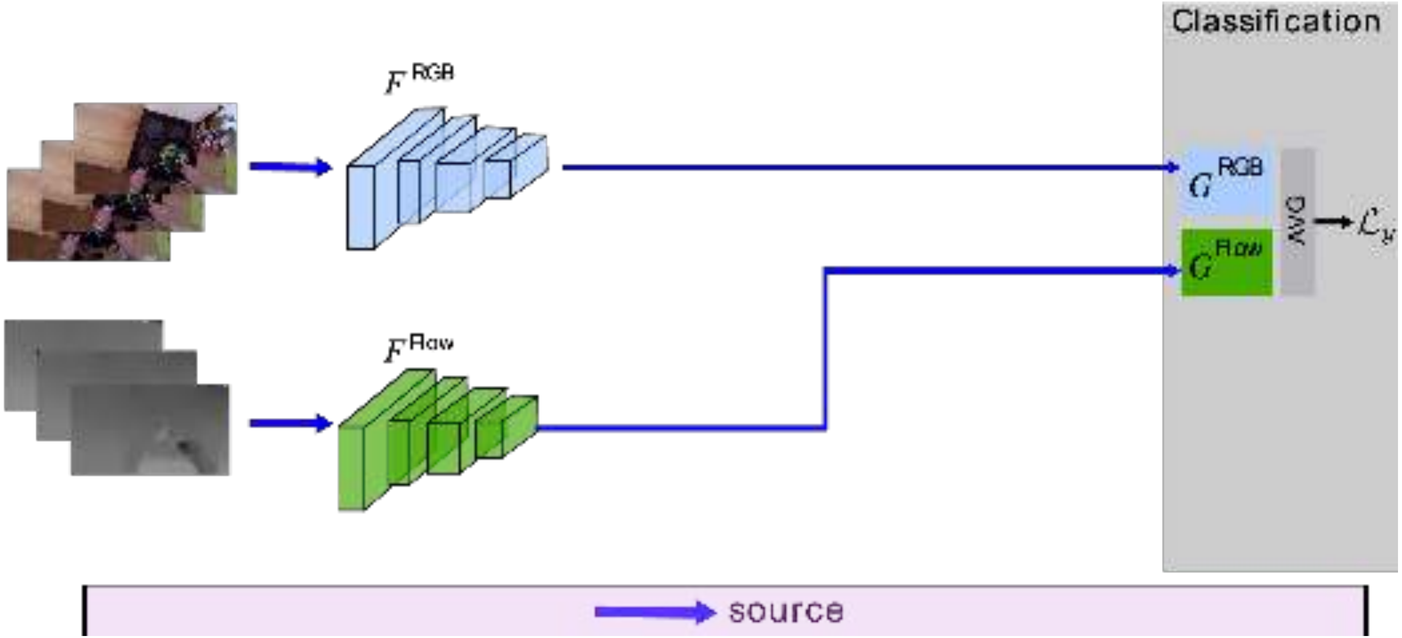
Detection, Recognition, 3D Mapping, Tracking, VOS …

Solutions prove more rewarding

Weak supervision, Domain Adaptation, Audio-Visual, long-term understanding

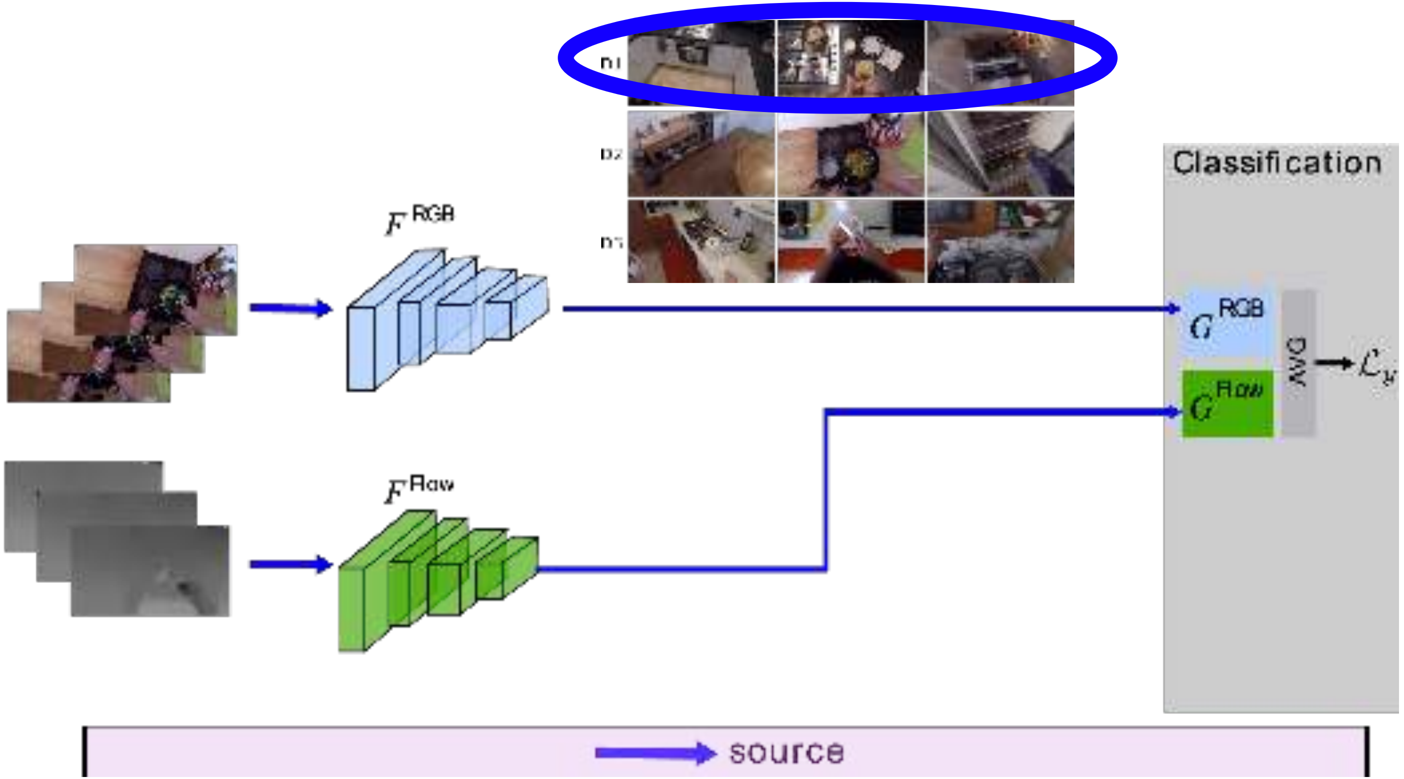
Multi-modal UDA

with: Jonathan Munro



Multi-modal UDA

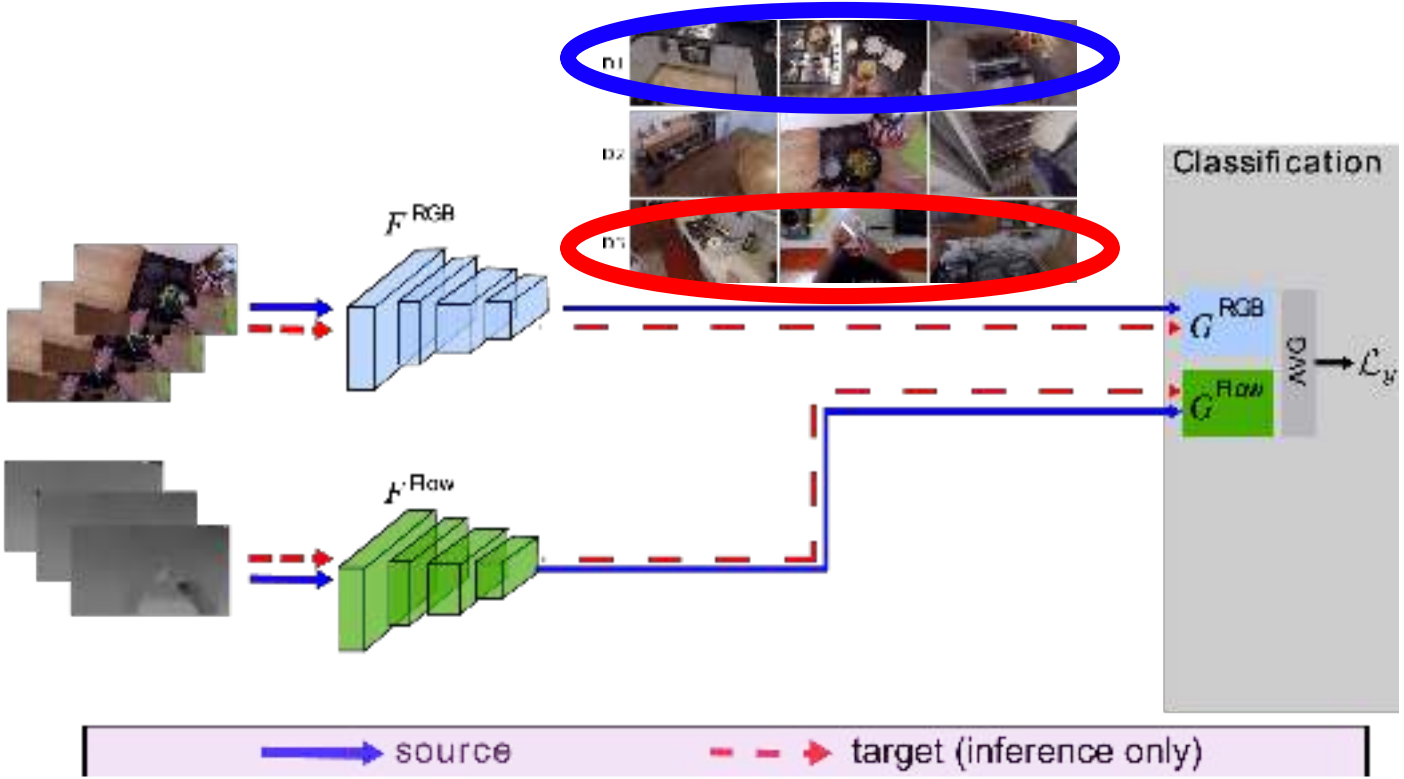
with: Jonathan Munro



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

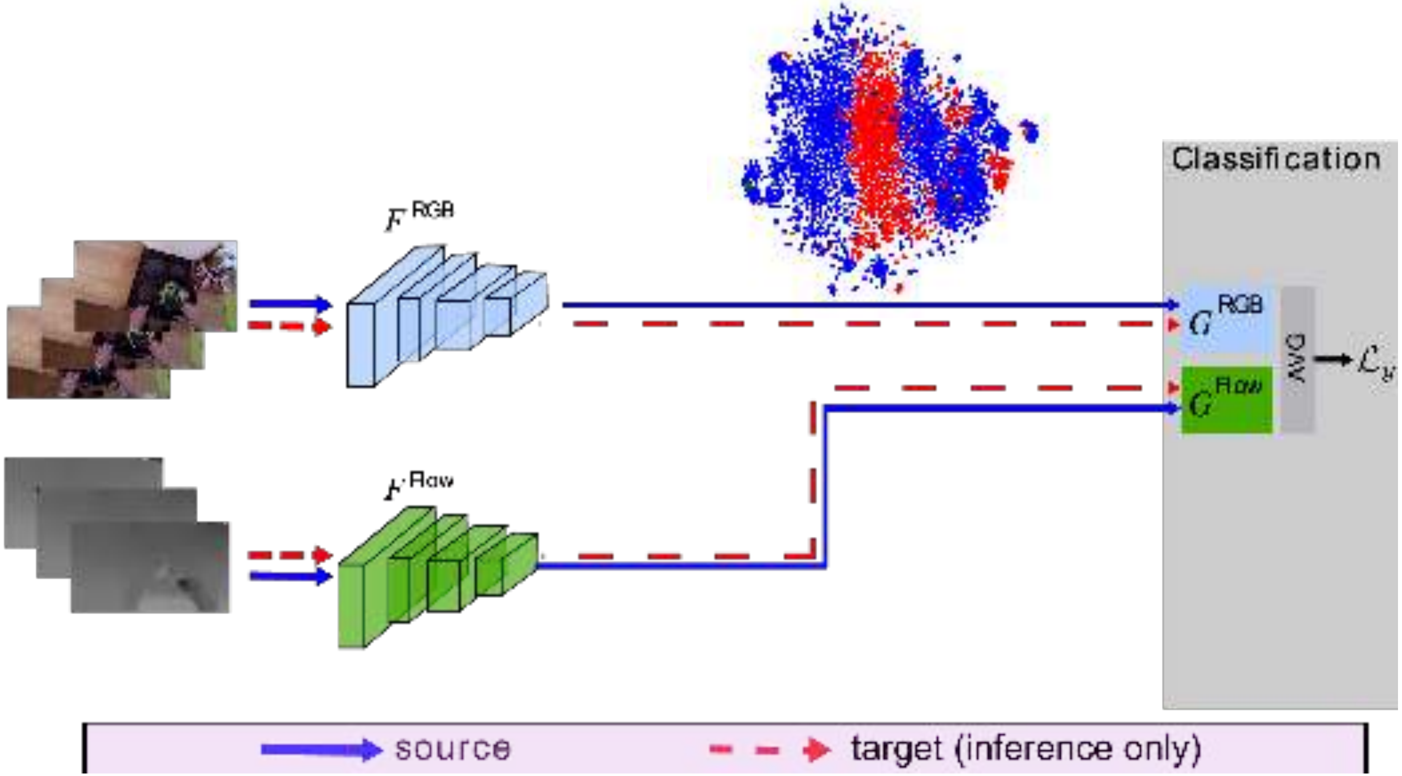


University of BRISTOL

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

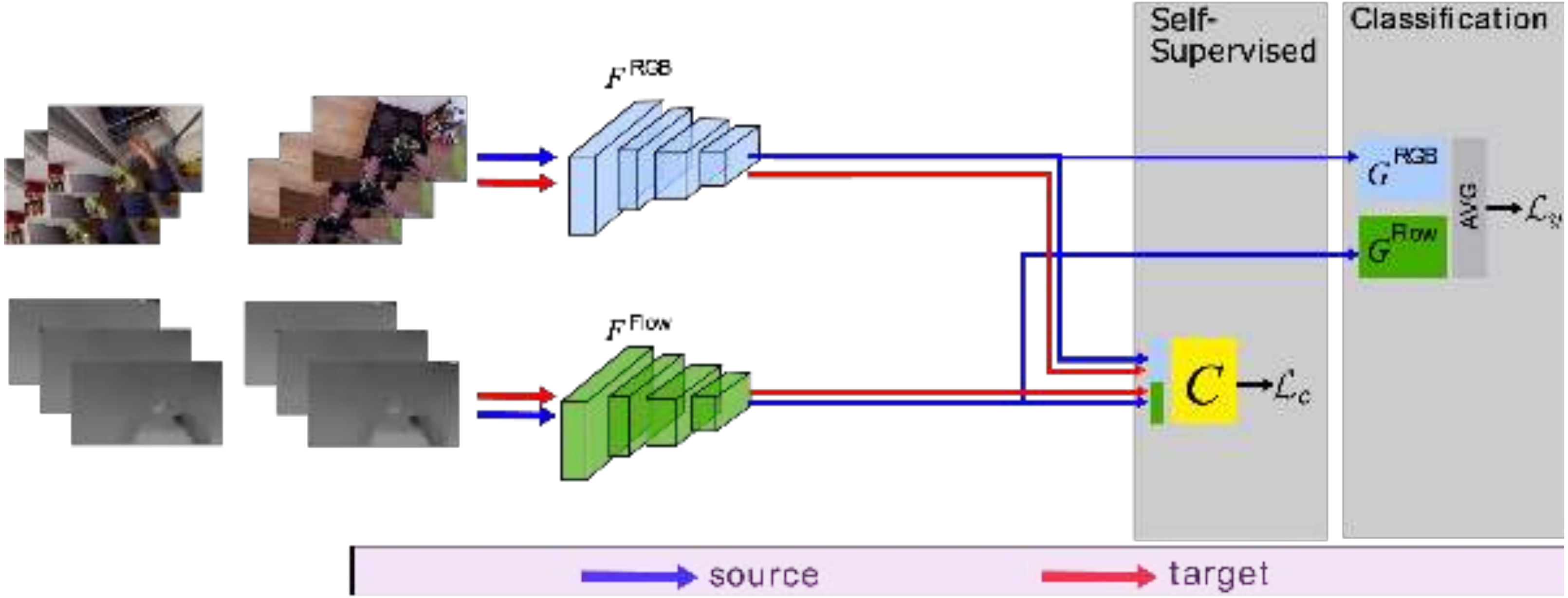
Multi-modal UDA

with: Jonathan Munro



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

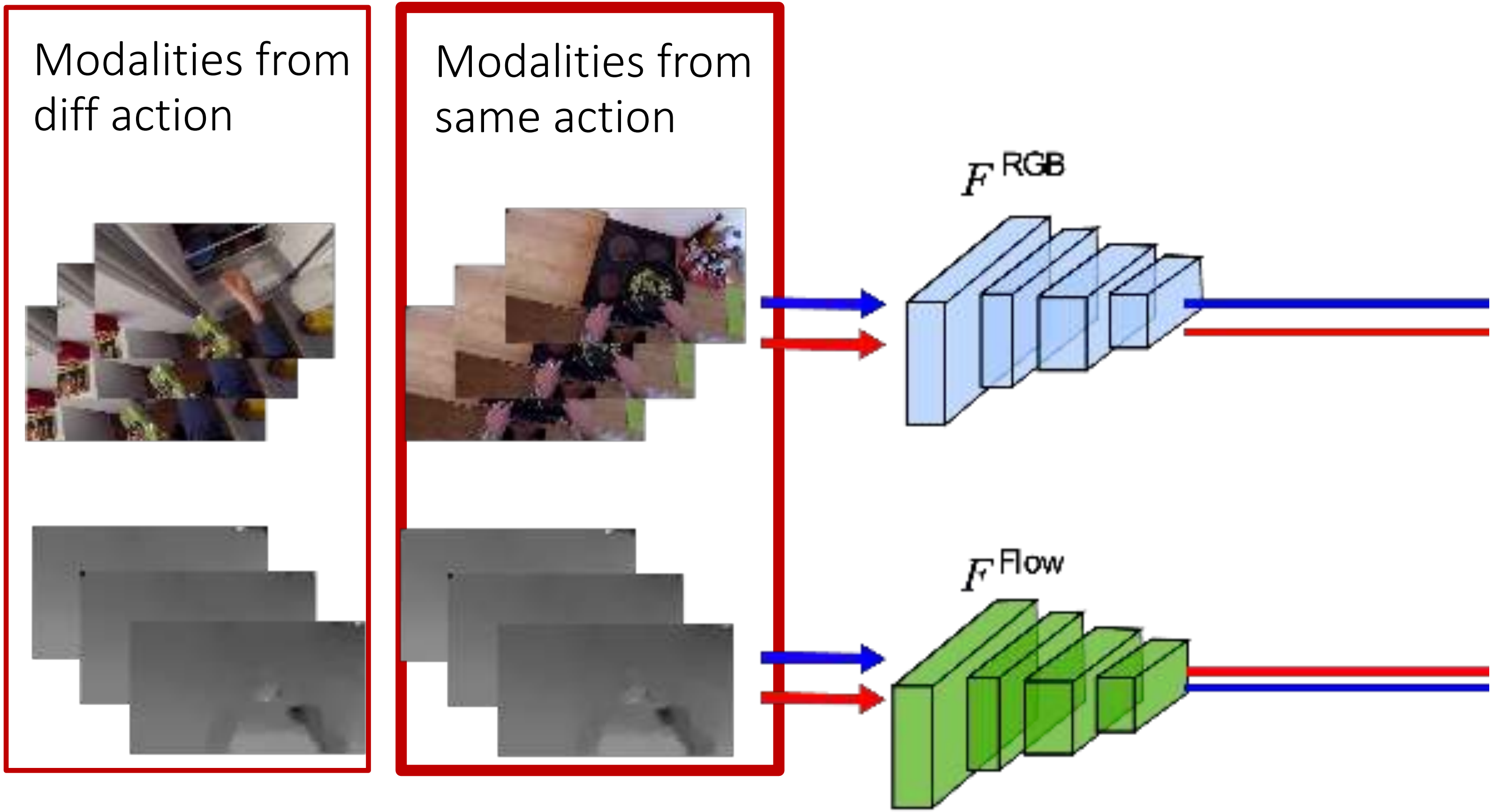
Multi-modal UDA



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

Multi-modal UDA

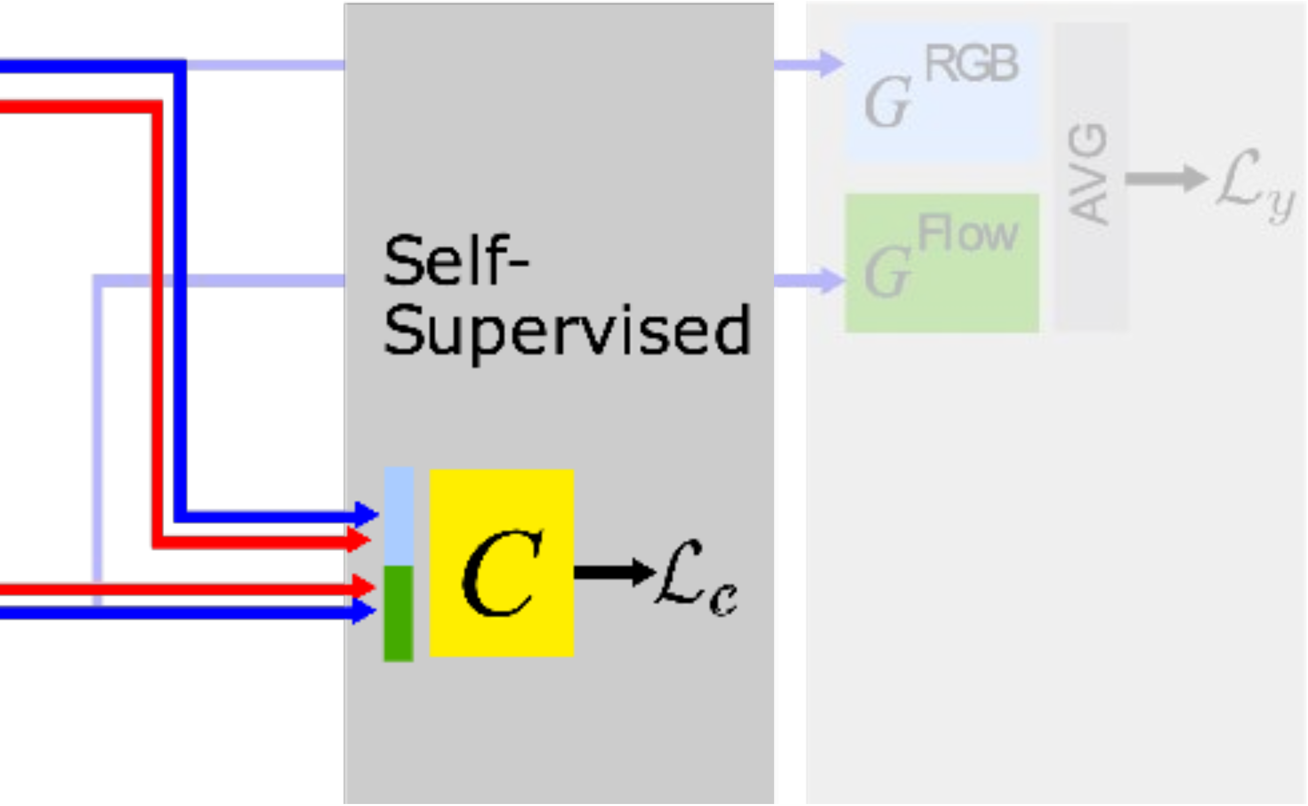
with: Jonathan Munro



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

Multi-modal UDA

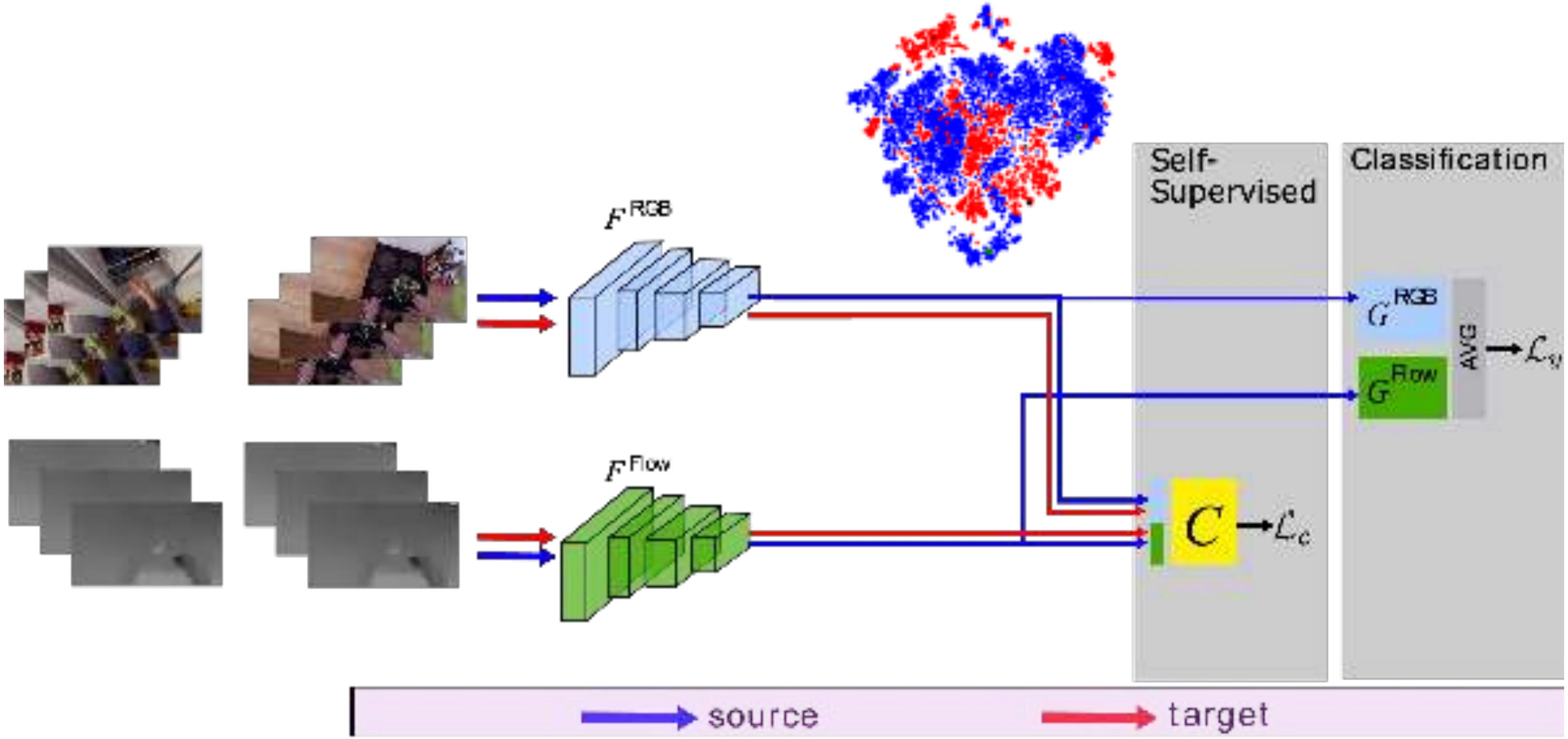
with: Jonathan Munro



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

Multi-modal UDA

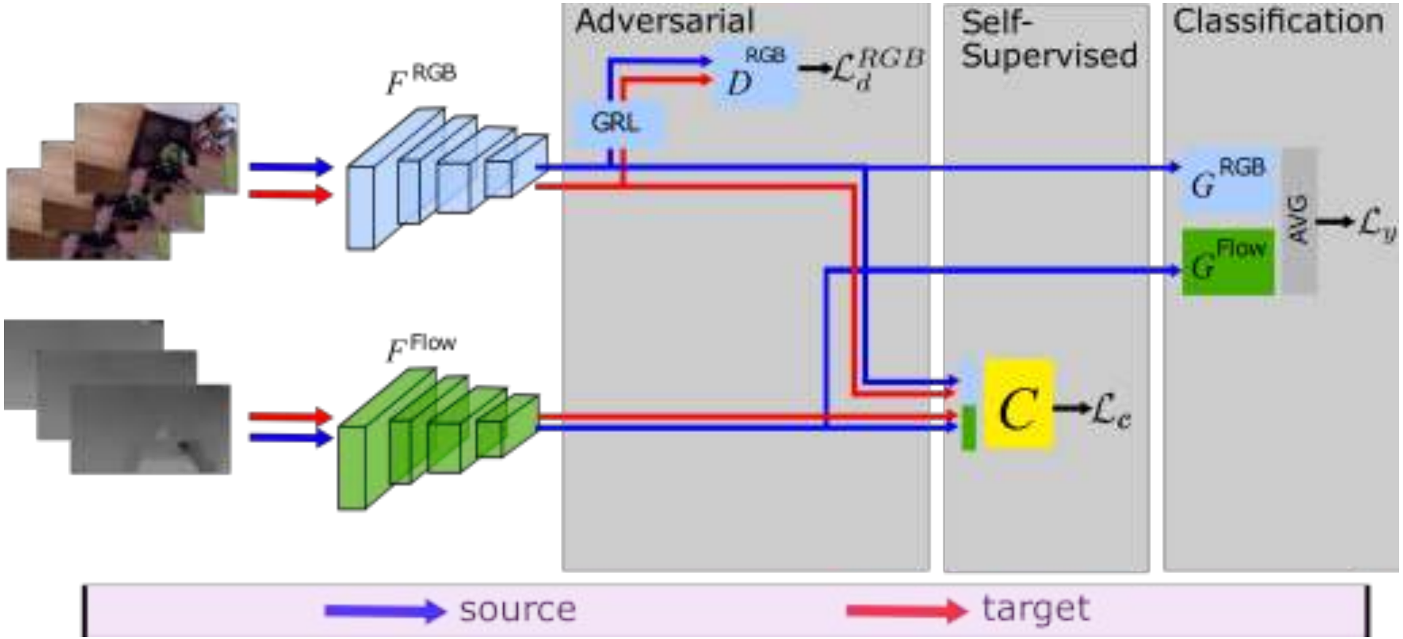
with: Jonathan Munro



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

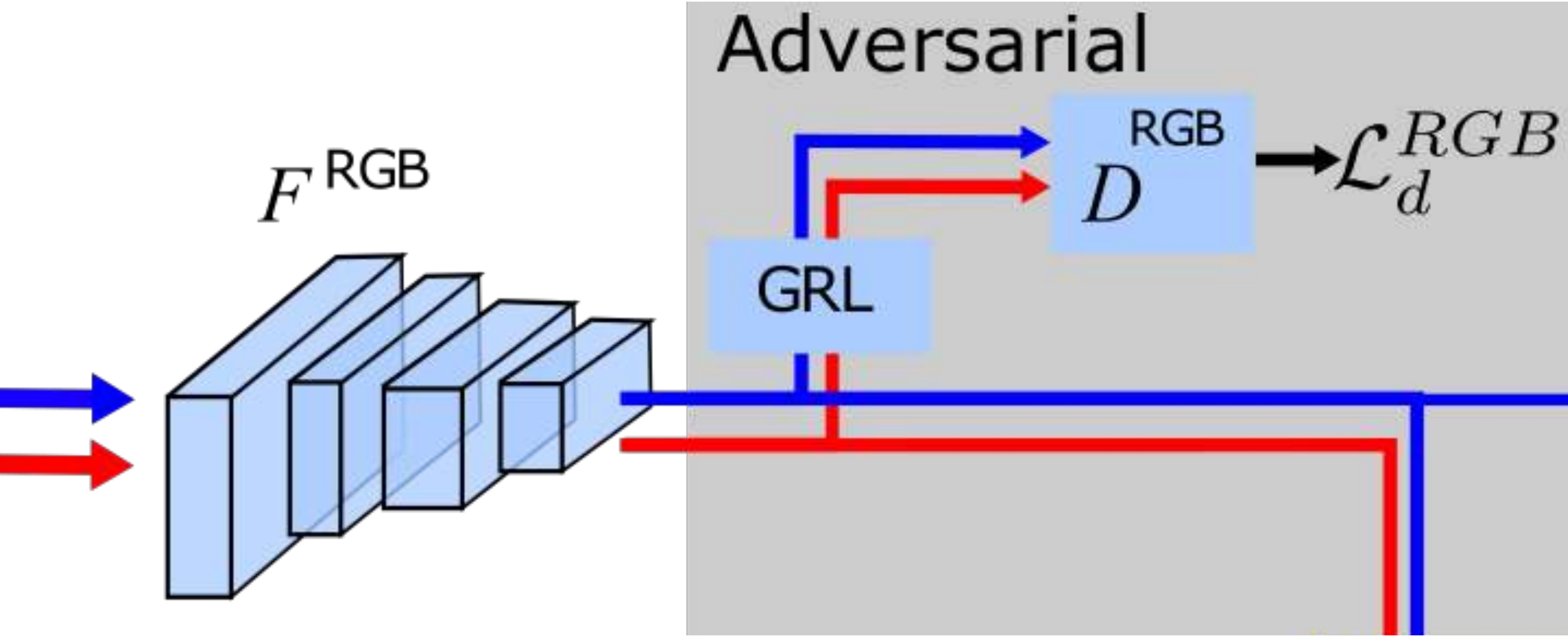
Multi-modal UDA

with: Jonathan Munro



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

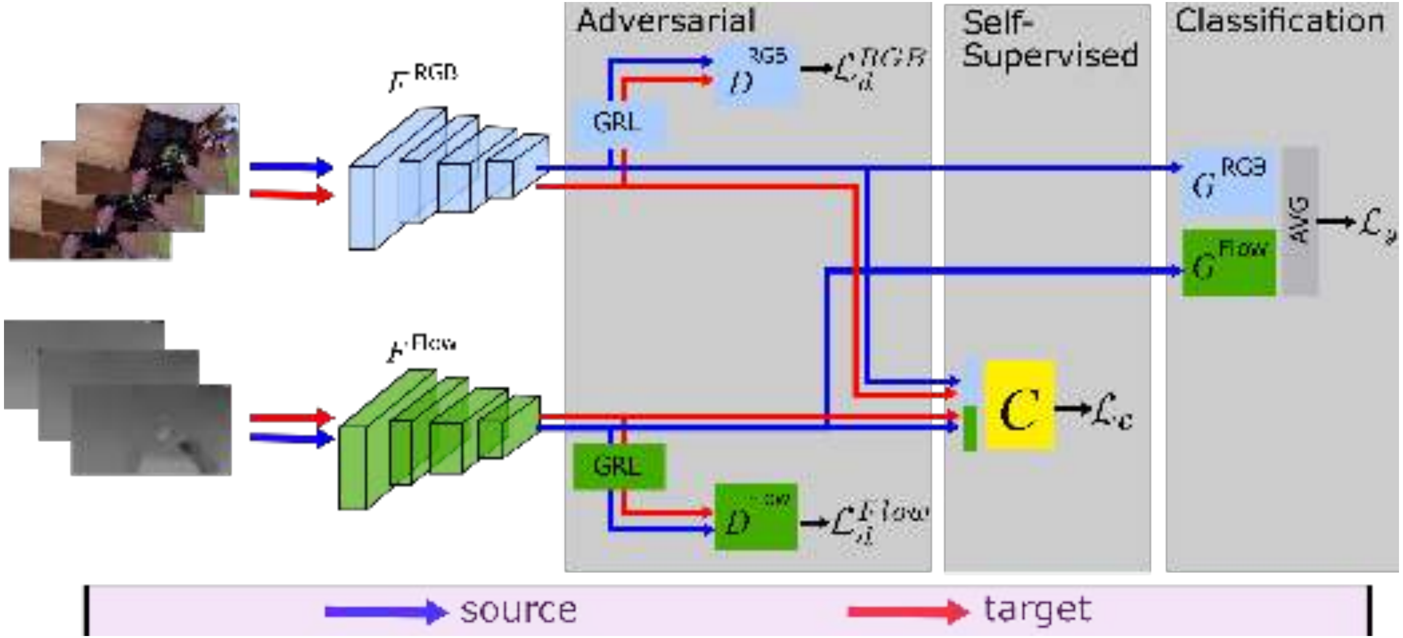
Multi-modal UDA



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

Multi-modal UDA

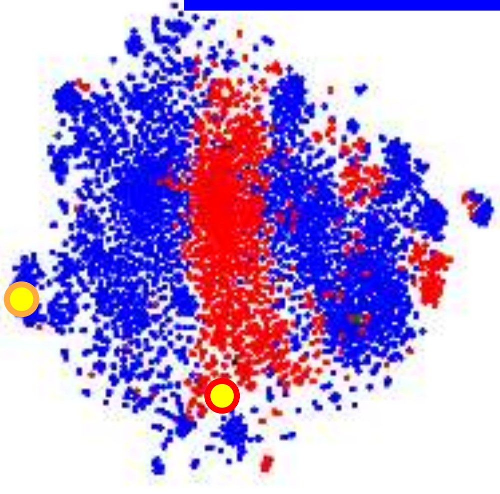
with: Jonathan Munro



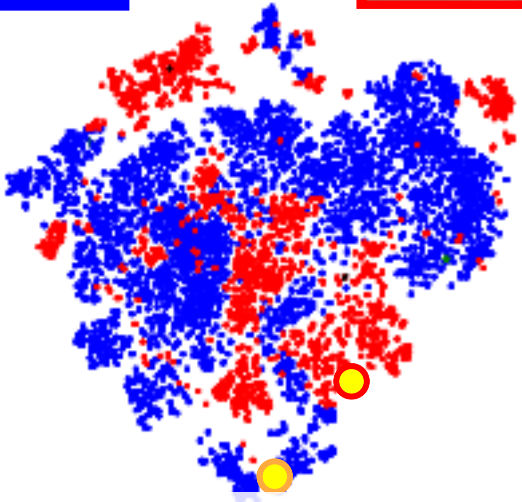
J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

Multi-modal UDA

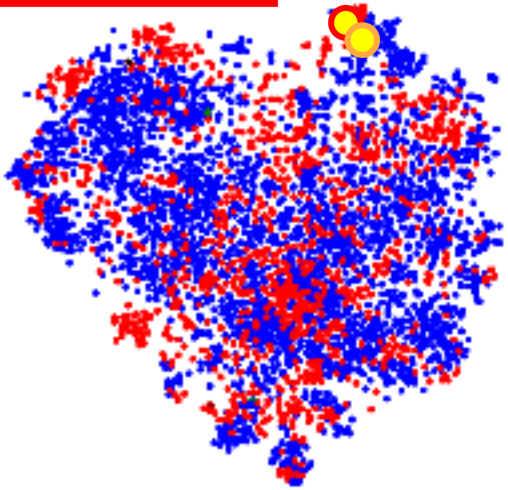
with: Jonathan Munro



Source-Only



Self-Supervision

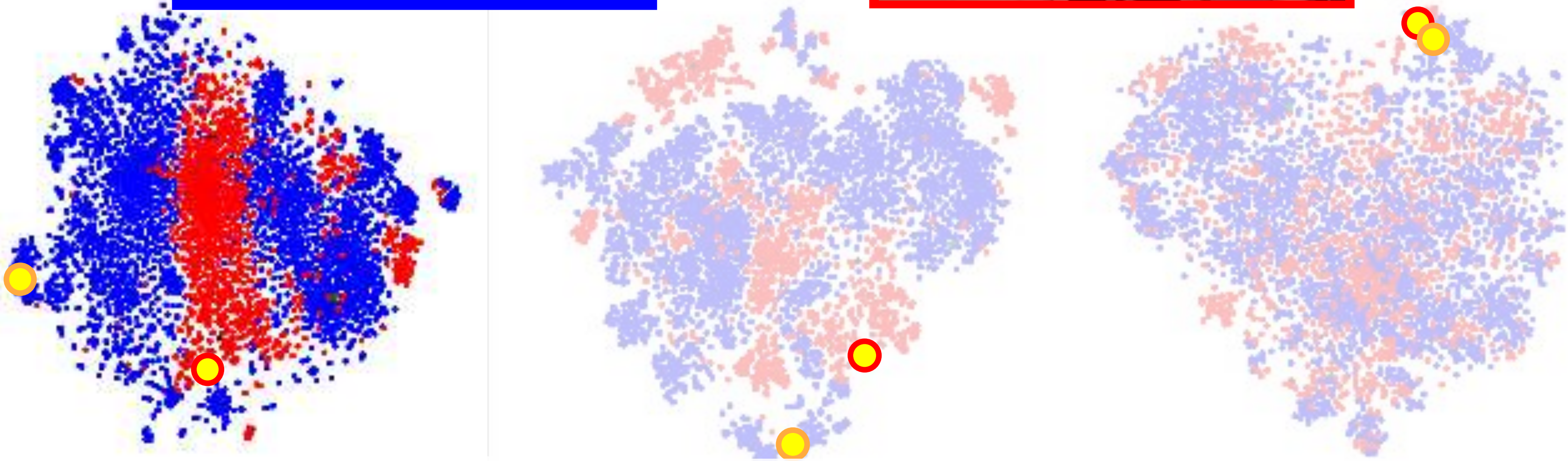


MM-SADA

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

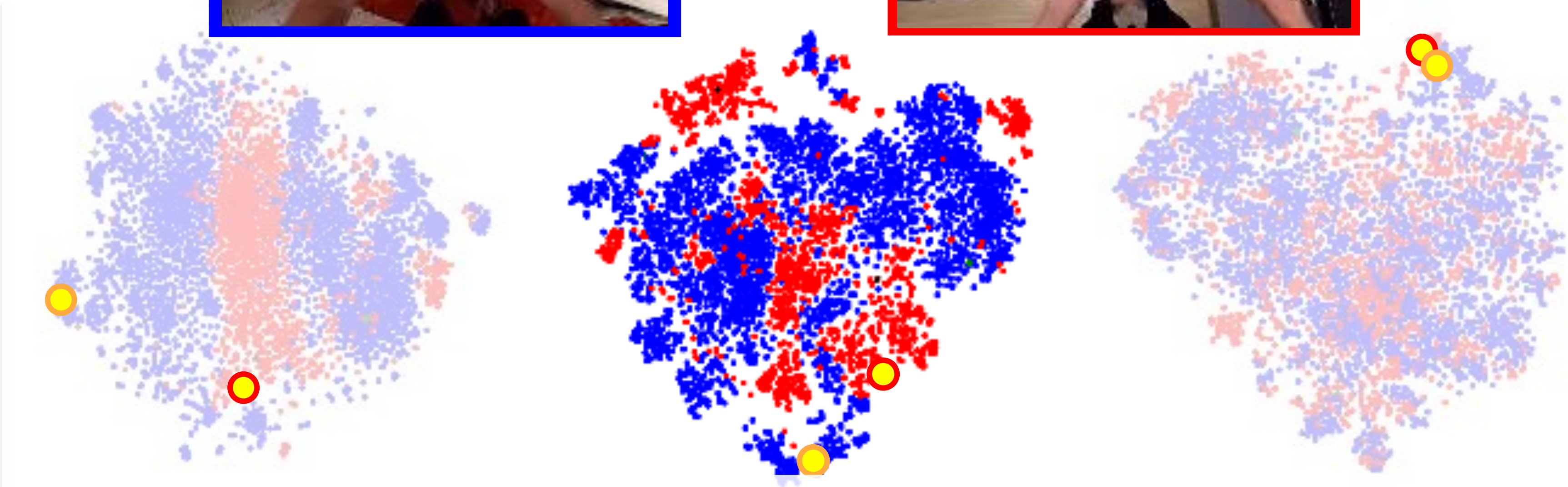
Self-Supervision

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

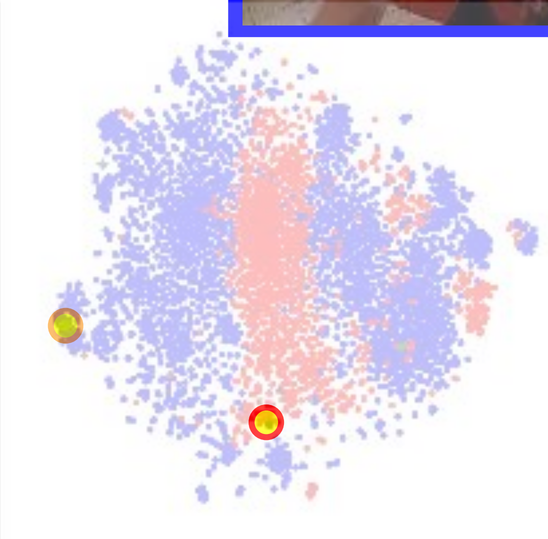


Source-Only | Self-Supervision | **MM-SADA**

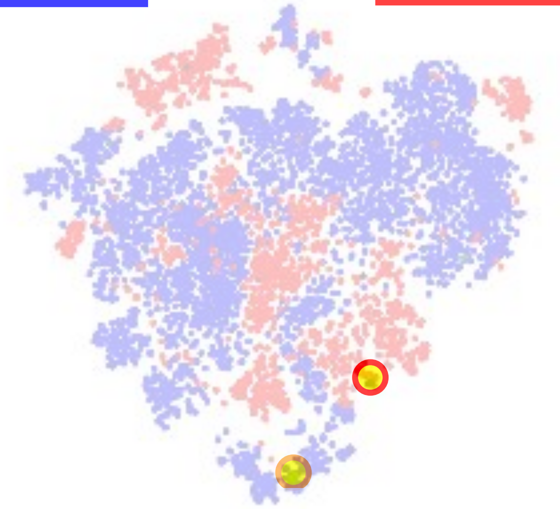
J Munro, D Damen (2020). Multi-Modal Domain Adaptation for Fine-Grained Action Recognition. *Computer Vision and Pattern Recognition (CVPR)*

Multi-modal UDA

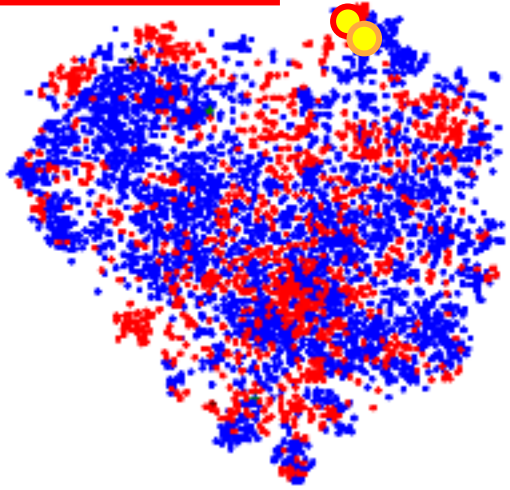
with: Jonathan Munro



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

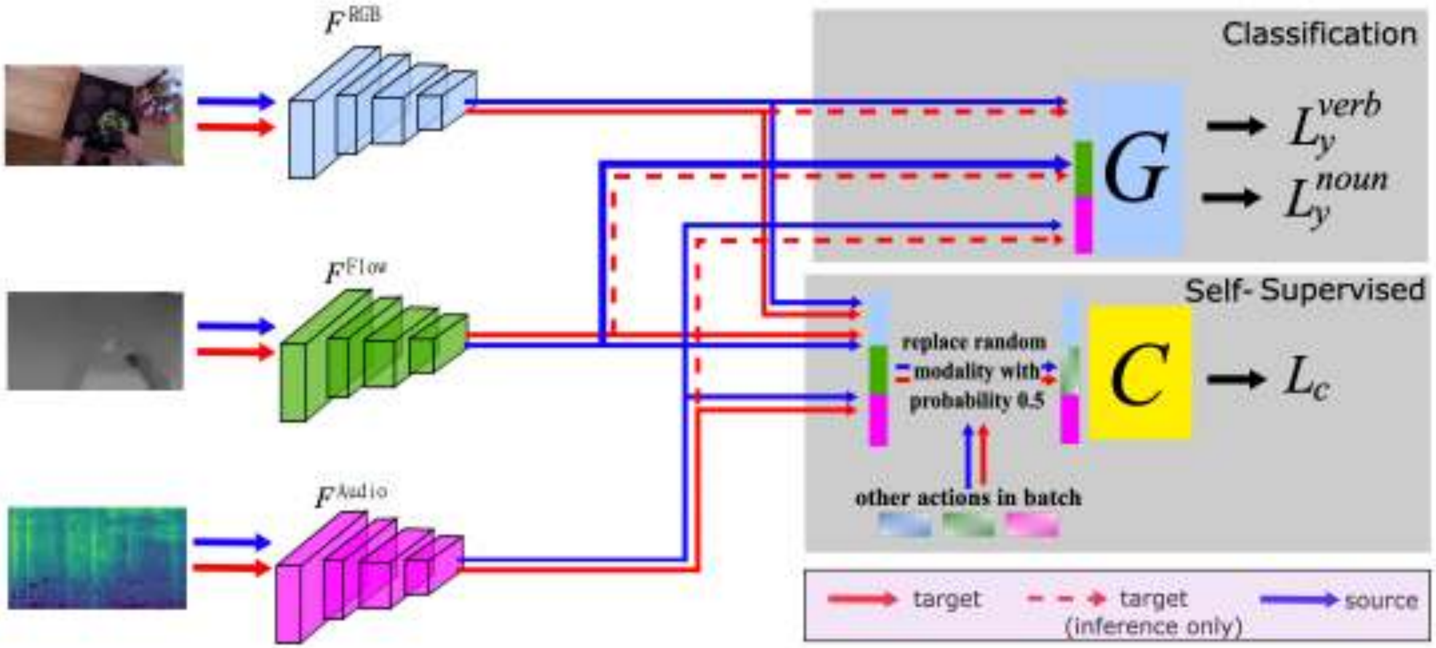


Figure 4.12: Temporal Binding Window Network (TBN) [23] with multi-modal self-supervision. A random modality is replaced with another instance from the batch to generate non-corresponding examples.

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

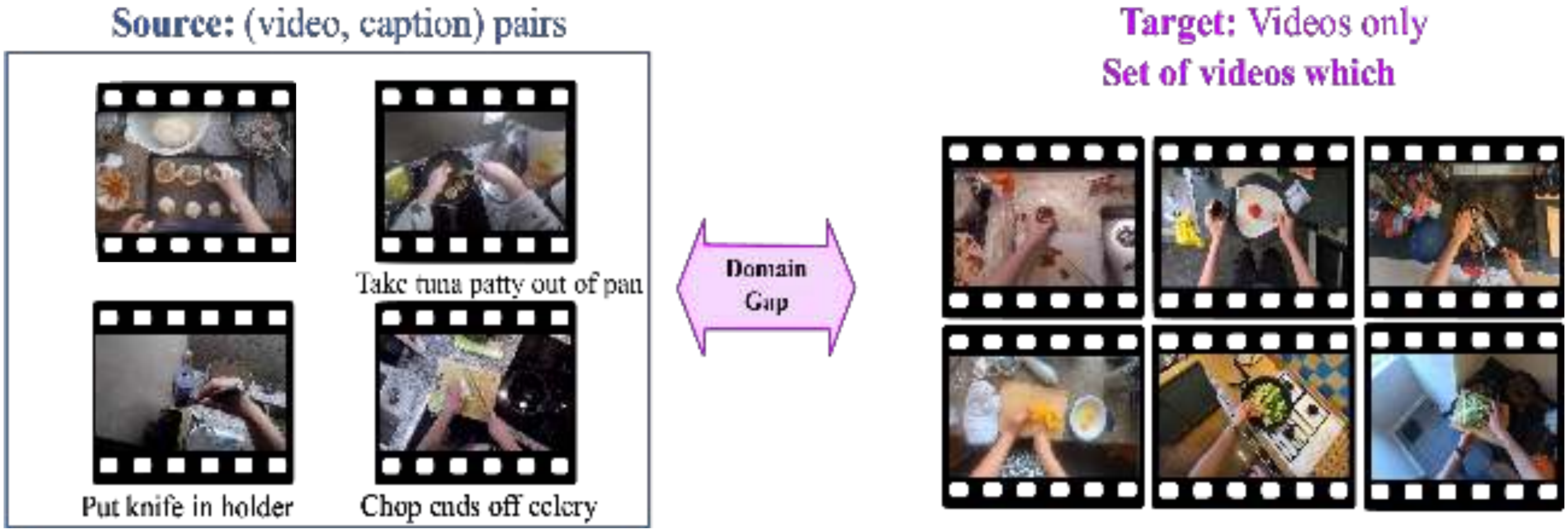
Metric	Method	$D2 \to D1$	$D3 \to D1$	$D1 \to D2$	$D3 \to D2$	$D1 \to D3$	$D2 \to D3$	mean
Verb	Source-only	28.3	27.0	32.1	45.8	27.6	42.1	33.8
	MM-SADA	25.4	33.2	44.1	49.3	35.6	43.7	38.5▲+4.70
Noun	Source-only	11.3	10.2	11.8	19.0	10.3	19.3	13.6
	MM-SADA	12.7	14.0	14.1	24.2	10.7	21.5	16.2▲+2.60

Table 4.8: Impact of multi-modal self-supervision (RGB, flow and audio) on the open-set domain adaption benchmarks. Both verb and noun classification improve with the self-supervised loss.

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

Multi-modal UDA for Retrieval

with: Jonathan Munro Diane Larlus
Michael Wray Gabriela Csurka



J Munro, et al (2021). Domain Adaptation in Multi-View Embedding for Cross-Modal Video Retrieval. *ArXiv*

Opportunities in Egocentric Vision

Tasks are harder

Detection, Recognition, 3D Mapping, Tracking, VOS …

Solutions prove more rewarding

Weak supervision, Domain Adaptation, 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

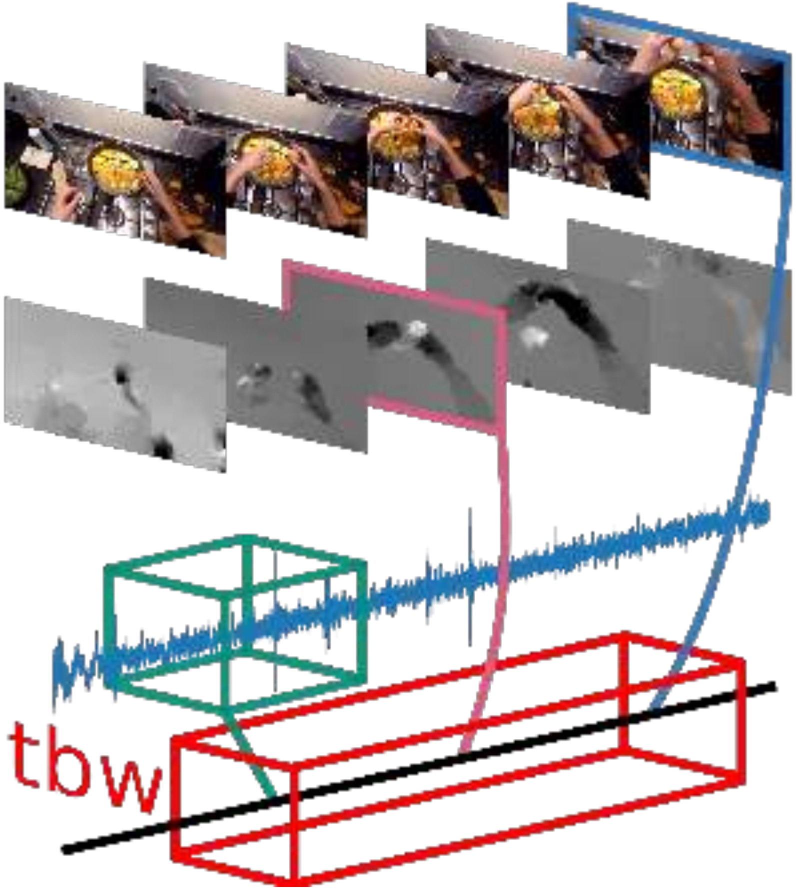
Multi-modal learning…

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

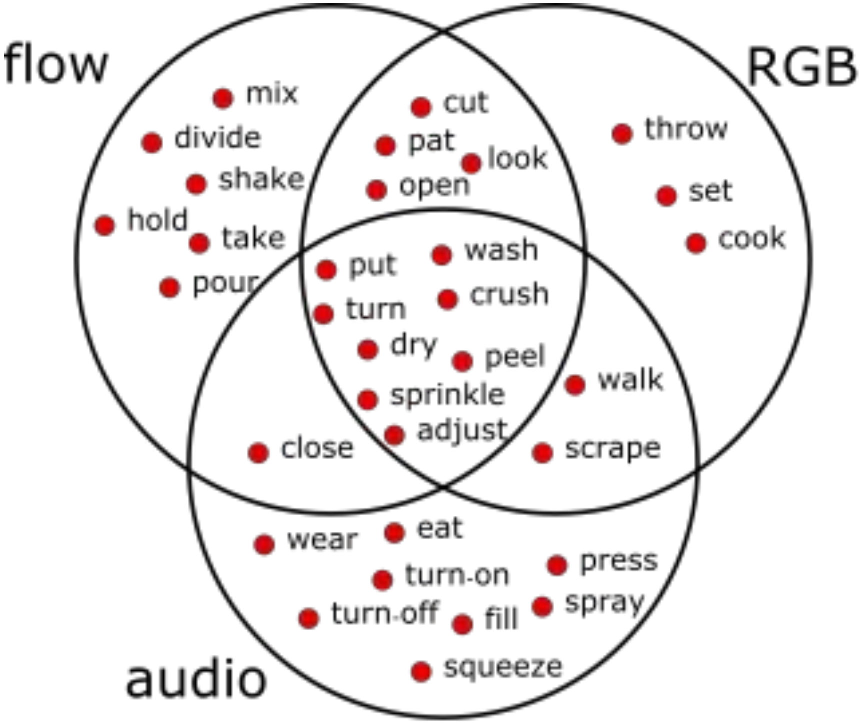
Audio-Visual Temporal Binding

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



Audio-Visual Temporal Binding

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



E Kazakos, A Nagrani, A Zisserman, D Damen (2019). EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. ICCV

Audio-Visual Temporal Binding

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

		Top-1 Accuracy			Top-5 Accuracy			Avg Class Precision			Avg Class Recall		
		VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION	VERB	NOUN	ACTION
S1	RGB	45.68	36.80	19.86	85.56	64.19	41.89	61.64	34.32	09.96	23.81	31.62	08.81
	Flow	55.65	31.17	20.10	85.99	56.00	39.30	48.83	26.84	09.02	27.58	24.15	07.89
	Audio	43.56	22.35	14.21	79.66	43.68	27.82	32.28	19.10	07.27	25.33	18.16	06.17
	TBN (RGB+Flow)	60.87	42.93	30.31	89.68	68.63	51.81	**61.93**	39.68	18.11	39.99	38.37	16.90
	TBN (All)	**64.75**	**46.03**	**34.80**	**90.70**	**71.34**	**56.65**	55.67	**43.65**	**22.07**	**45.55**	**42.30**	**21.31**
S2	RGB	34.89	21.82	10.11	74.56	45.34	25.33	19.48	14.67	04.77	11.22	17.24	05.67
	Flow	48.21	22.98	14.48	77.85	45.55	29.33	23.00	13.29	05.63	19.61	16.09	07.61
	Audio	35.43	11.98	06.45	69.20	29.49	16.18	22.46	09.41	04.59	18.02	09.79	04.19
	TBN (RGB+Flow)	49.61	25.68	16.80	78.36	50.94	32.61	30.54	20.56	09.89	21.90	20.62	11.21
	TBN (All)	**52.69**	**27.86**	**19.06**	**79.93**	**53.78**	**36.54**	**31.44**	**21.48**	**12.00**	**28.21**	**23.53**	**12.69**

E Kazakos, A Nagrani, A Zisserman, D Damen (2019). EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition. ICCV

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

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

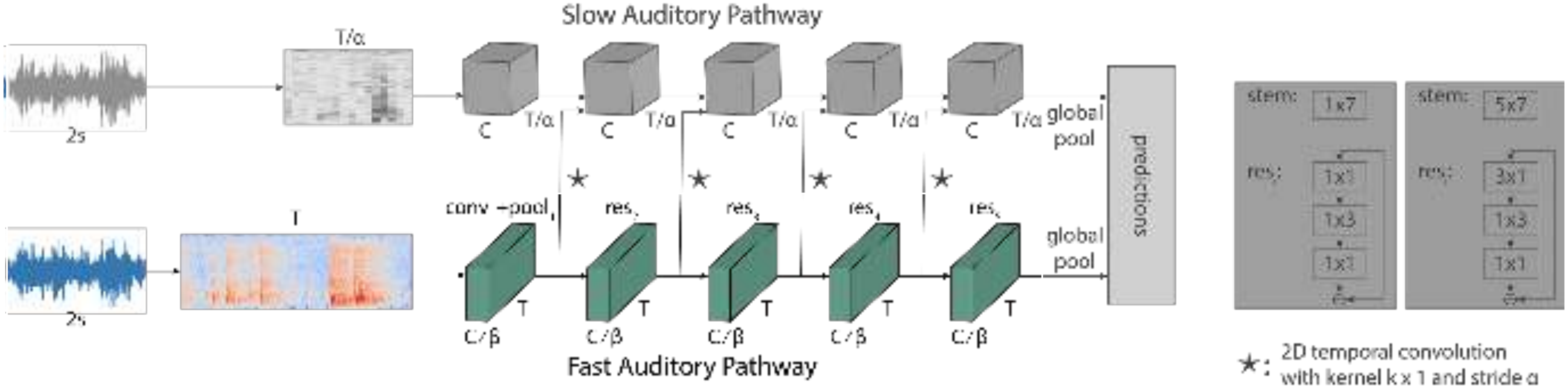
with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

Auditory Slow-Fast

Outstanding Paper Award – ICASSP 2021

Audio Slow-Fast

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

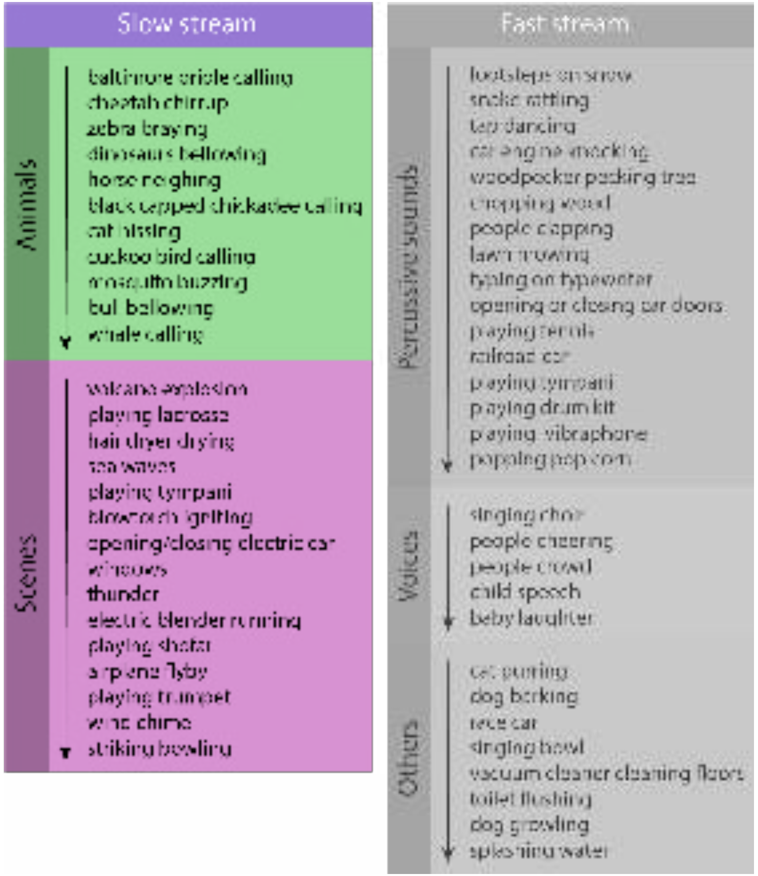


- Slow has low temporal precision and large amount of channels
- Fast has fewer channels but high temporal resolution
- Multi-level lateral connections
- Separable convolutions

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

Audio Slow-Fast

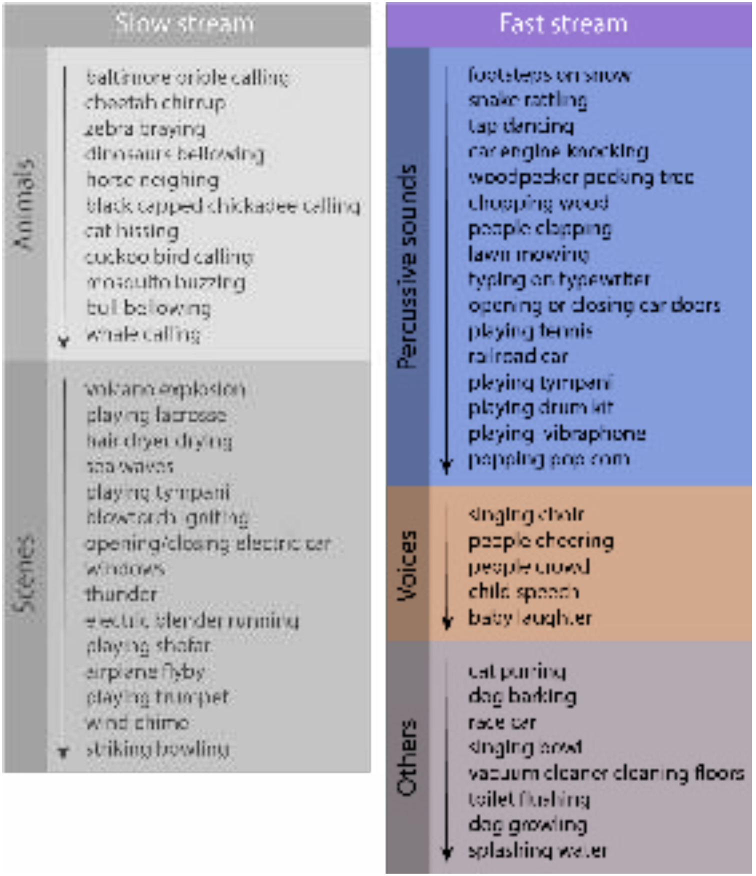
with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



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



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 [6, 12], 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

Tasks are harder

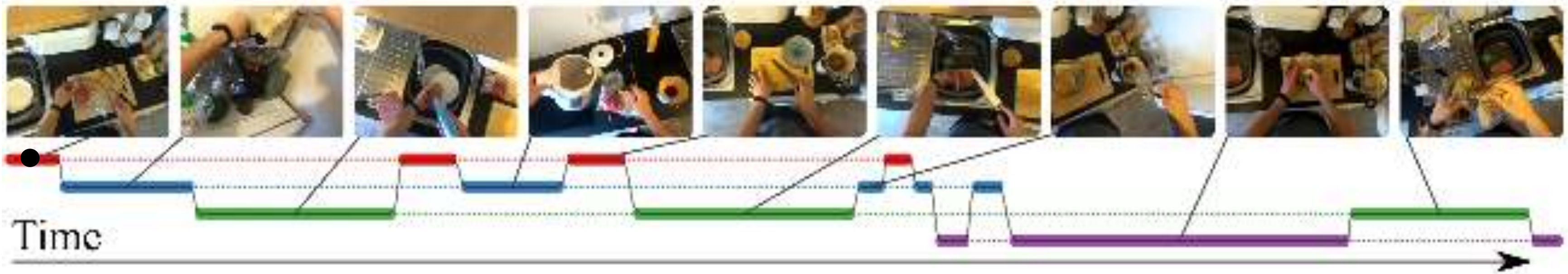
Detection, Recognition, 3D Mapping, Tracking, VOS …

Solutions prove more rewarding

Weak supervision, Domain Adaptation, Audio-Visual, long-term understanding

Goals…

with: Will Price
Carl Vondrick



W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick

W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick

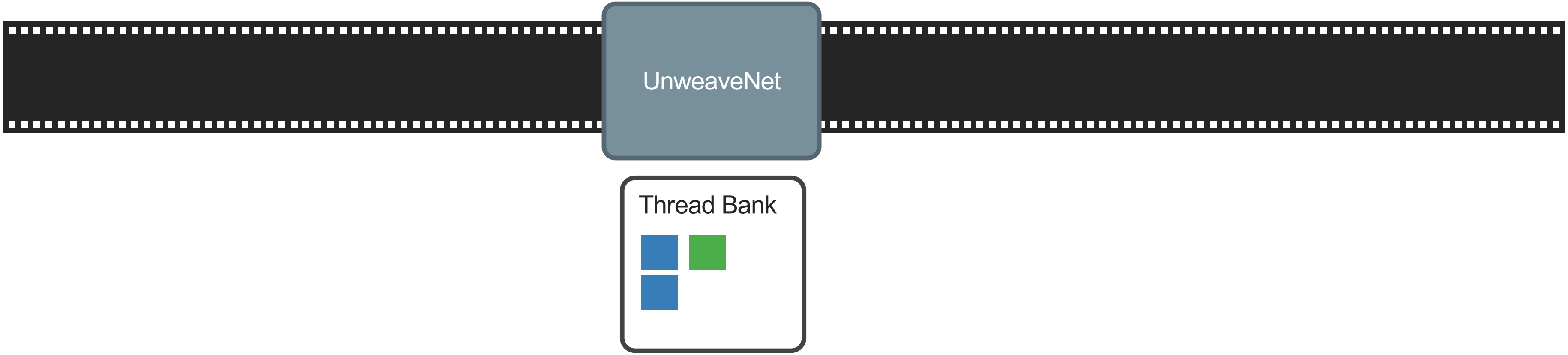
W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick

UnweaveNet

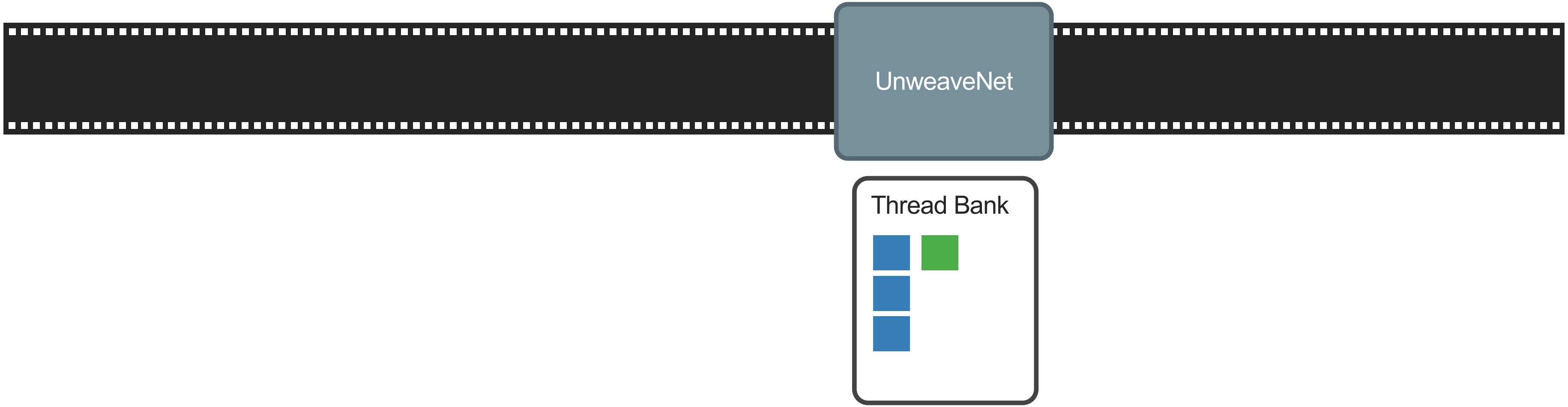
with: Will Price
Carl Vondrick



W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick



UnweaveNet

with: Will Price
Carl Vondrick

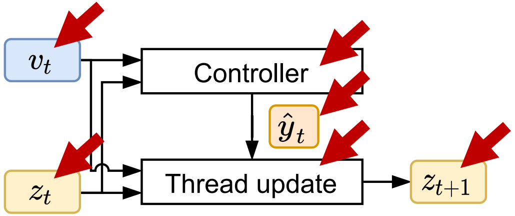
W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick

UnweaveNet

with: Will Price
Carl Vondrick

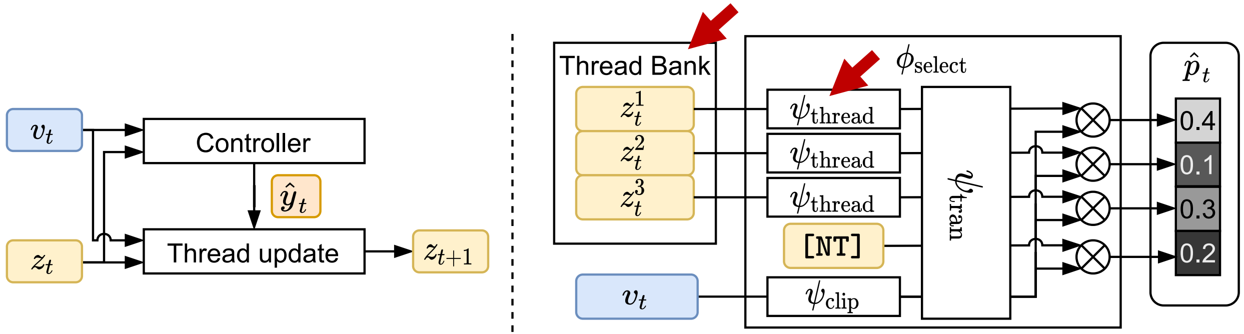


(a) UnweaveNet Overview

W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick



(a) UnweaveNet Overview

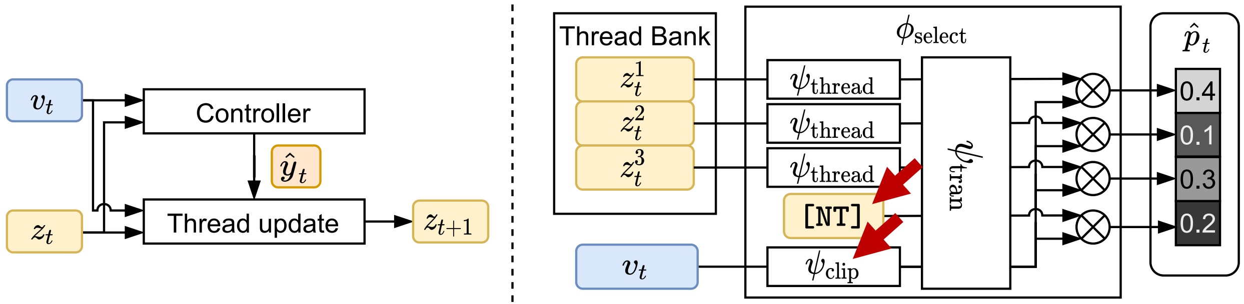
(b) Controller Architecture

Two learnt embeddings

$$\psi_{\text{thread}} : \mathbb{R}^D \rightarrow \mathbb{R}^E$$

UnweaveNet

with: Will Price
Carl Vondrick



(a) UnweaveNet Overview

(b) Controller Architecture

Two learnt embeddings

$$\psi_{\text{thread}} : \mathbb{R}^D \rightarrow \mathbb{R}^E$$

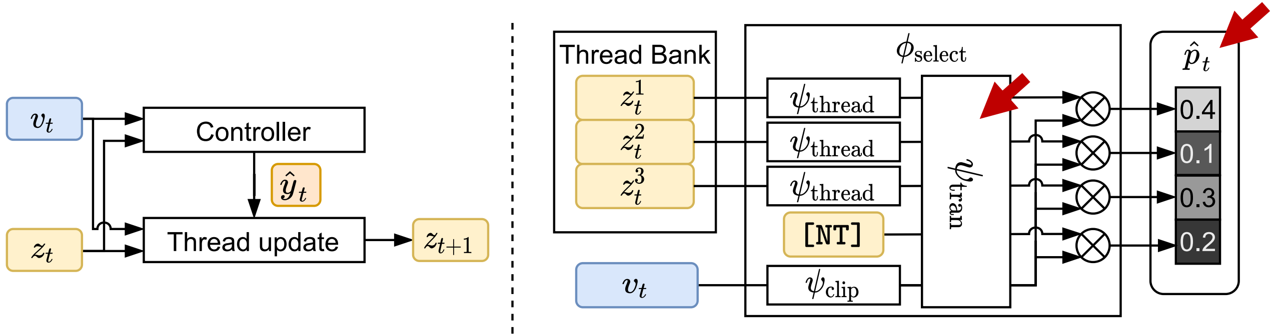
$$\psi_{\text{clip}} : \mathbb{R}^C \rightarrow \mathbb{R}^E$$

Learnt Encoding $\texttt{[NT]} \in \mathbb{R}^E$

W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick

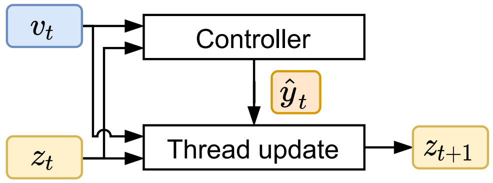


(a) UnweaveNet Overview (b) Controller Architecture

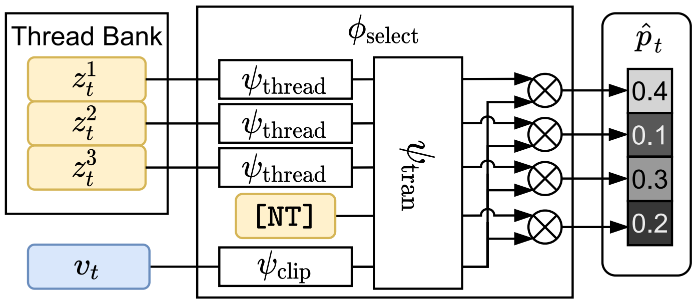
W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

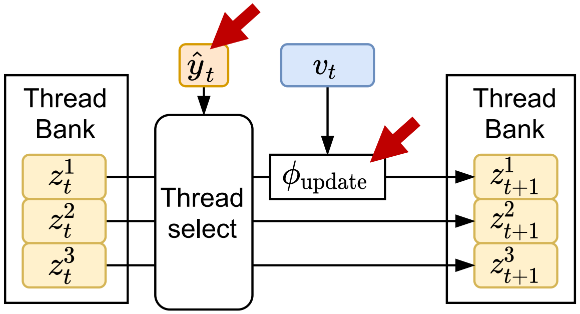
with: Will Price
Carl Vondrick



(a) UnweaveNet Overview



(b) Controller Architecture

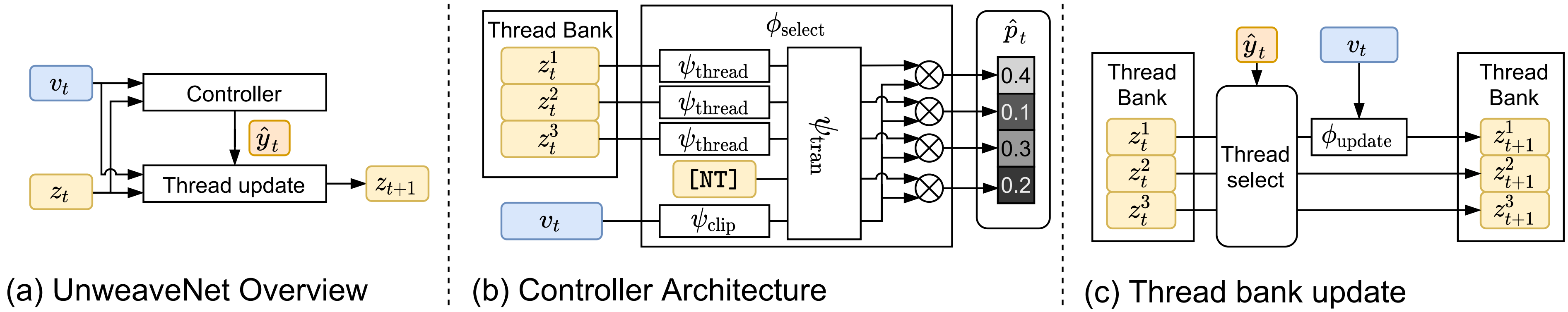


(c) Thread bank update

W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick

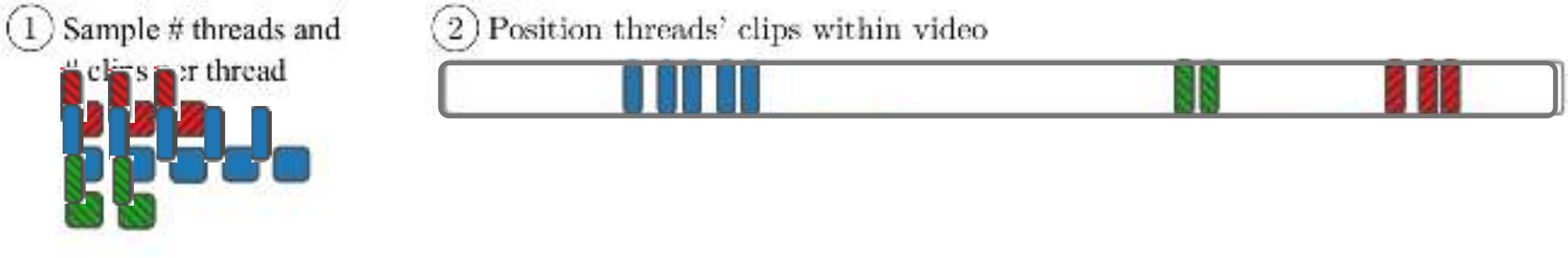


- Trained **end-to-end** including the backbone for clip features
- decisions made by ϕ_{select} are supervised using **teacher forcing**
  - at each time step, z_t is populated according to the ground-truth assignments $y_{1:t-1}$
  - A loss is then imposed on the output p_t given the correct decision y_t

UnweaveNet

with: Will Price
Carl Vondrick

- We propose self-supervised pretraining for UnweaveNet that samples threads from different parts of a long video and synthetically forms woven activity stories.

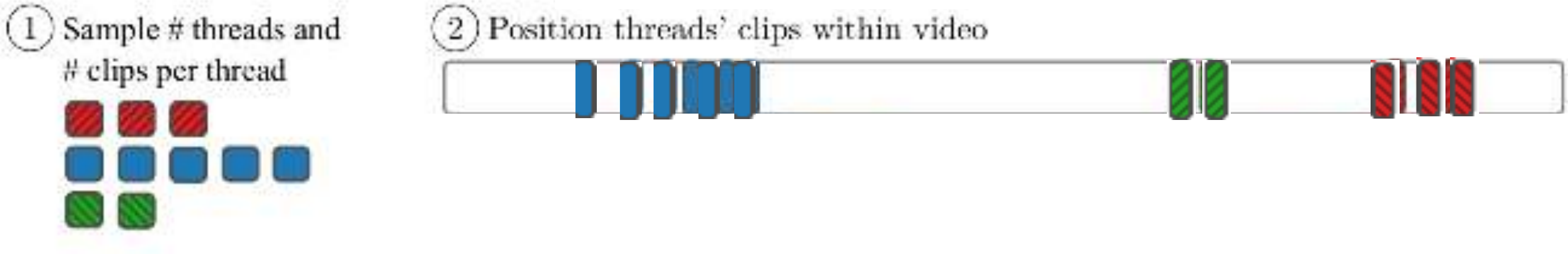


W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick

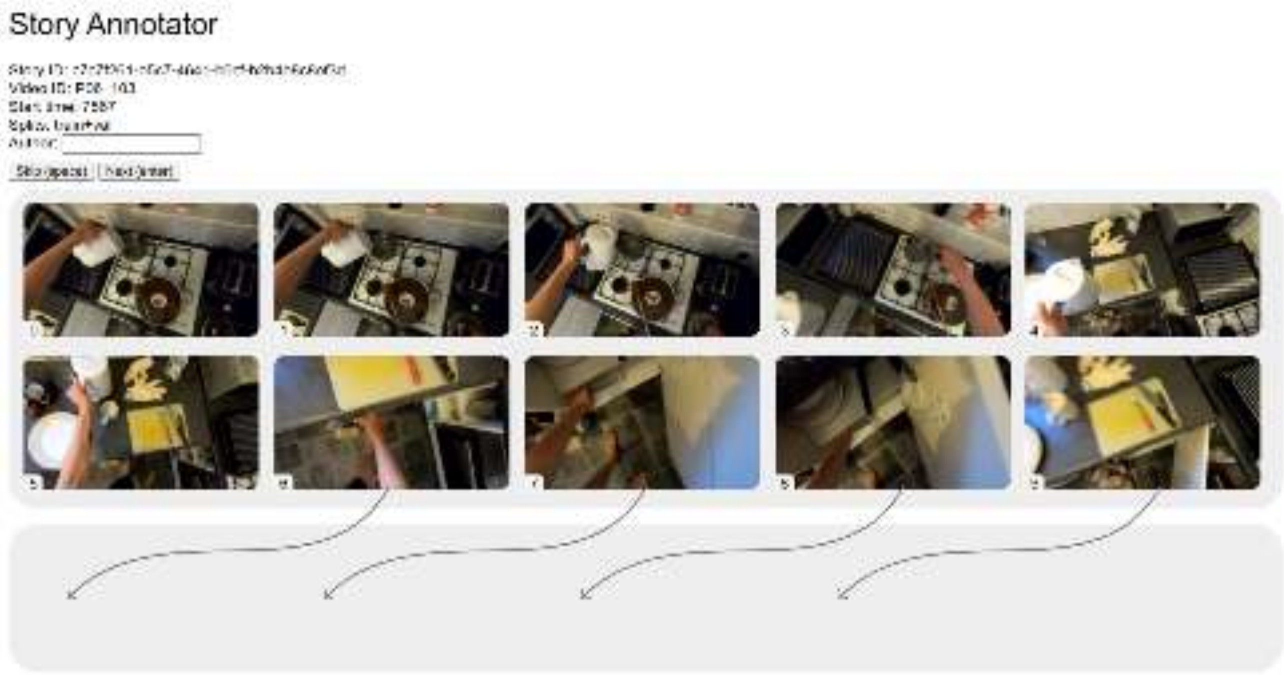
- We propose self-supervised pretraining for UnweaveNet that samples threads from different parts of a long video and synthetically forms woven activity stories.



UnweaveNet

with: Will Price
Carl Vondrick

- Labelled Sequences



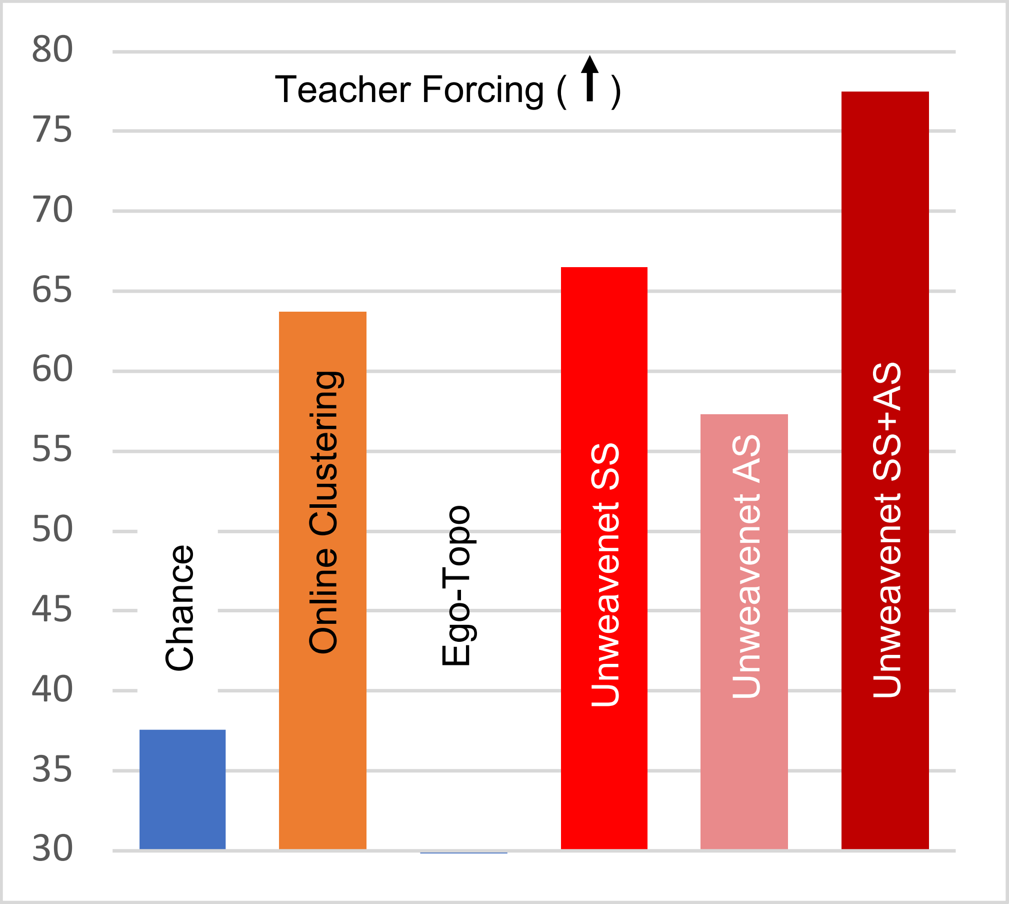
	# Threads		
Split	1	2	3
Train	718	201	32
Val	211	94	46
Test	50	50	50
Total	979	345	128

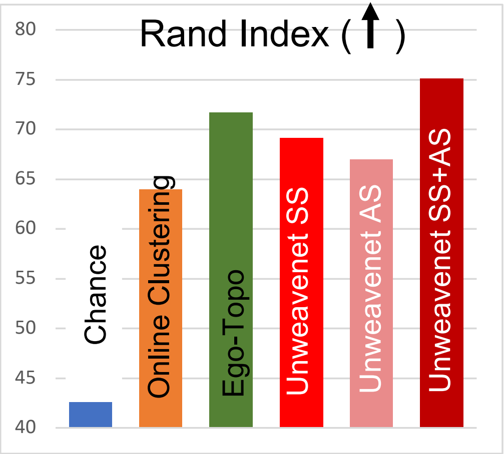
Table 1. EPIC-KITCHENS activity-story dataset by # of threads.

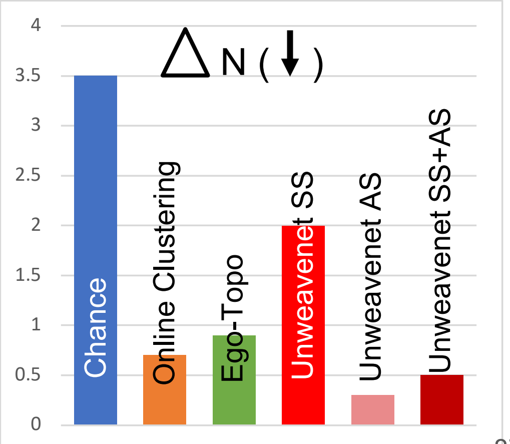
W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick







W Price, C Vondrick, D Damen (2022). UnweaveNet: Unweaving Activity Stories. CVPR

UnweaveNet

with: Will Price
Carl Vondrick

Clip index

UnweaveNet predicted thread

Ground truth thread

UnweaveNet's Decision scores

UnweaveNet

with: Will Price
Carl Vondrick

Opportunities in Egocentric Vision

Tasks are harder

Detection, Recognition, 3D Mapping, Tracking, VOS …

Solutions prove more rewarding

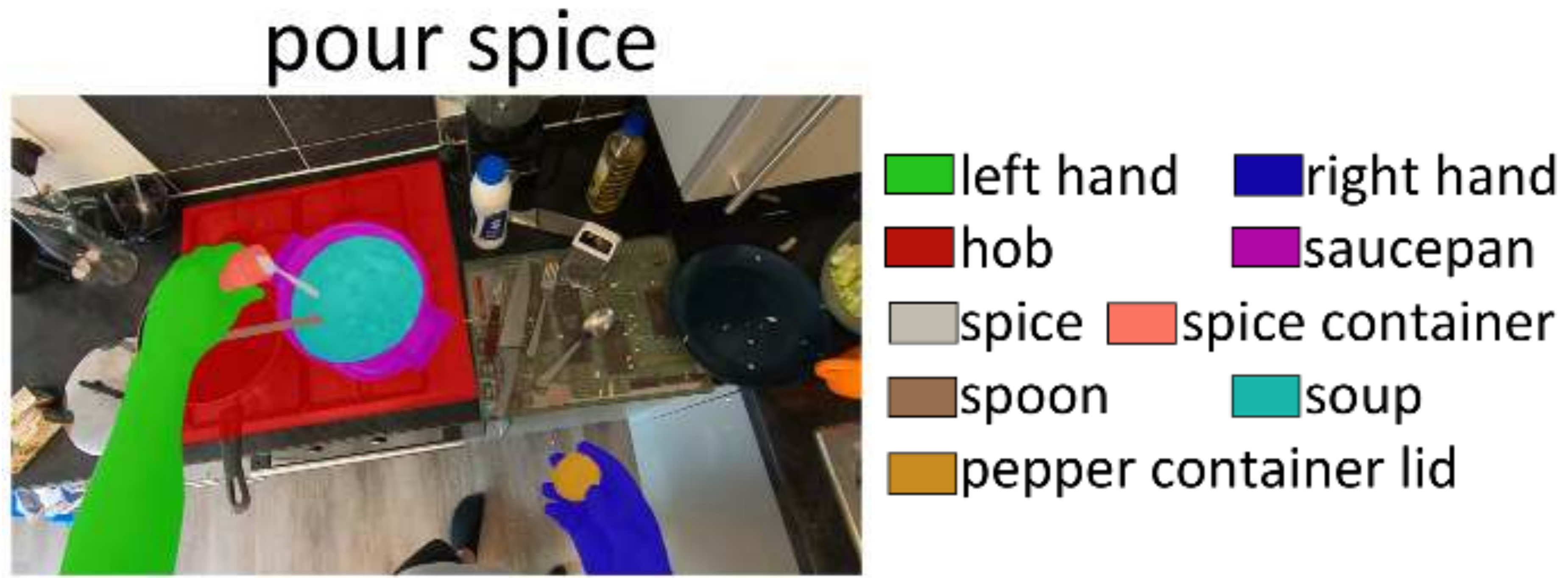
Weak supervision, Domain Adaptation, Audio-Visual, long-term understanding

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 is….

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



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

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

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

pour spice

left hand | right hand
hob | saucepan
spice | spice container
spoon | soup
pepper container lid

saucepan → pan → cookware

spoon → spoon → cutlery

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

VISOR is….

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

action

pour spice

left hand
right hand
hob
saucepan
spice
spice container
spoon
soup
pepper container lid

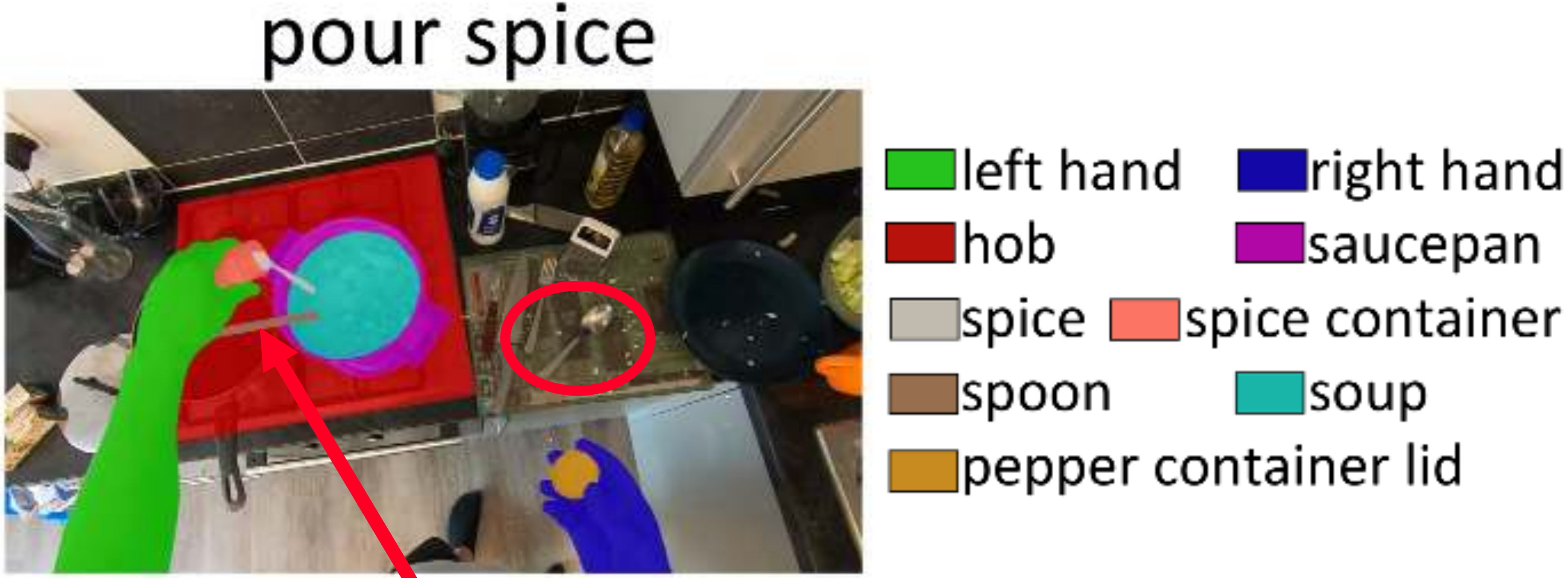
in-contact (spice container)

in-contact (container lid)

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

VISOR is….

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

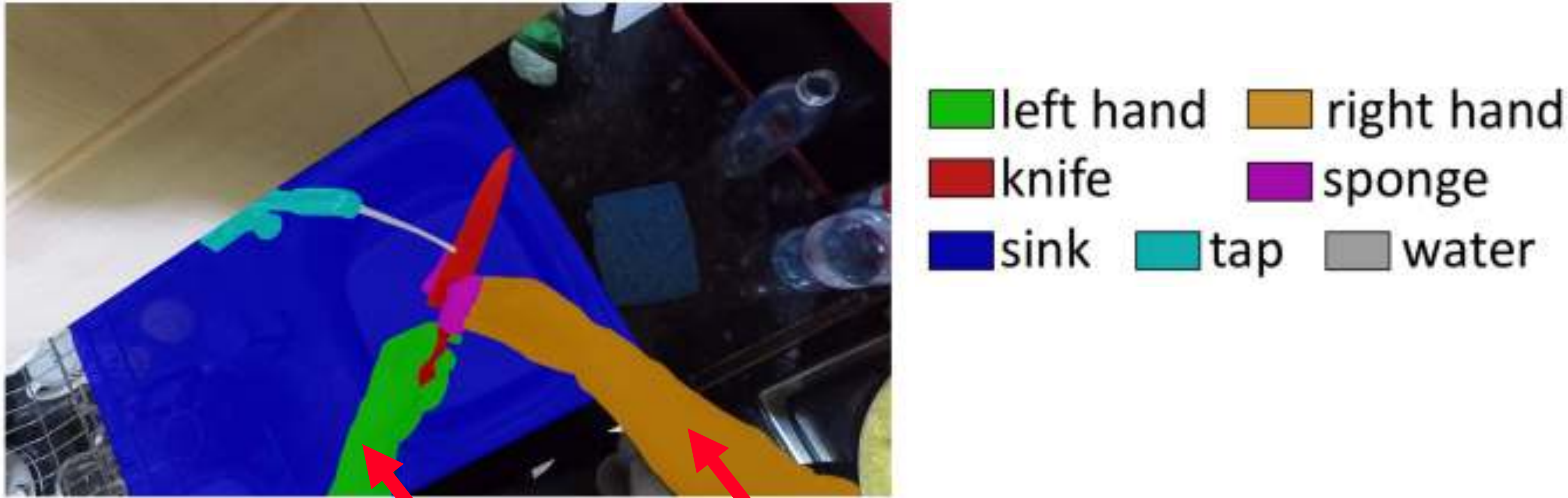


spoon (non-exhaustive)

VISOR is….

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

wash knife



in-contact (knife)

in-contact (sponge)

Comparative Stats

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

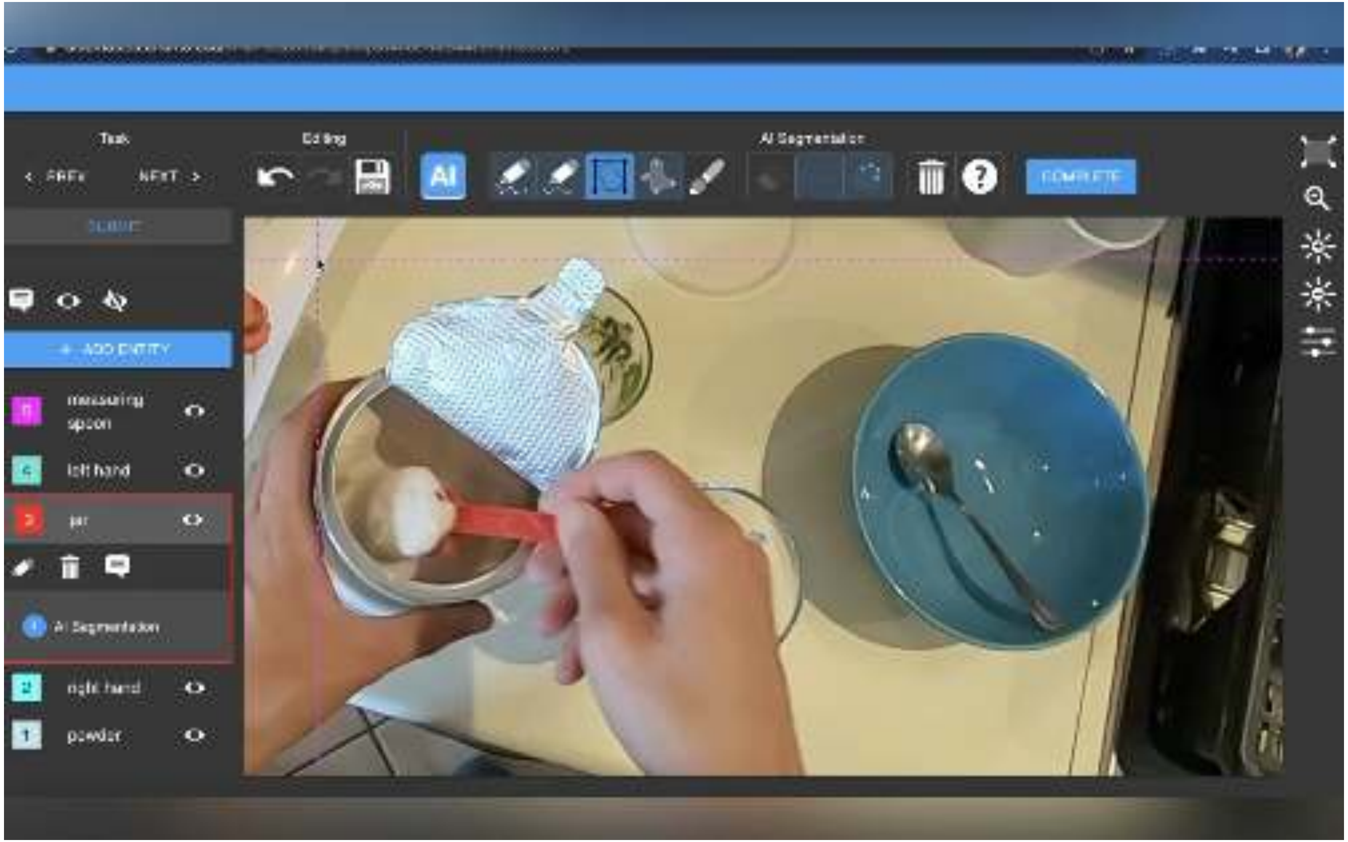
	Basic Statistics		Pixel-Level Annotations	Action Annotations		
Dataset	Total Mins	Avg Seq Ln	Total Masks	Actions	#Action Classes	#Entity Classes
EgoHand [3]	72	-	15.1K	-	-	2
DAVIS [6]	8	3s	32.0K	-	-	-
YTVOS [43]	335	5s	197.2K	-	-	94
UVOv0.5 (Sparse) [41]	511	3s	*200.6K	10,213	300	-
VISOR (Ours)	**2,180**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

University of BRISTOL

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

TORAS annotation tool

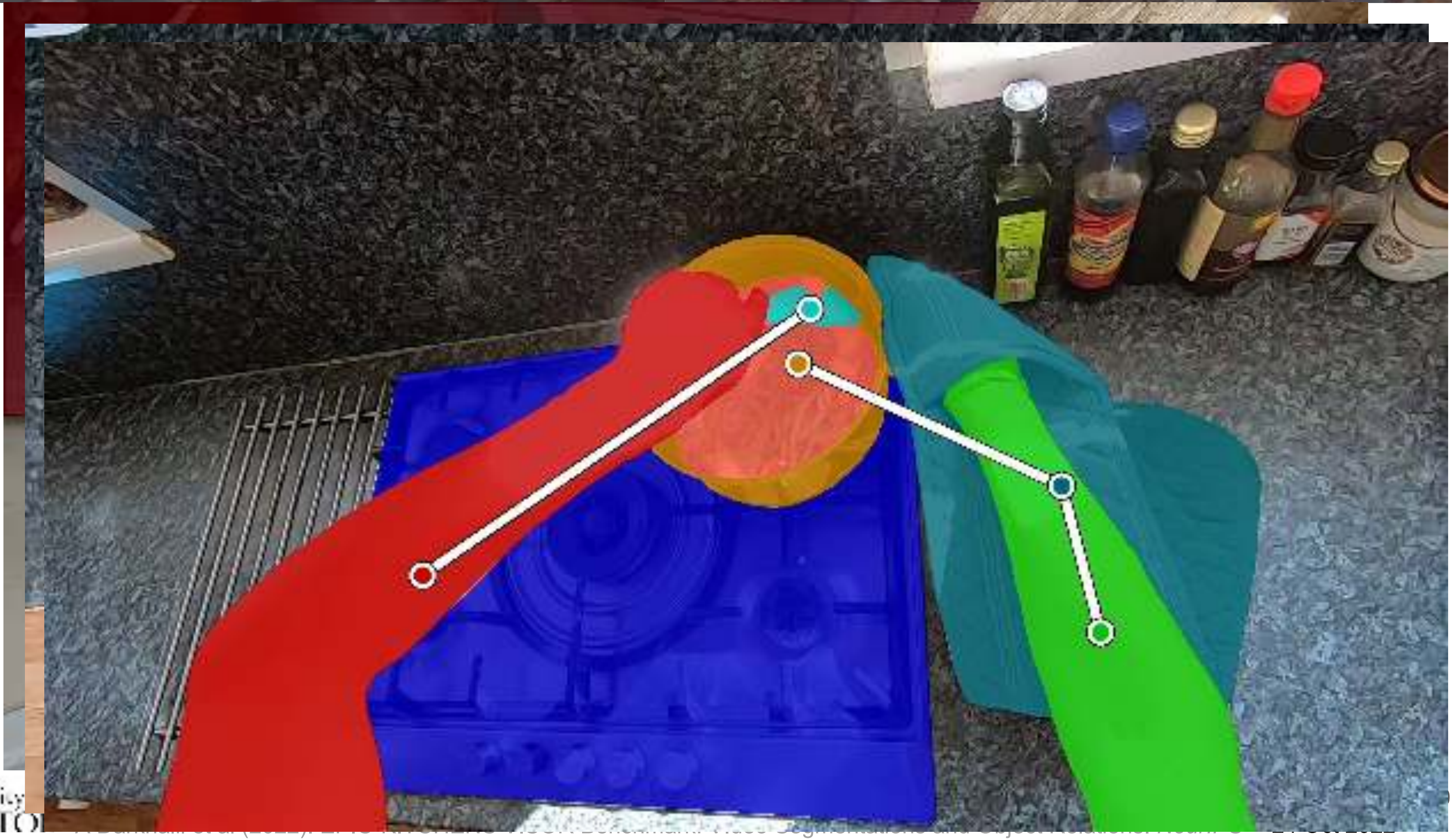
with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler, Dima Damen



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

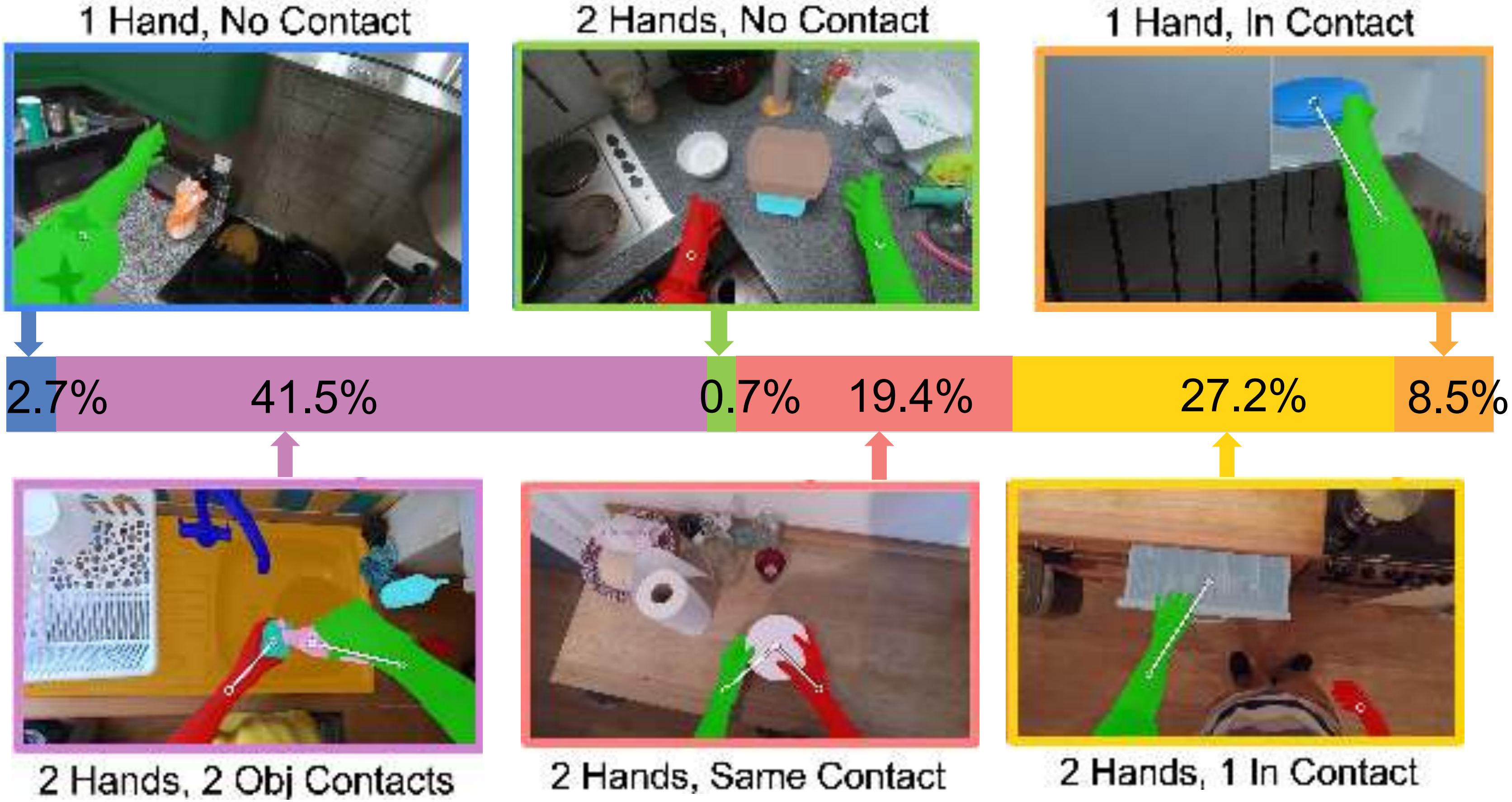
VISOR Relations

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



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

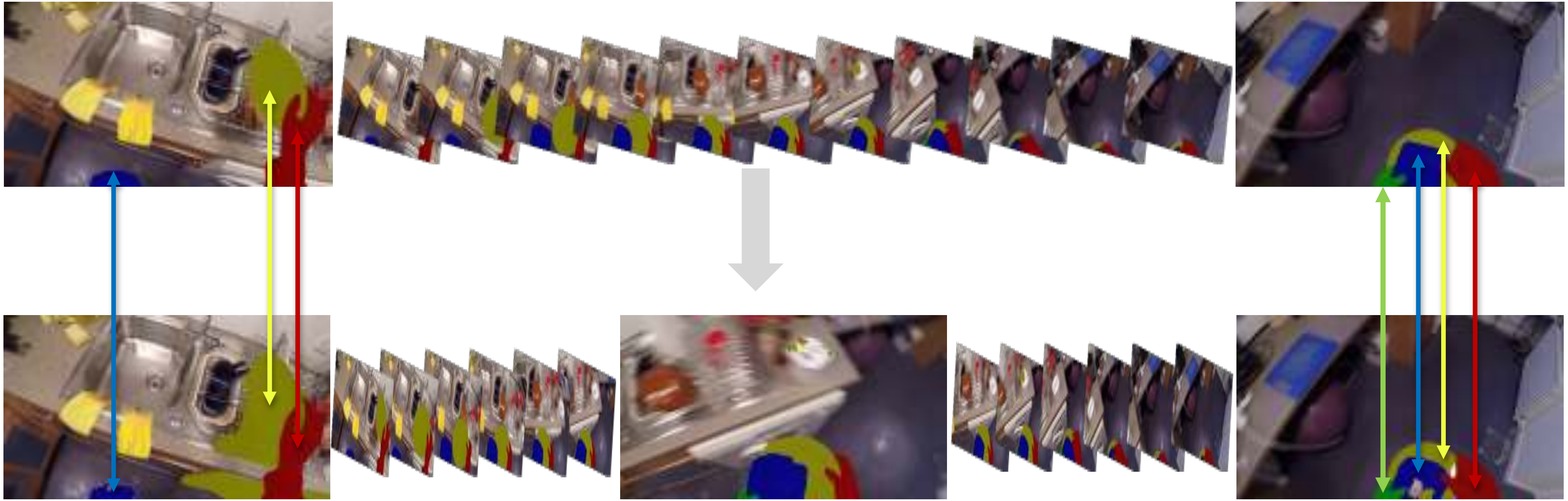
Dense Annotations

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

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

Dense Annotations

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



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

Dense Annotations

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

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*

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

Semi-supervised video object segmentation

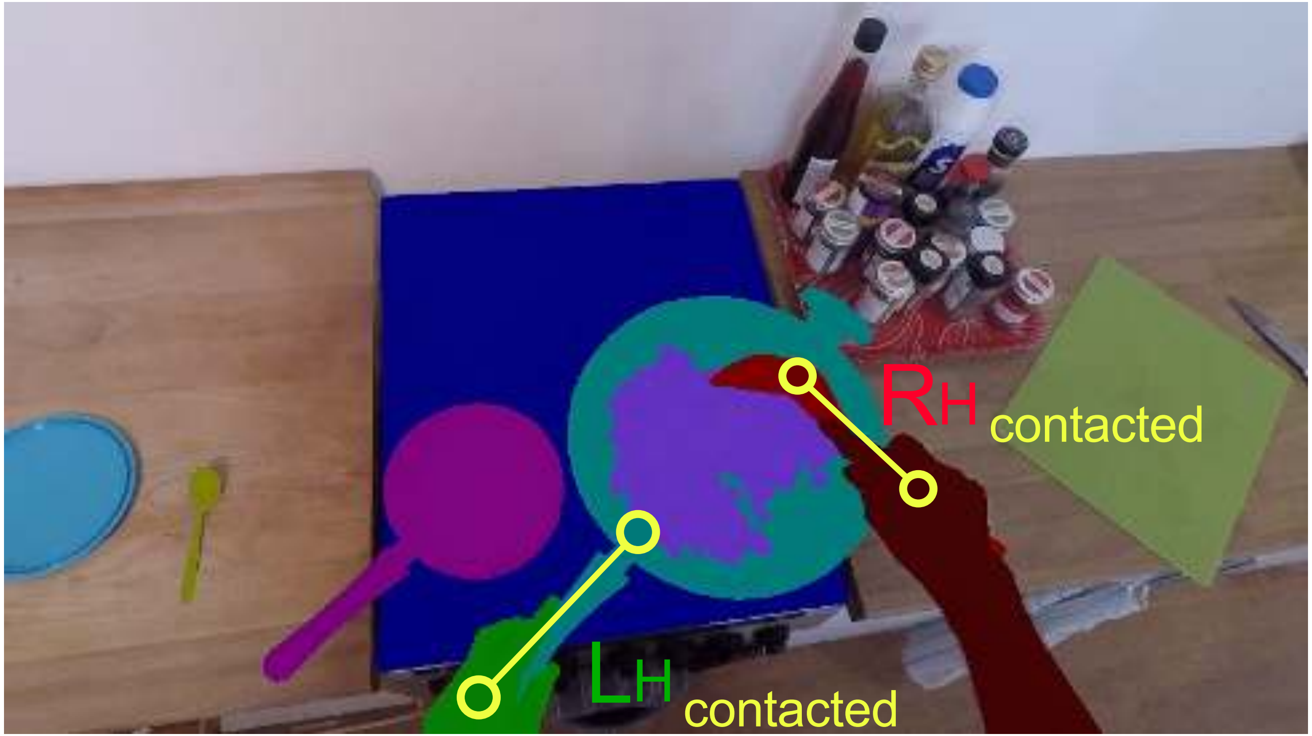
First frame

Propagate

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

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

Hand object segmentation



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

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

Where did *this* come from?

Query

21:45

Trace back

Evidence

03:41

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

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*

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*

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*

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*

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*

EPIC-KITTCHENS VISOR

Dataset Released: 16th of Aug 2022…
Code and Models: 27th of Sep 2022…

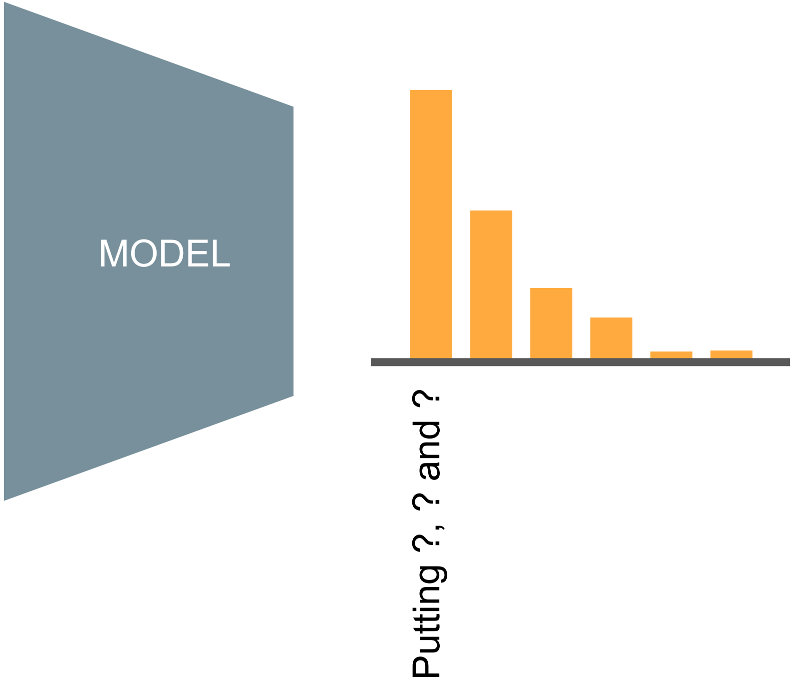
http://epic-kitchens.github.io/VISORFurther

And…

Explainable?

Frame Attributions in Video Models

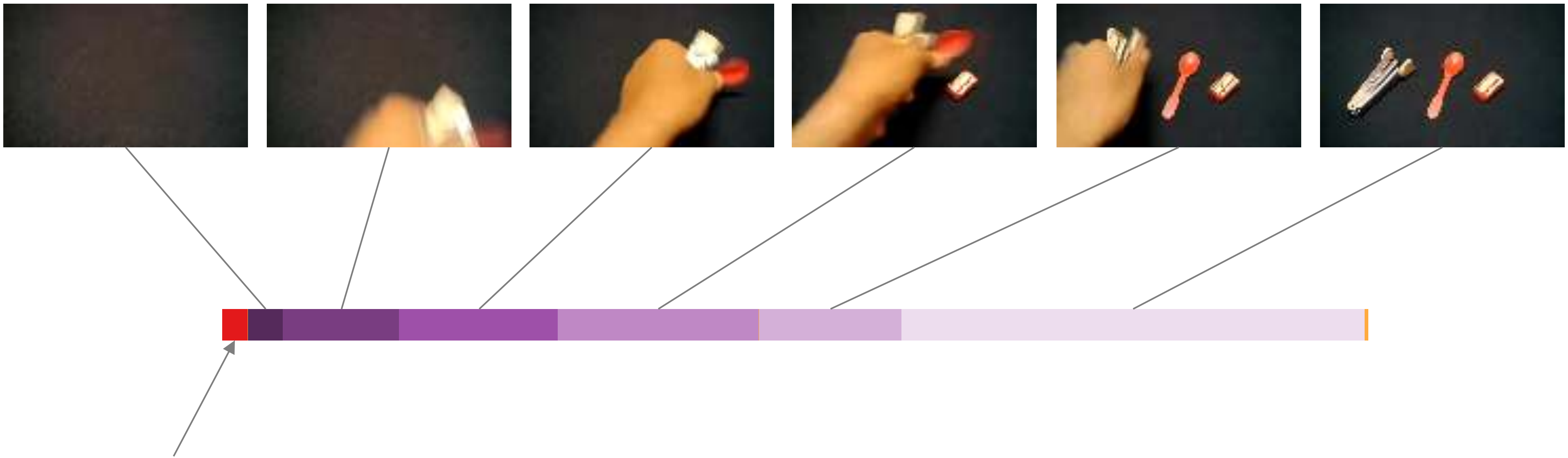
with: Will Price



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

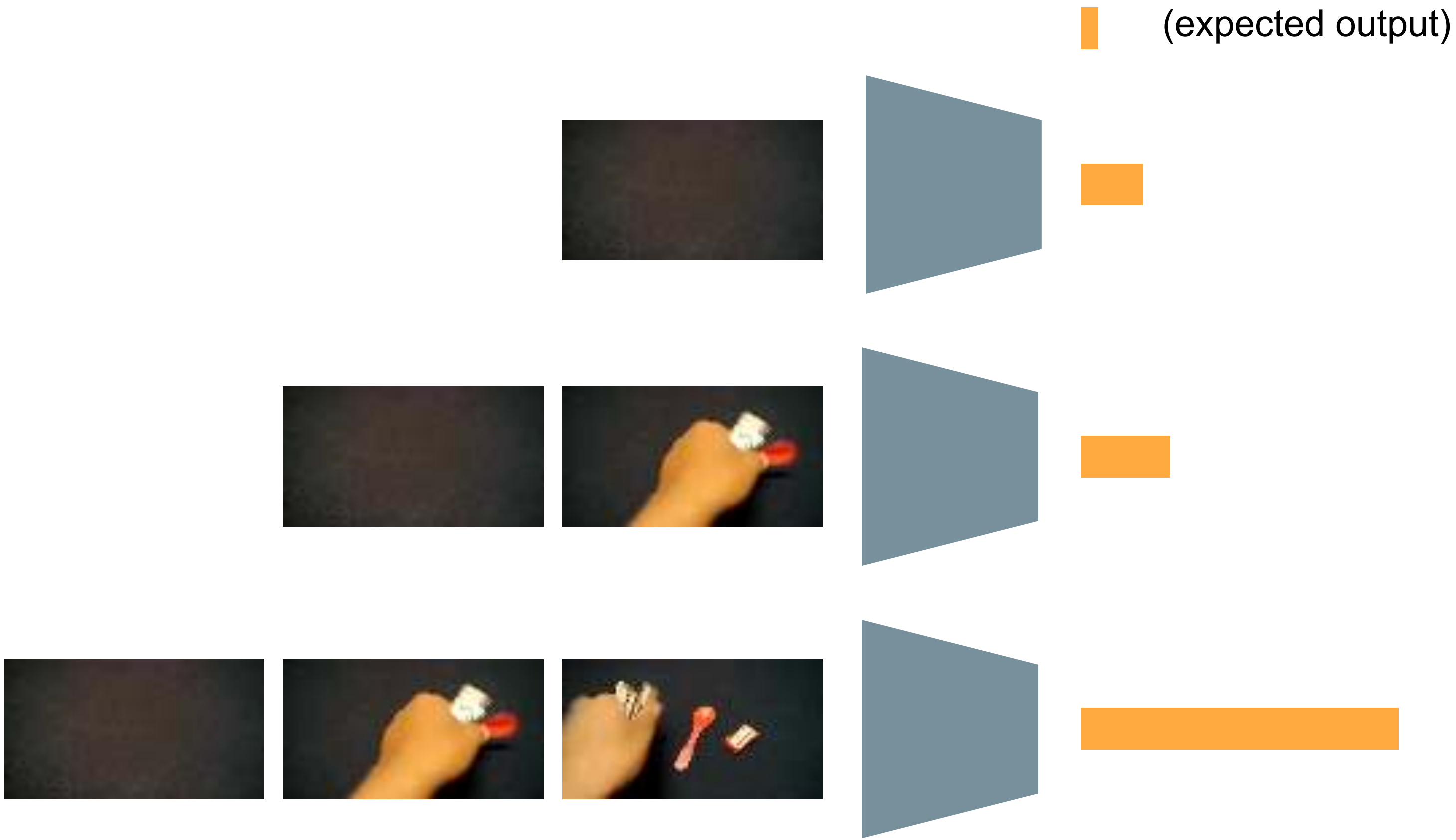


Expected output
(Prior probability for
classification model)

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

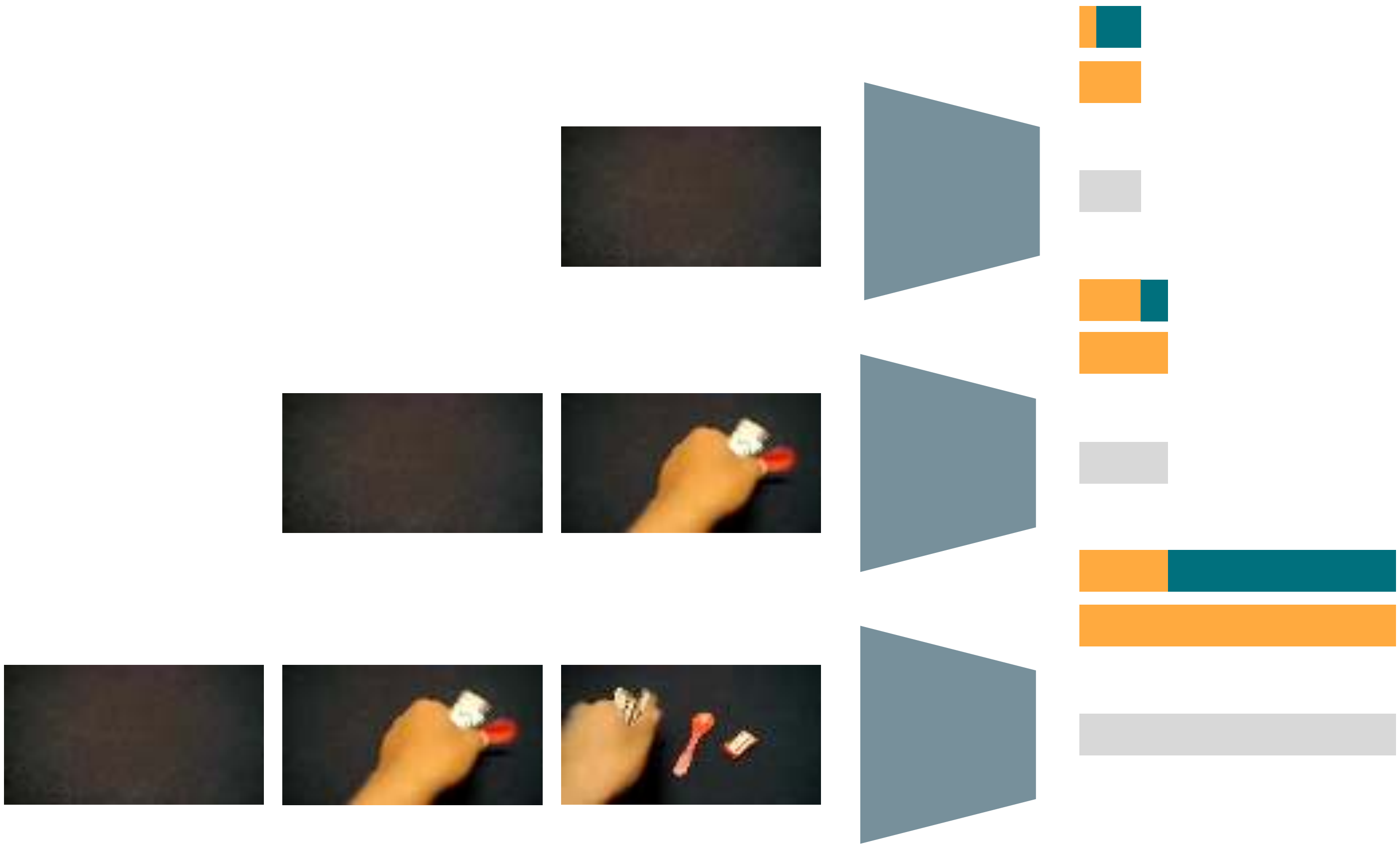
Frame Attributions in Video Models

with: Will Price



Frame Attributions in Video Models

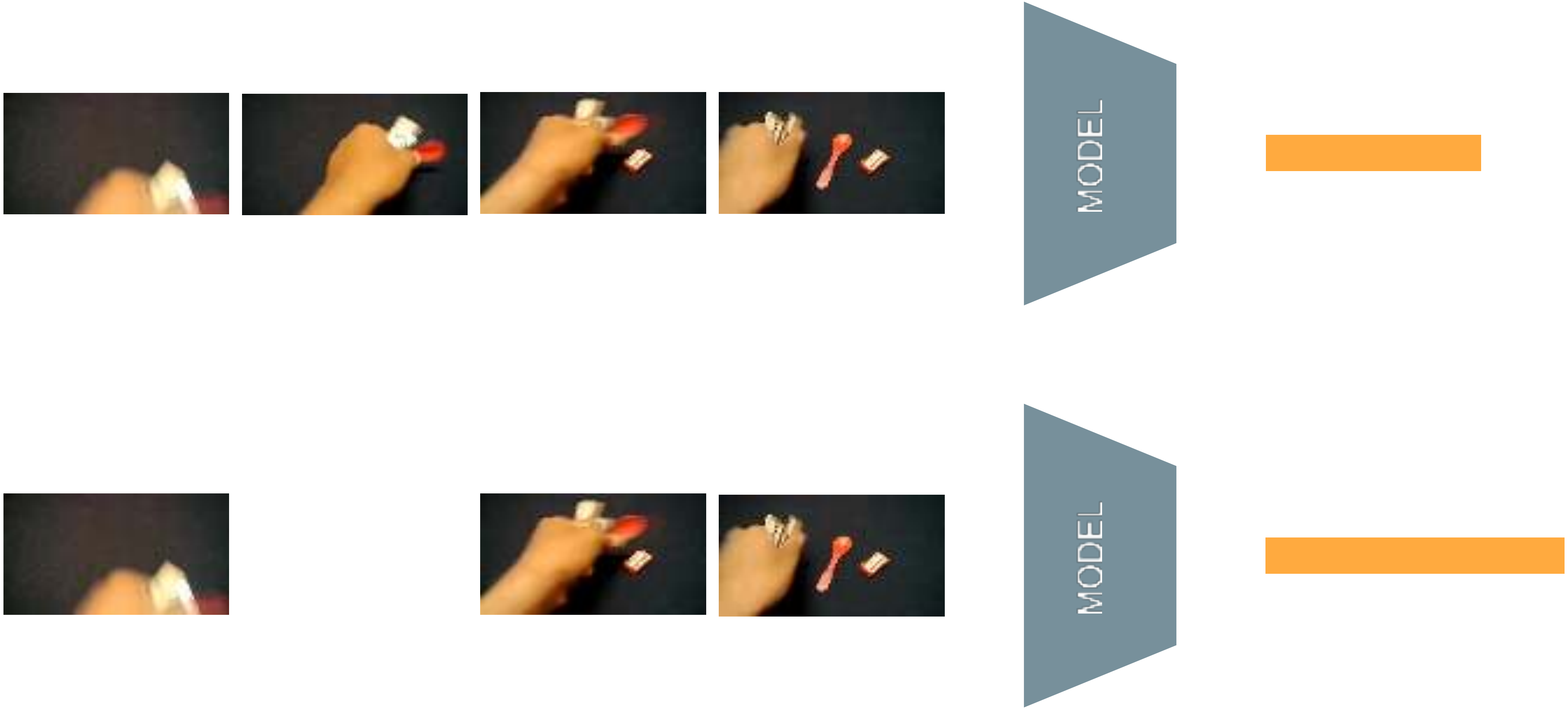
with: Will Price



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

Frame Attributions in Video Models

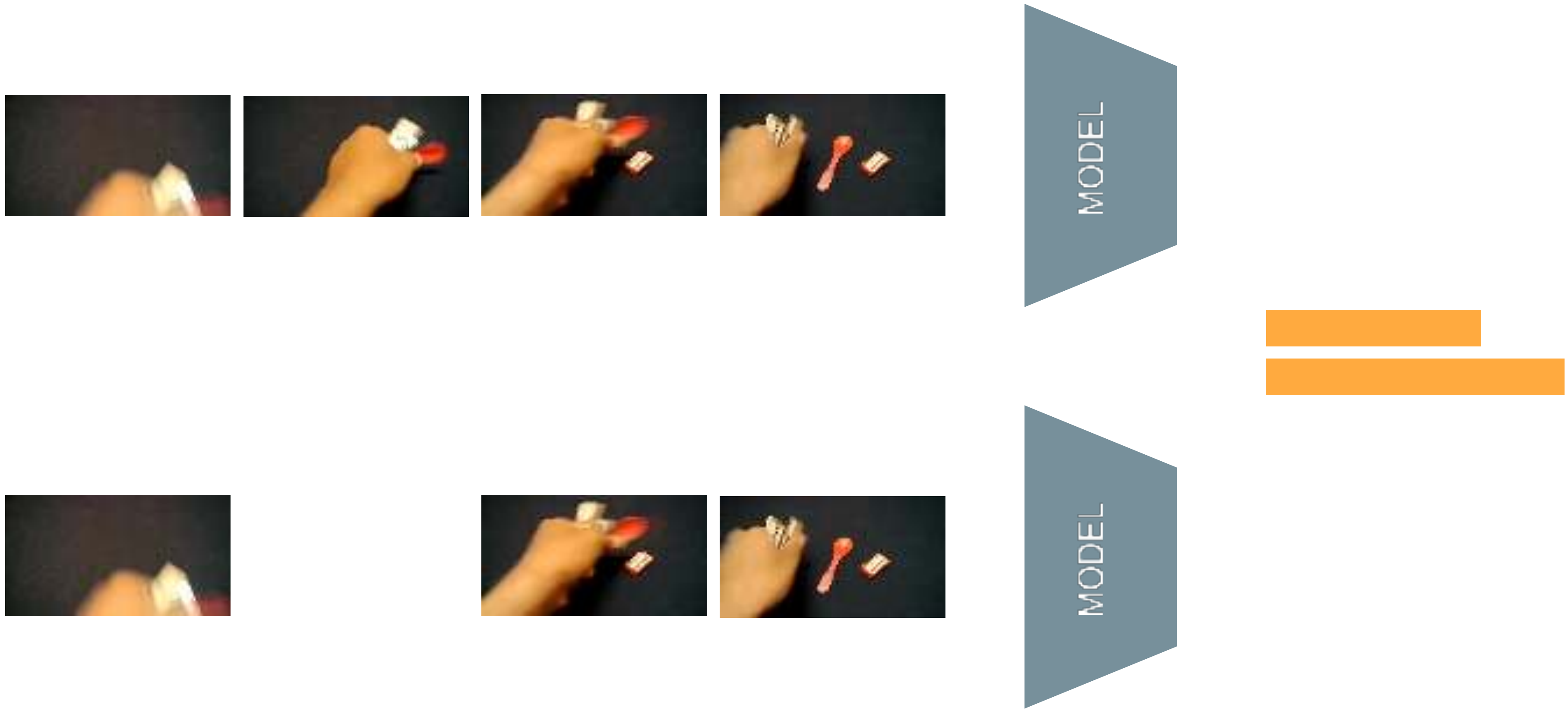
with: Will Price



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

Frame Attributions in Video Models

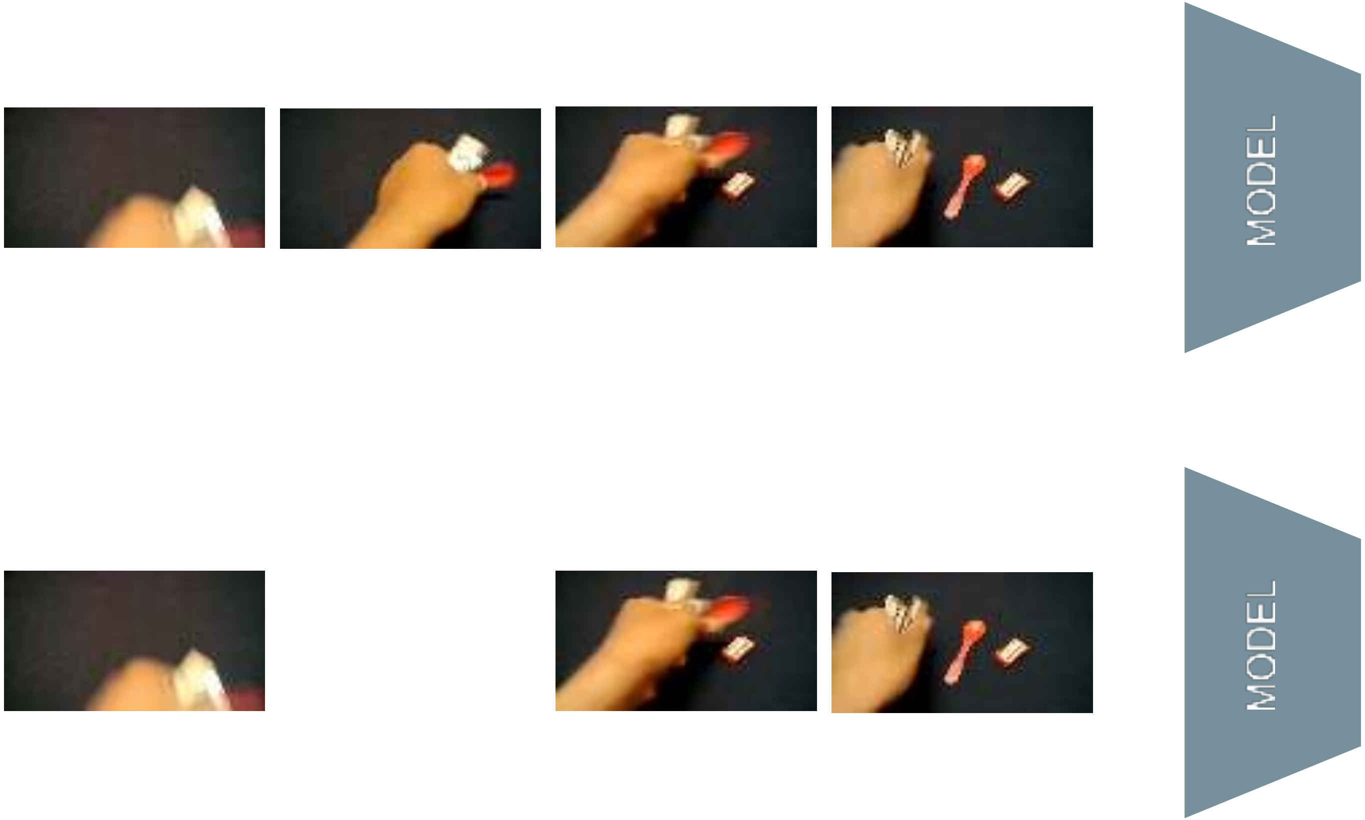
with: Will Price



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

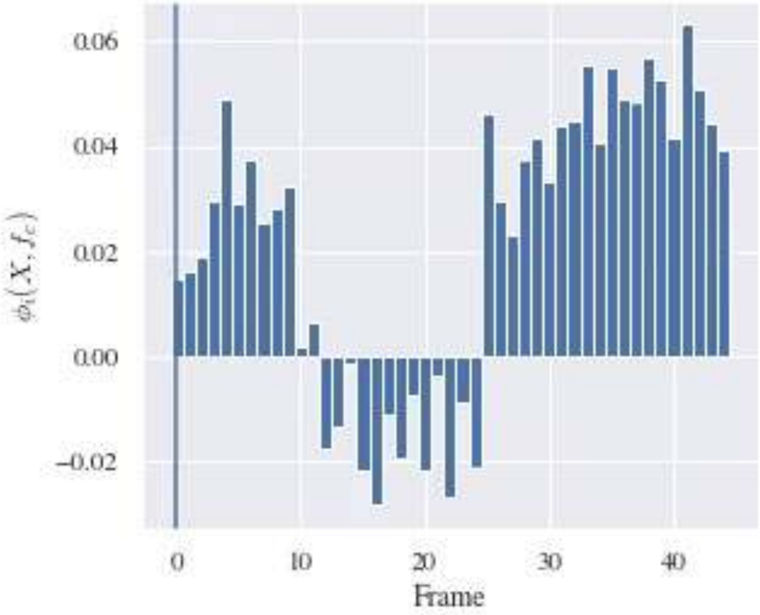


$\Delta_3(\{1,2,4,5\}) = -.2$

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

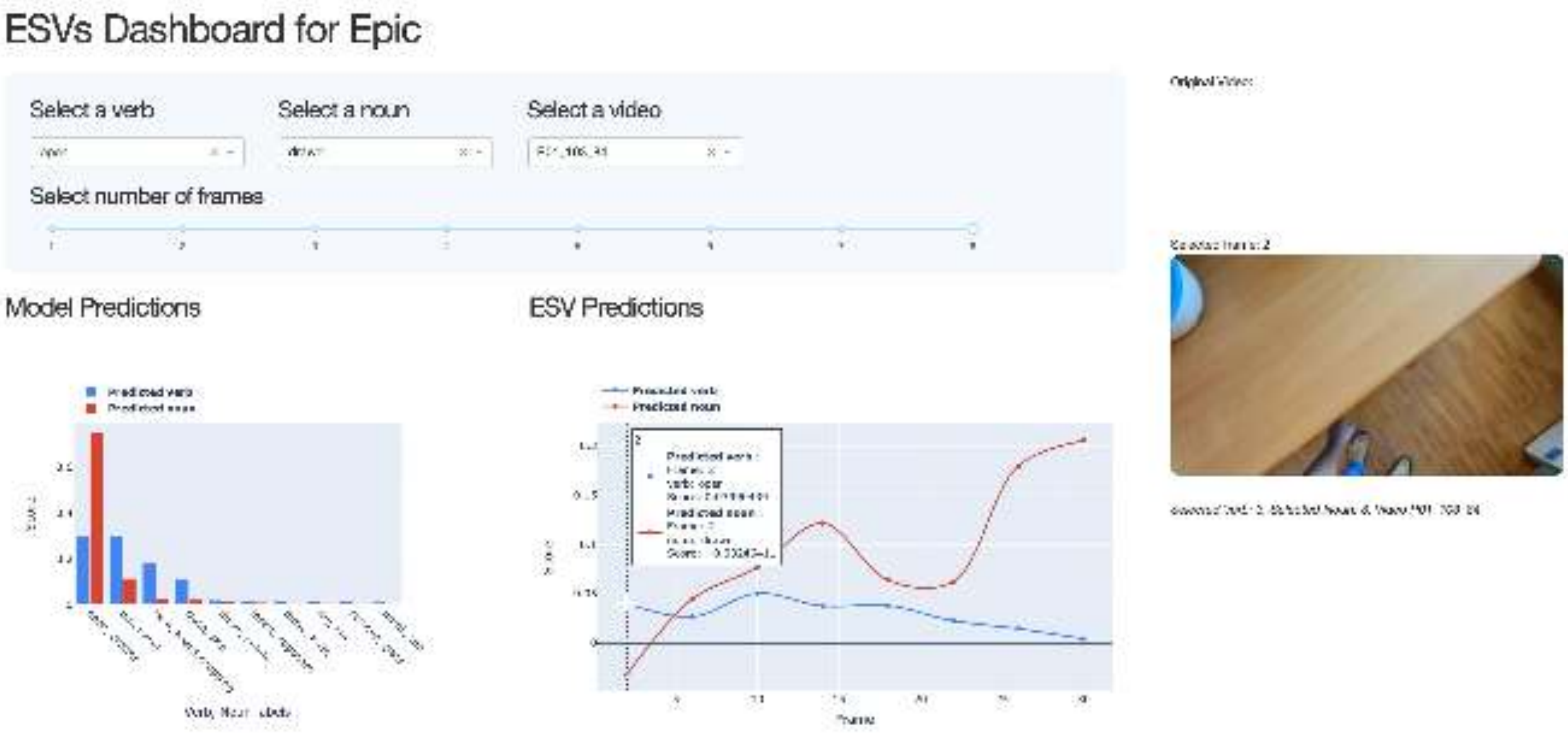


Showing that something is empty

Dashboard

Frame Attributions in Video Models

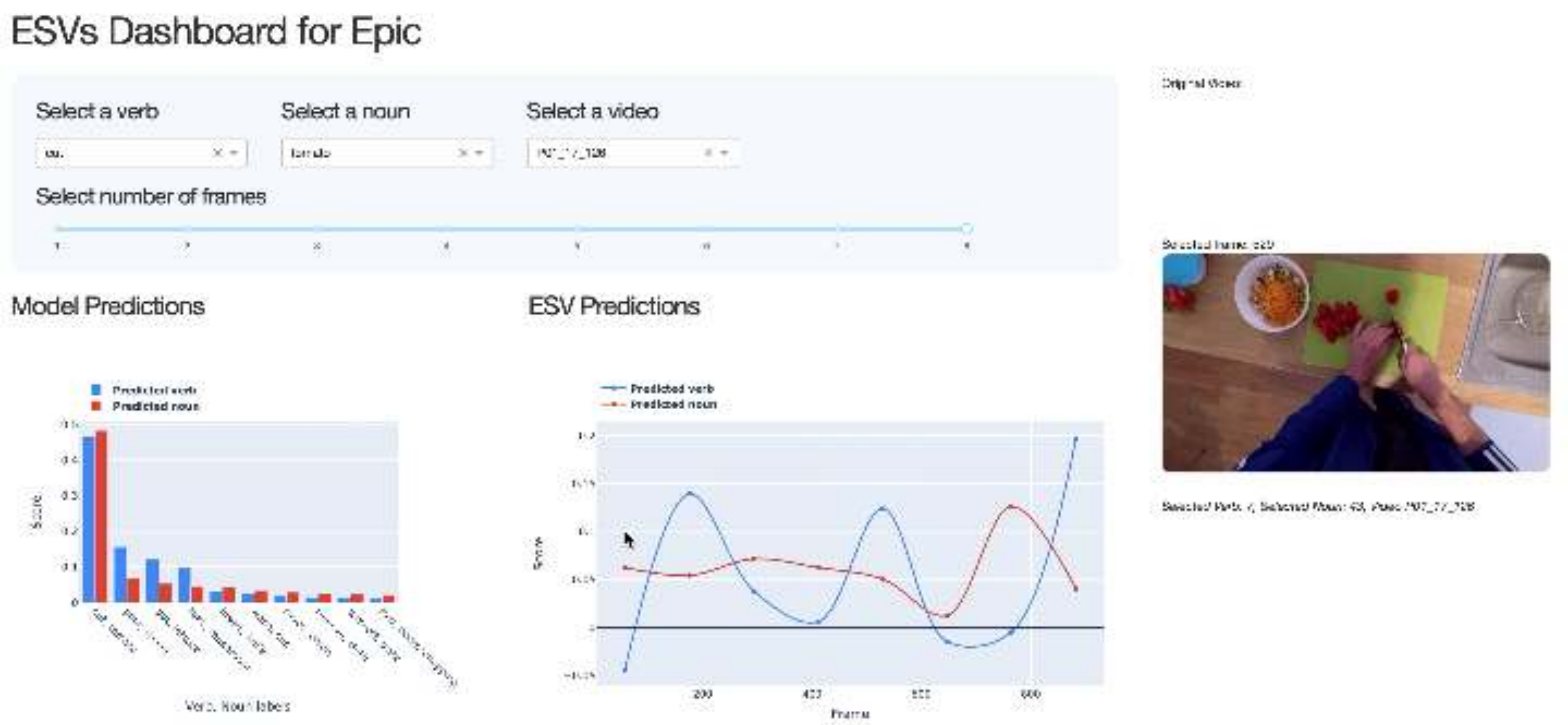
with: Will Price
Tom Stark



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



University of BRISTOL

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

The Team

Thank you

For further info, datasets, code, publications…

http://dimadamen.github.io

@dimadamen

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

Q&A