A Fine-Grained Perspective onto Object Interactions

University of BRISTOL

Dima Damen

26 December 2019

Natural interactions

University of BRISTOL
UNIVERSITY OF TORONTO
UNIVERSITÀ degli STUDI di CATANIA

Scaling Egocentric Vision: The EPIC-KITCHENS Dataset

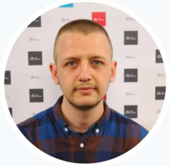

Dima Damen


Hazel Doughty


Giovanni M. Farinella


Sanja Fidler


Antonino Furnari


Evangelos Kazakos


Davide Moltisanti


Jonathan Munro


Toby Perrett


Will Price


Michael Wray

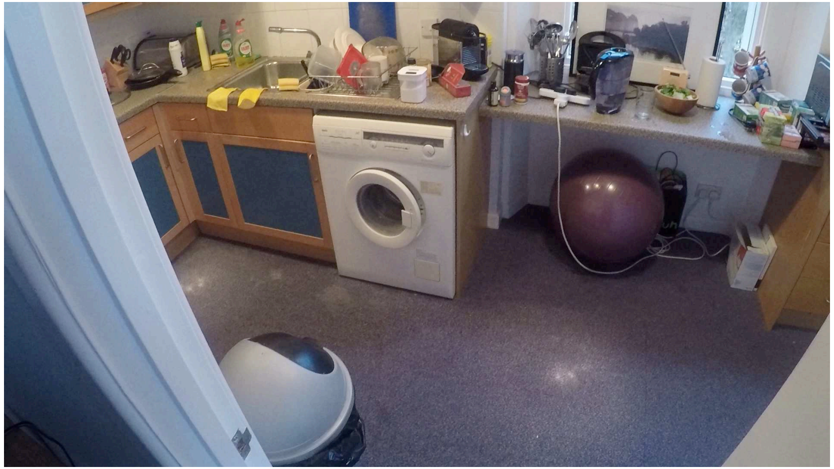
Scaling…

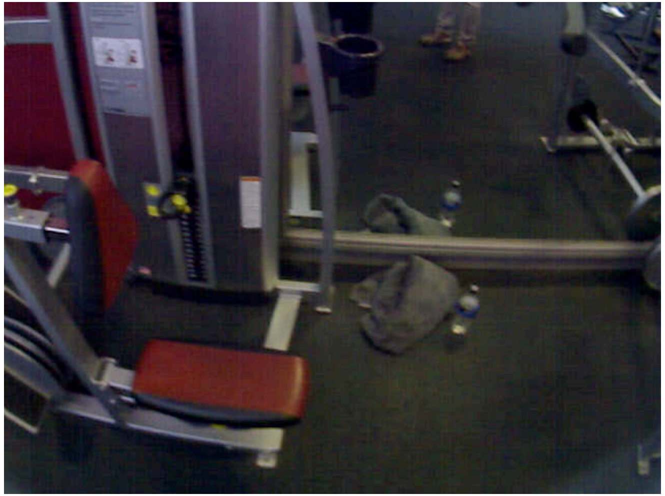
PASCAL2
Pattern Analysis, Statistical Modelling and Computational Learning

COCO
Common Objects in Context

EPIC KITCHENS

… Egocentric Vision

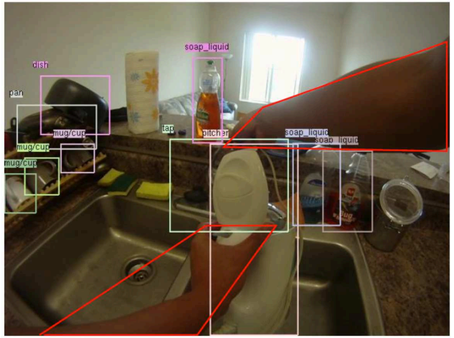




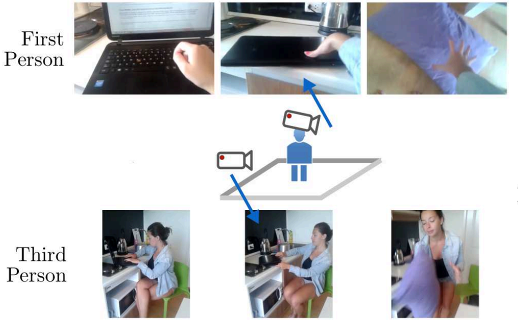
EPIC KITCHENS

Scaling Egocentric Vision

CMU (2009)



ADL (2012)



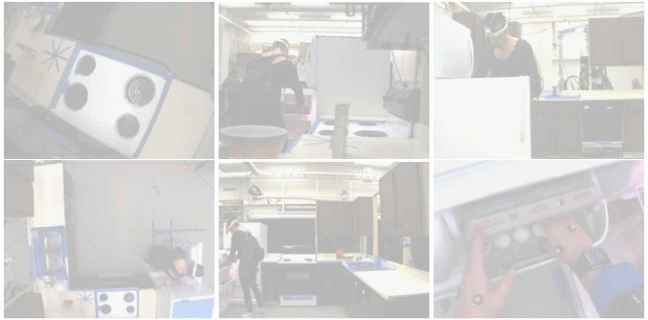
Charades-Ego (2018)

BEOID (2014)

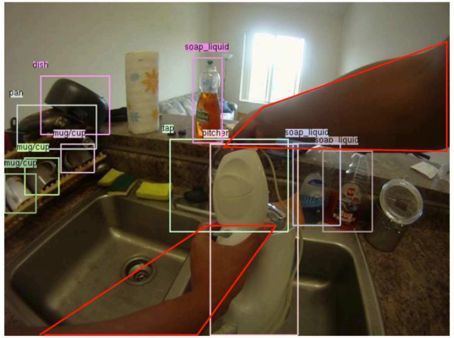
GTEA … (2011-)

EGTEA+ (2018)

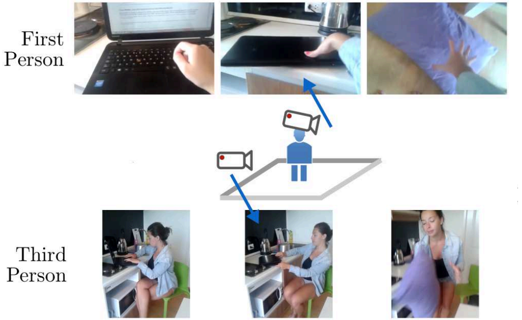
Native & Multiple Environments?



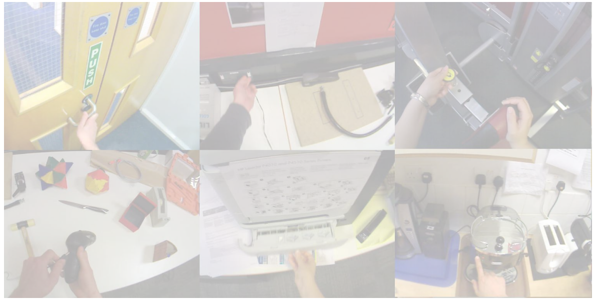
CMU (2009)



ADL (2012)



Charades-Ego (2018)



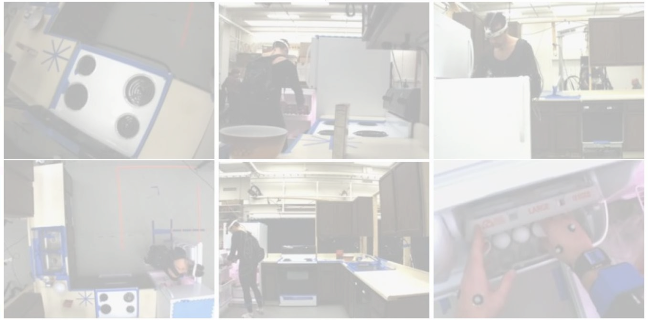
BEOID (2014)

GTEA … (2011-)

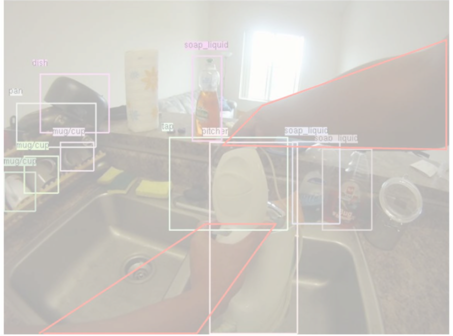
EGTEA+ (2018)

EPIC KITCHENS

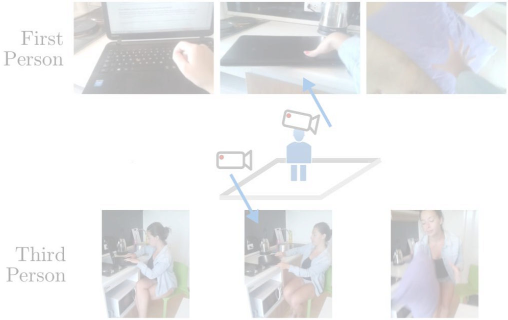
Non-scripted?



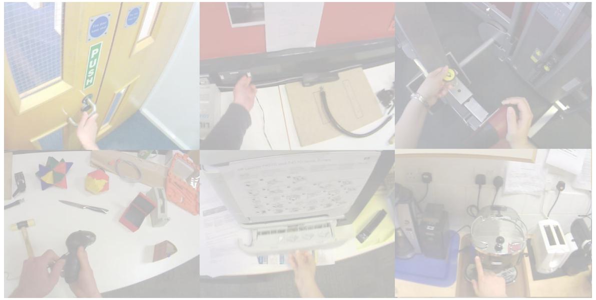
CMU (2009)



ADL (2012)



Charades-Ego (2018)

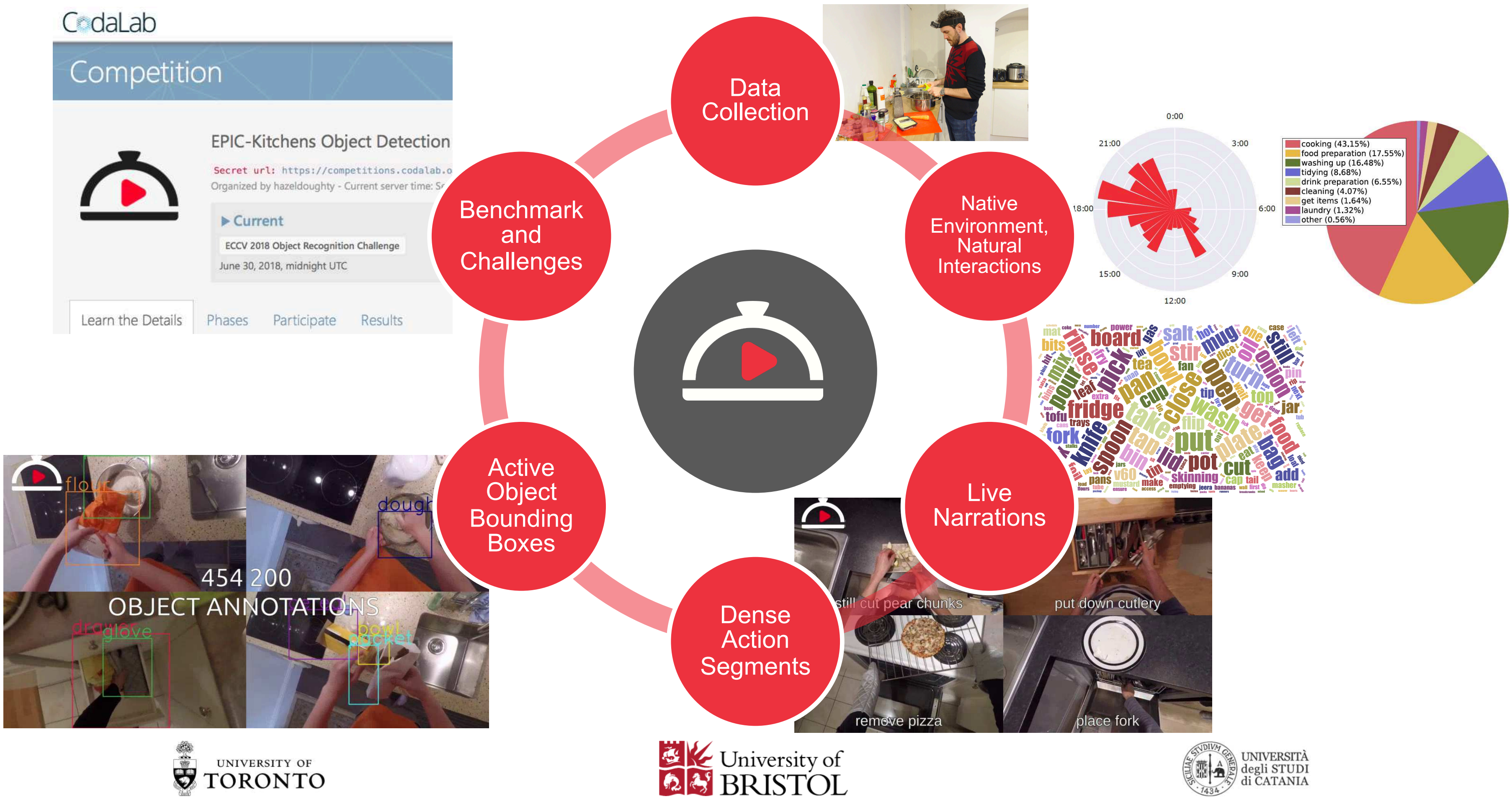


BEOID (2014)

GTEA … (2011-)

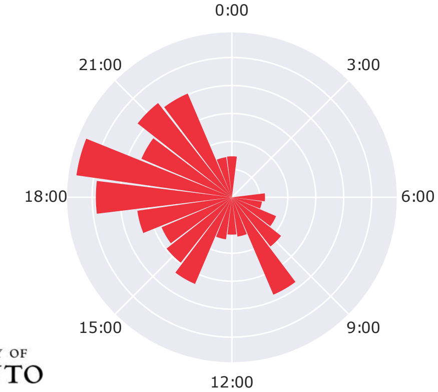
EGTEA+ (2018)

Scaling Egocentric Vision



Data Collection

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



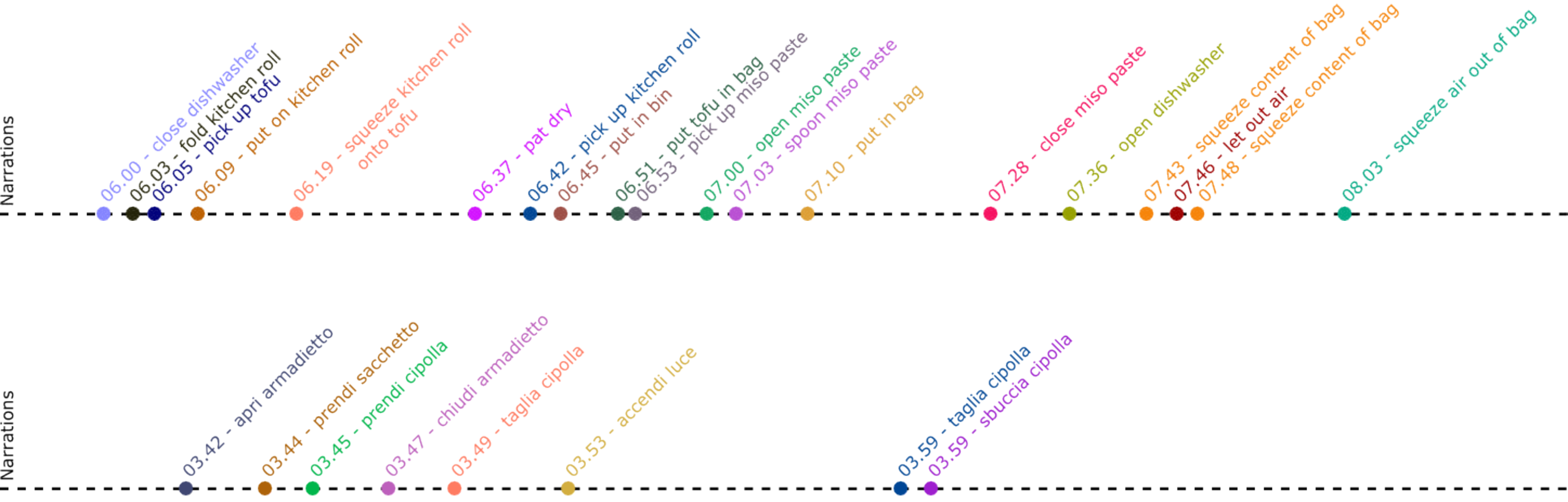
UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

EPIC KITCHENS

Annotations (1) - Narrations

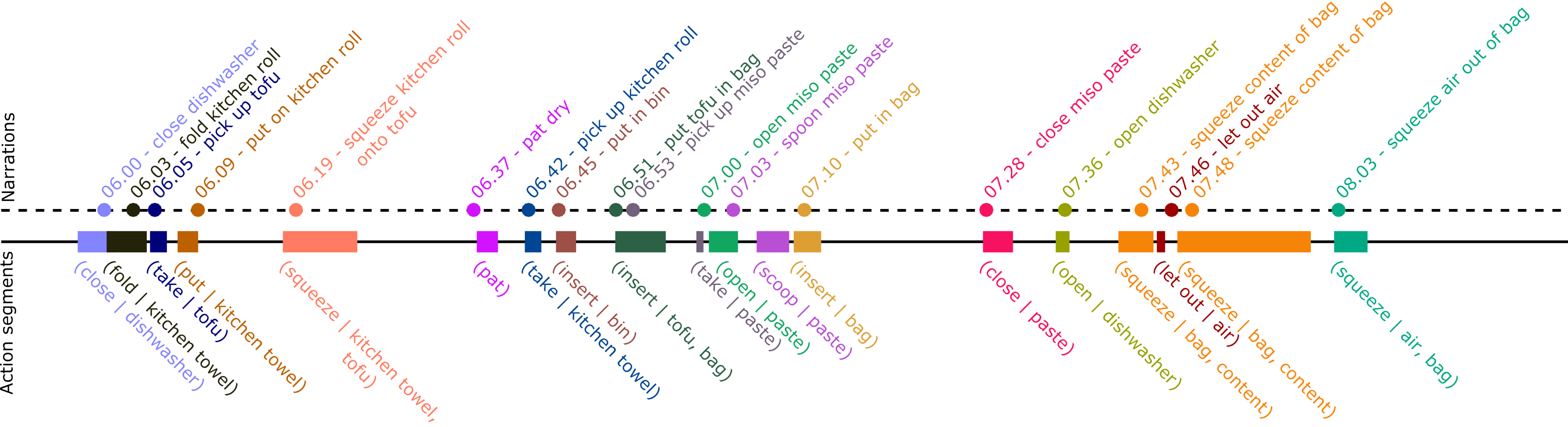


UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

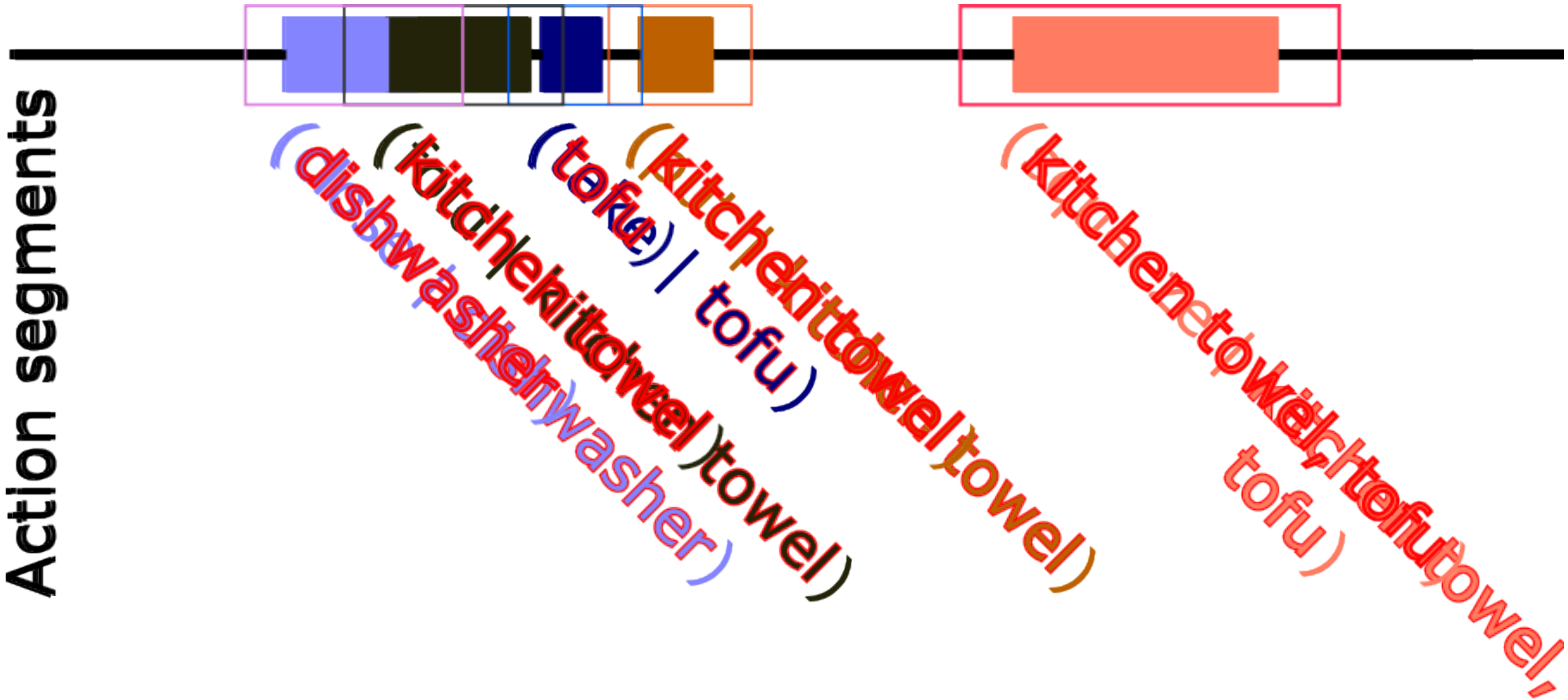
Narrations to Action Segments



UNIVERSITY OF TORONTO · University of BRISTOL · UNIVERSITÀ degli STUDI di CATANIA

39 000
ACTION SEGMENTS

Annotations (3) – Object Bounding Boxes



454 200
OBJECT ANNOTATIONS

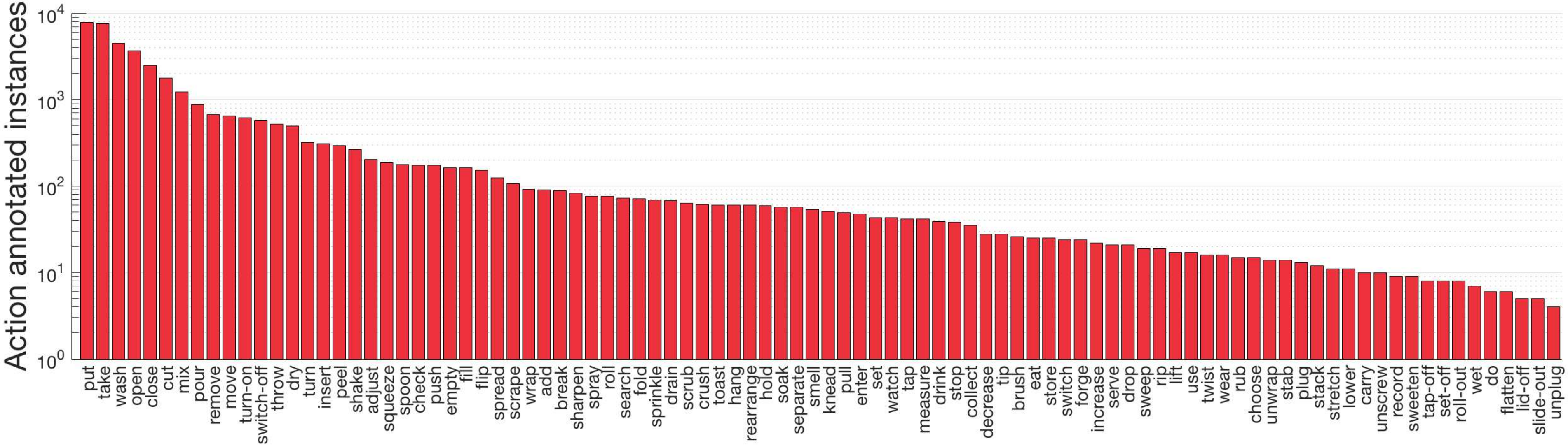
EPIC KITCHENS

Annotations (4) – Verb and Noun Classes

take, grab, pick, get, fetch, pick-up, ...

- 125 verb classes
- 331 noun classes

Annotations Statistics

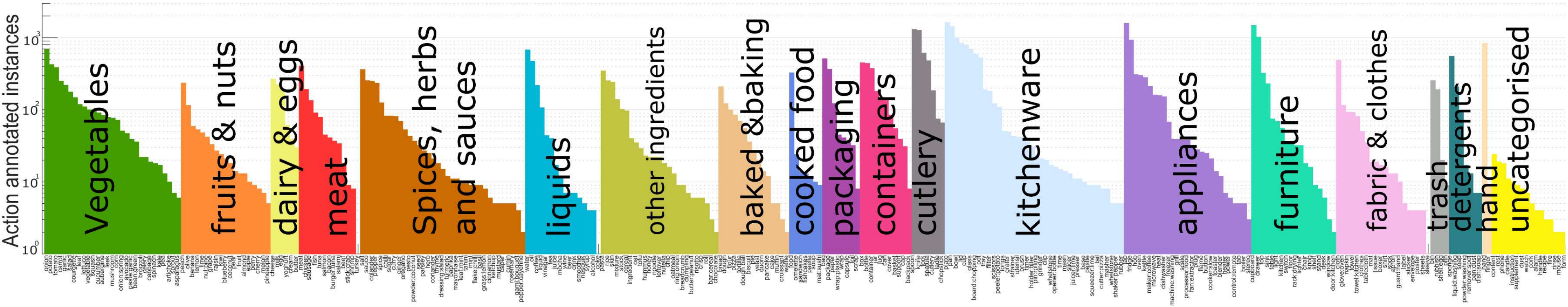


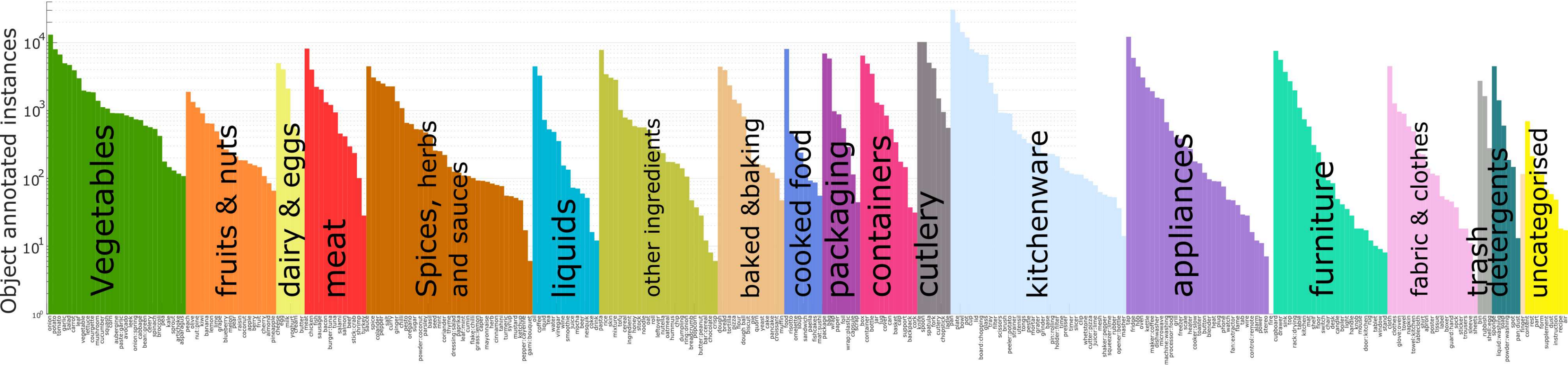
UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

Annotations Statistics





UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

Train/Test Splits

- 20% - Seen Test Set
  - 28 Kitchens
- 7% - Unseen Test Set
  - 4 Kitchens

Table 4: Statistics of test splits: seen (S1) and unseen (S2) kitchens

	#Subjects	#Sequences	Duration (s)	%	Narrated Segments	Action Segments	Bounding Boxes
Train/Val	28	272	141731		28,587	28,561	326,388
S1 Test	28	106	39084	20%	8,069	8,064	97,872
S2 Test	4	54	13231	7%	2,939	2,939	29,995

UNIVERSITY OF TORONTO

University of BRISTOL

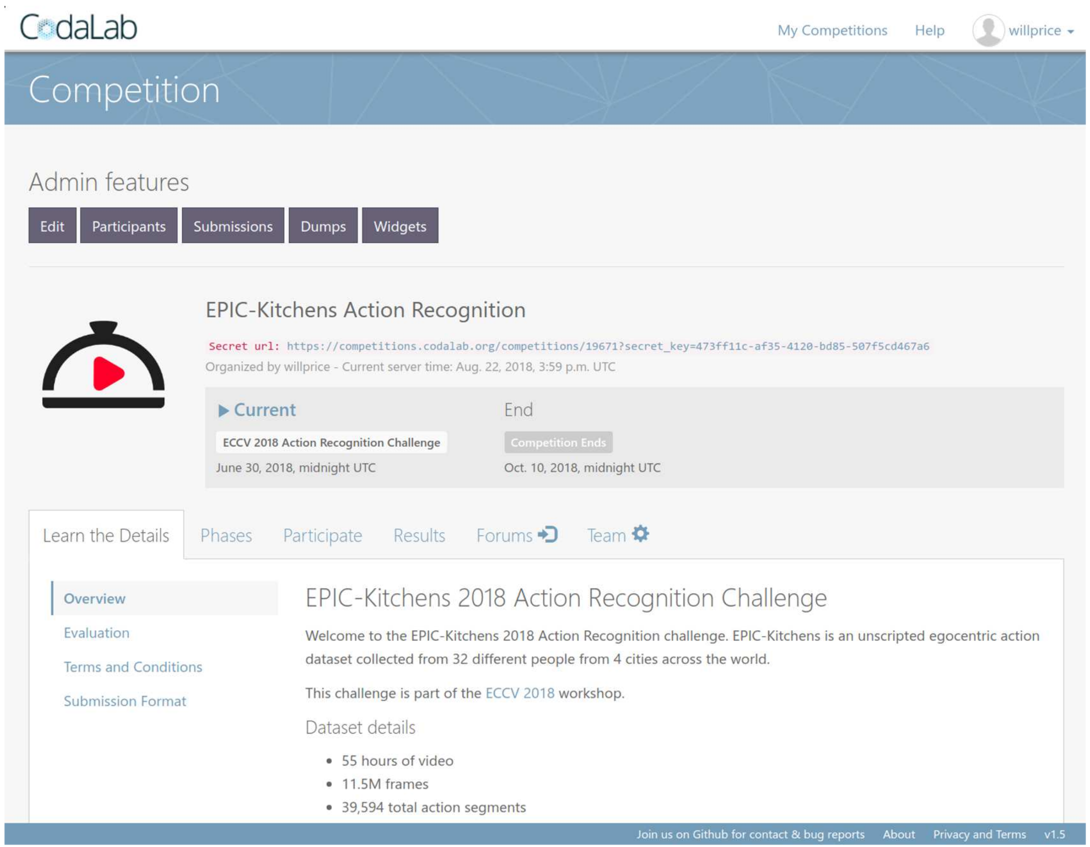
UNIVERSITÀ degli STUDI di CATANIA

Open Challenges

EPIC KITCHENS

Three open challenges:

- Action Recognition
- Action Anticipation
- Object Detection



UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

Action Recognition Challenge

Action Recognition Challenge

Given a trimmed action segment:

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

classify the action within.

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

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

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

UNIVERSITY OF TORONTO

University of BRISTOL

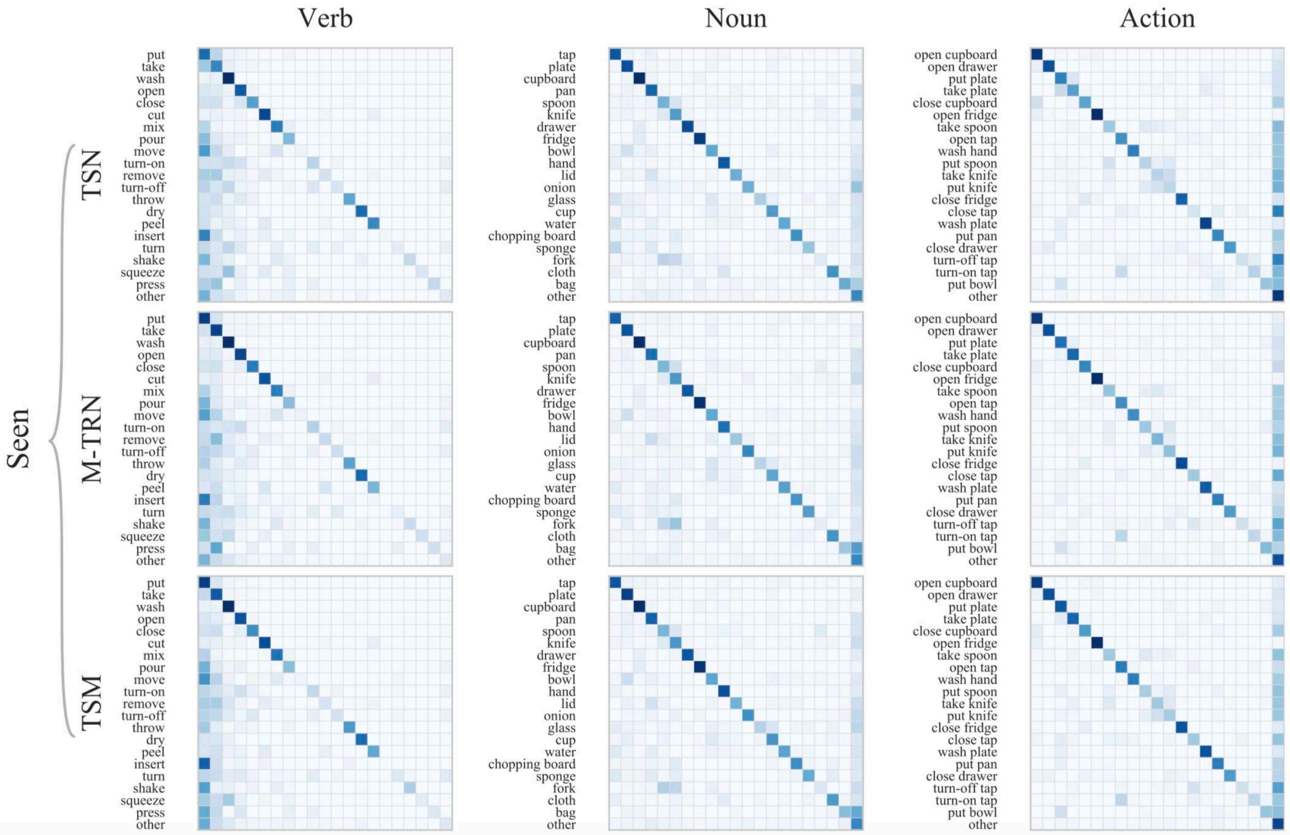
UNIVERSITÀ degli STUDI di CATANIA

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	3	08/12/19	UTS_BAIDU	69.80 (1)	52.27 (1)	41.37 (1)	90.95 (2)	76.71 (1)	63.59 (1)	63.55 (1)	46.86 (1)	25.13 (1)	46.94 (1)	49.17 (1)	26.39 (1)
2	action_banks	2	11/11/19		63.23 (6)	47.00 (5)	39.34 (2)	86.94 (8)	69.55 (6)	62.23 (2)	54.86 (5)	43.43 (3)	24.47 (2)	40.59 (6)	43.46 (5)	25.43 (2)
3	aptx4869lm	10	09/27/19	GT-WISC-MPI	68.14 (2)	49.62 (2)	38.18 (3)	88.93 (5)	72.77 (3)	58.30 (4)	53.45 (7)	42.88 (4)	22.47 (4)	41.77 (4)	46.55 (2)	23.73 (3)
4	weiyaowang	13	11/14/19		65.91 (4)	48.48 (3)	36.76 (4)	89.51 (3)	71.36 (4)	56.17 (6)	51.76 (9)	41.26 (6)	20.84 (5)	46.73 (3)	44.92 (3)	21.98 (5)
5	TBN_Ensemble	1	07/20/19	Bristol-Oxford	66.10 (3)	47.88 (4)	36.66 (5)	91.28 (1)	72.80 (2)	58.62 (3)	60.73 (4)	44.89 (2)	24.01 (3)	46.81 (2)	43.88 (4)	22.92 (4)
6	Sudhakaran	11	08/10/19	FBK_HuPBA	63.34 (5)	44.75 (6)	35.54 (6)	89.01 (4)	69.88 (5)	57.18 (5)	63.21 (2)	42.26 (5)	19.76 (6)	37.77 (8)	41.28 (6)	21.19 (6)
7	antoninofurnari	1	07/19/19		56.93 (9)	43.05 (7)	33.06 (7)	85.68 (11)	67.12 (7)	55.32 (7)	50.42 (10)	39.84 (7)	18.91 (7)	37.82 (7)	38.11 (7)	19.12 (7)
8	georkap	6	08/29/19	uunol	58.36 (7)	40.87 (8)	30.79 (8)	87.49 (7)	64.71 (9)	52.27 (8)	53.57 (6)	35.73 (11)	15.86 (9)	37.33 (9)	36.23 (10)	17.20 (9)
9	bduke	2	11/16/19		57.52 (8)	38.00 (12)	26.98 (9)	88.27 (6)	62.66 (13)	45.56 (10)	49.22 (11)	35.40 (13)	18.25 (8)	40.95 (5)	36.80 (8)	18.63 (8)
10	hepic	6	12/04/19		49.96 (11)	39.24 (10)	25.30 (10)	85.80 (10)	64.05 (12)	49.19 (9)	42.39 (14)	35.92 (10)	13.93 (10)	35.68 (10)	35.50 (11)	13.65 (10)
11	Saptarshi_sinha	8	12/19/19		52.78 (10)	39.72 (9)	23.69 (11)	81.66 (18)	64.56 (10)	40.61 (13)	23.77 (20)	36.84 (9)	12.26 (15)	26.21 (13)	36.31 (9)	12.70 (12)
12	plnet	3	11/20/19		49.78	32.62	21.69	82.93	58.57	39.59	40.13	28.45	13.15	31.60	30.45	12.75

Evaluating Action Recognition Models

with: Will Price



W Price, D Damen (2019). An Evaluation of Action Recognition Models on EPIC-Kitchens. Arxiv

UNIVERSITY OF TORONTO
University of BRISTOL
UNIVERSITÀ degli STUDI di CATANIA

Evaluating Action Recognition Models

with: Will Price

Model	GFLOP/s RGB	GFLOP/s Flow	Params (M) RGB	Params (M) Flow
TSN	33.12	35.33	24.48	24.51
TRN	33.12	35.32	25.33	25.35
M-TRN	33.12	35.33	27.18	27.21
TSM	33.12	35.33	24.48	24.51

Table 3: Model parameter and FLOP/s count using a ResNet-50 backbone with 8 segments for a single video.

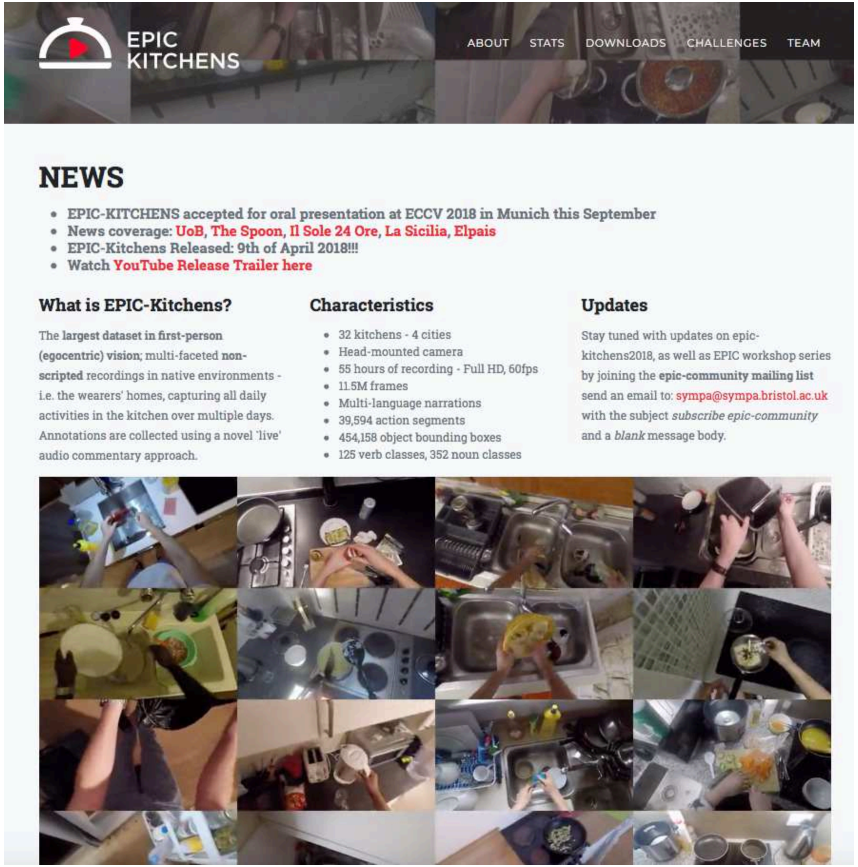
Models Released

W Price, D Damen (2019). An Evaluation of Action Recognition Models on EPIC-Kitchens. Arxiv

EPIC KITCHENS

More?

http://epic-kitchens.github.io



UNIVERSITY OF TORONTO

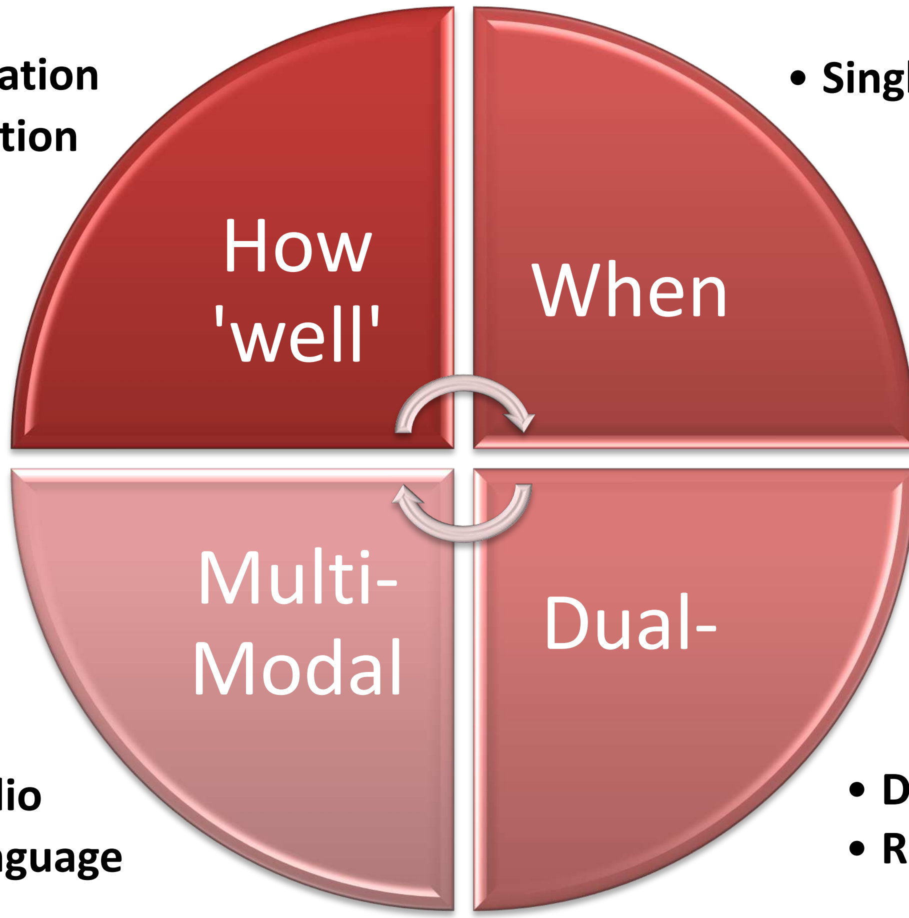
University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

Fine-Grained Object Interactions

- Skill Determination
- Action Completion

- Single-timestamp



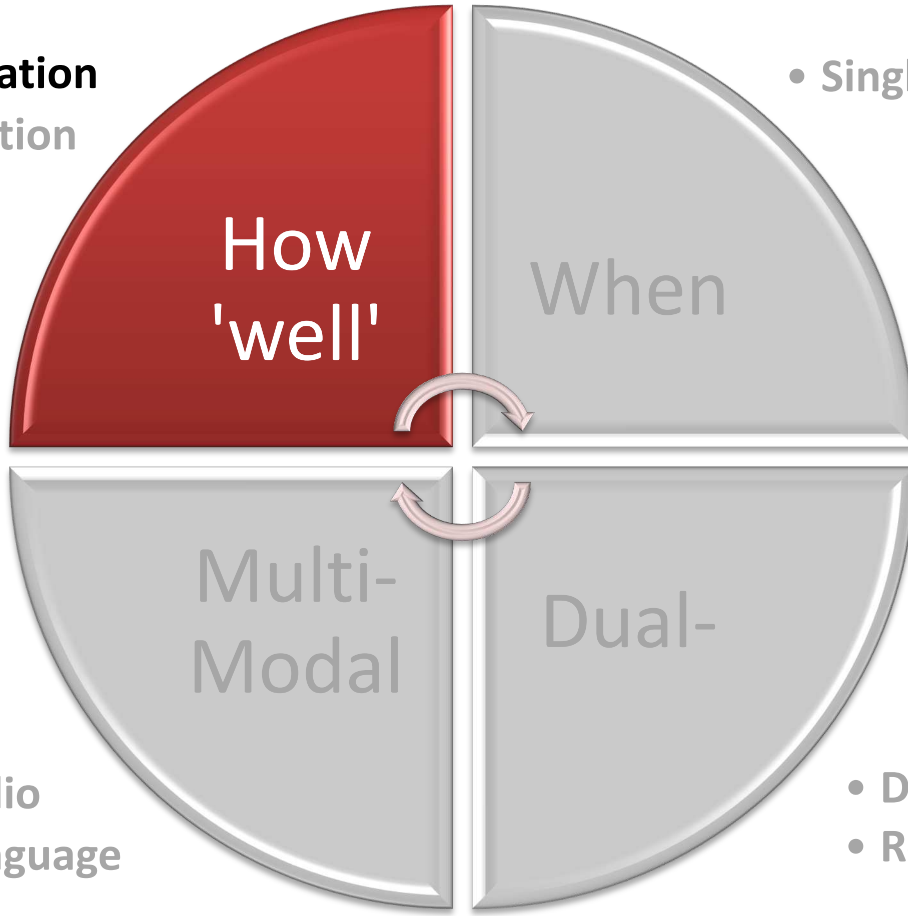
- Vision+Audio
- Vision+Language

- DDLSTM
- Retro-Actions

Fine-Grained Object Interactions

- **Skill Determination**
- Action Completion

- Single-timestamp



- Vision+Audio
- Vision+Language

- DDLSTM
- Retro-Actions

Who's Better? Who's Best? Skill Determination in Video using Deep Ranking

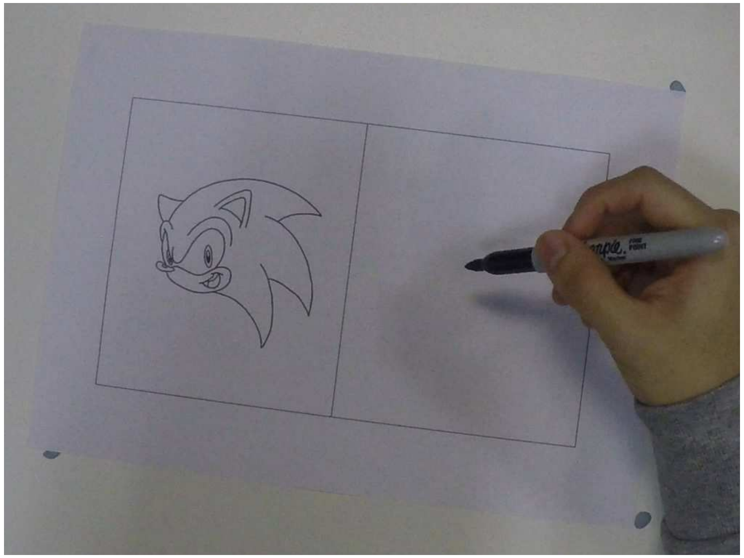
with: Hazel Doughty
Walterio Mayol-Cuevas

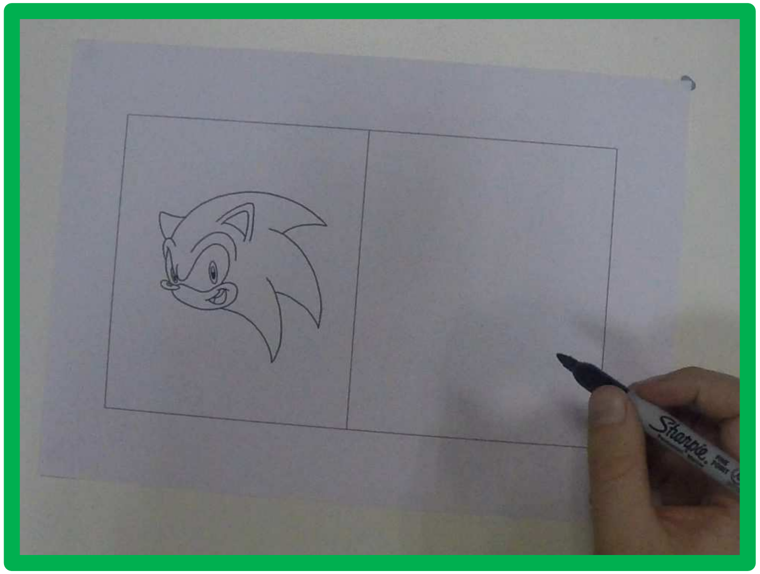
Assess relative skill for a collection of video sequences, applicable to a variety of tasks.

with: Hazel Doughty
Walterio Mayol-Cuevas

Skill Determination from Video

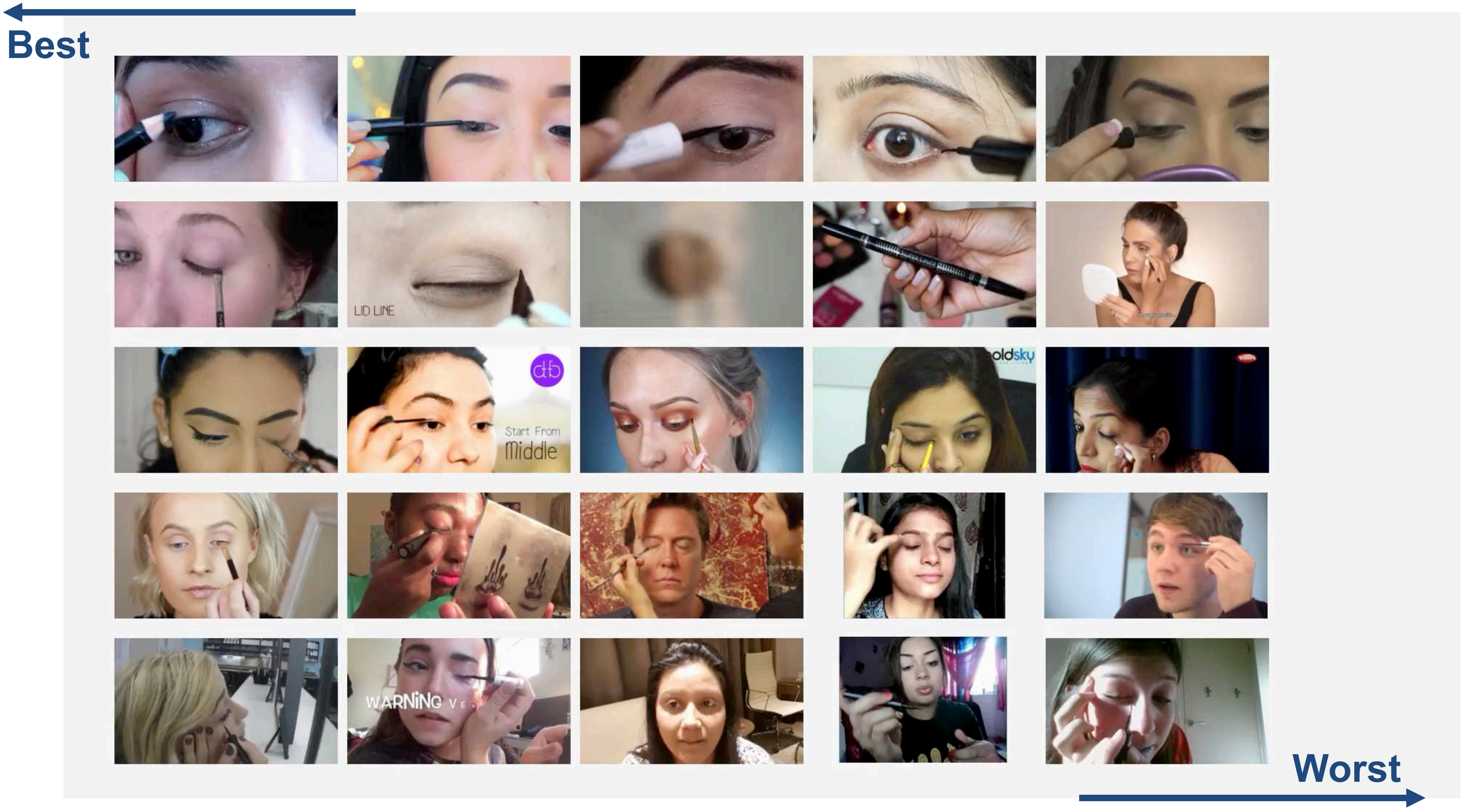
Input: Pairwise annotations of videos, indicating higher skill or no skill preference





Skill Determination in Video

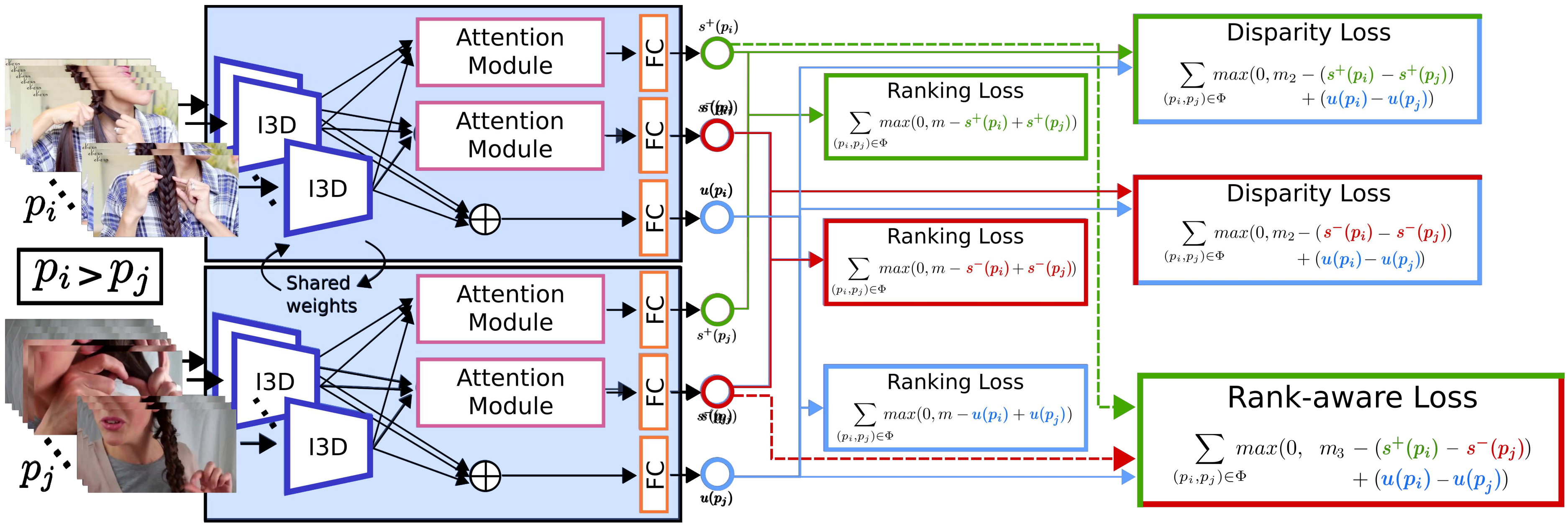
with: Hazel Doughty
Walterio Mayol-Cuevas



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)*

The Pros and Cons: Rank-Aware Temporal Attention

with: Hazel Doughty
Walterio Mayol-Cuevas



The Pros and Cons: Rank-Aware Temporal Attention

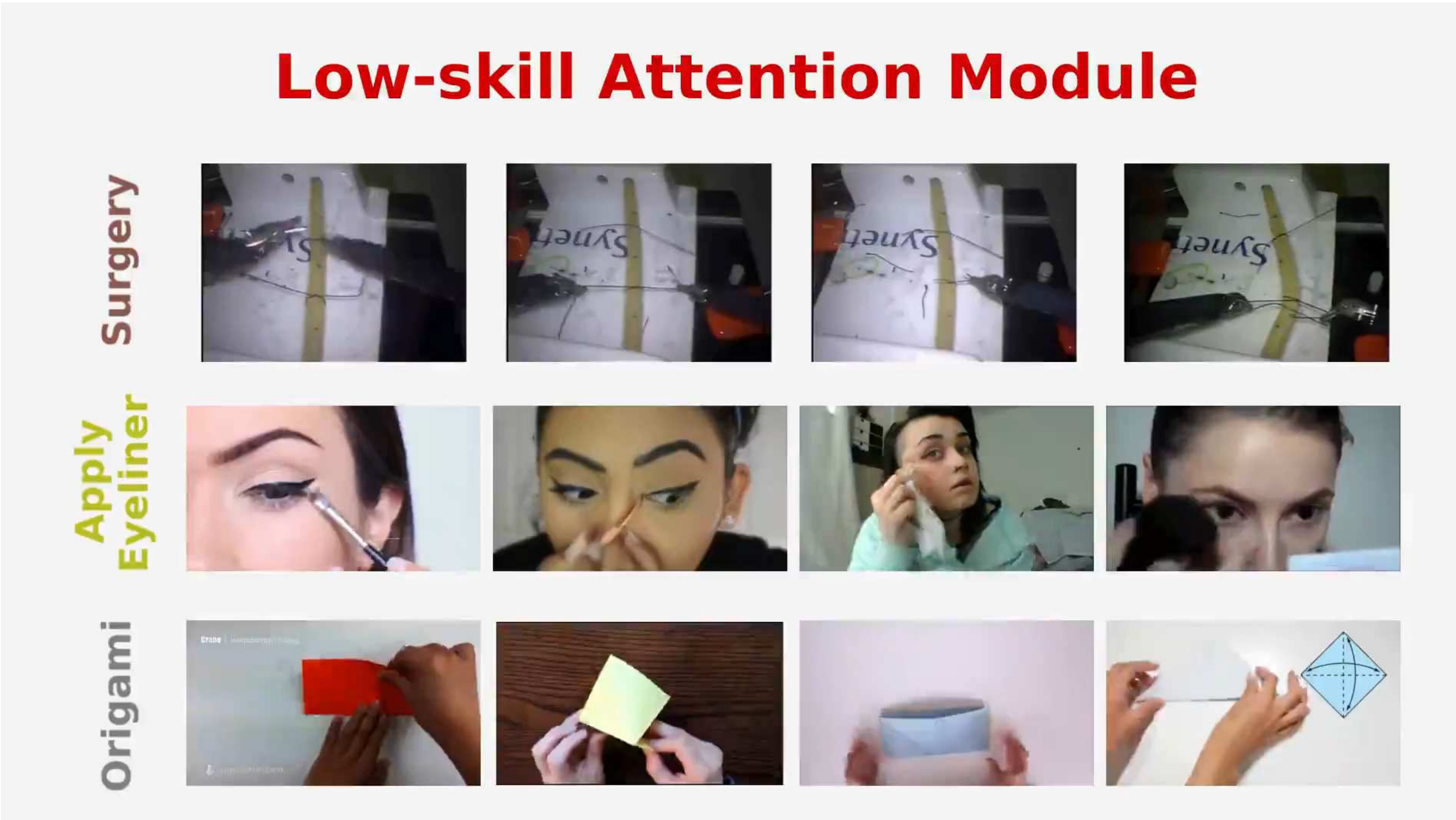
with: Hazel Doughty
Walterio Mayol-Cuevas

Method	EPIC Skills	BEST
Who's Better [7]	76.0	75.8
Last Segment	76.8	61.0
Uniform Weighting	78.8	73.6
Softmax Attention	74.5	72.3
STPN [22]	74.3	70.0
Ours (Rank Aware Attention)	**80.3**	**81.2**

Table 2. Results of our method in comparison to baseline. Our final method outperforms every baseline on both datasets.

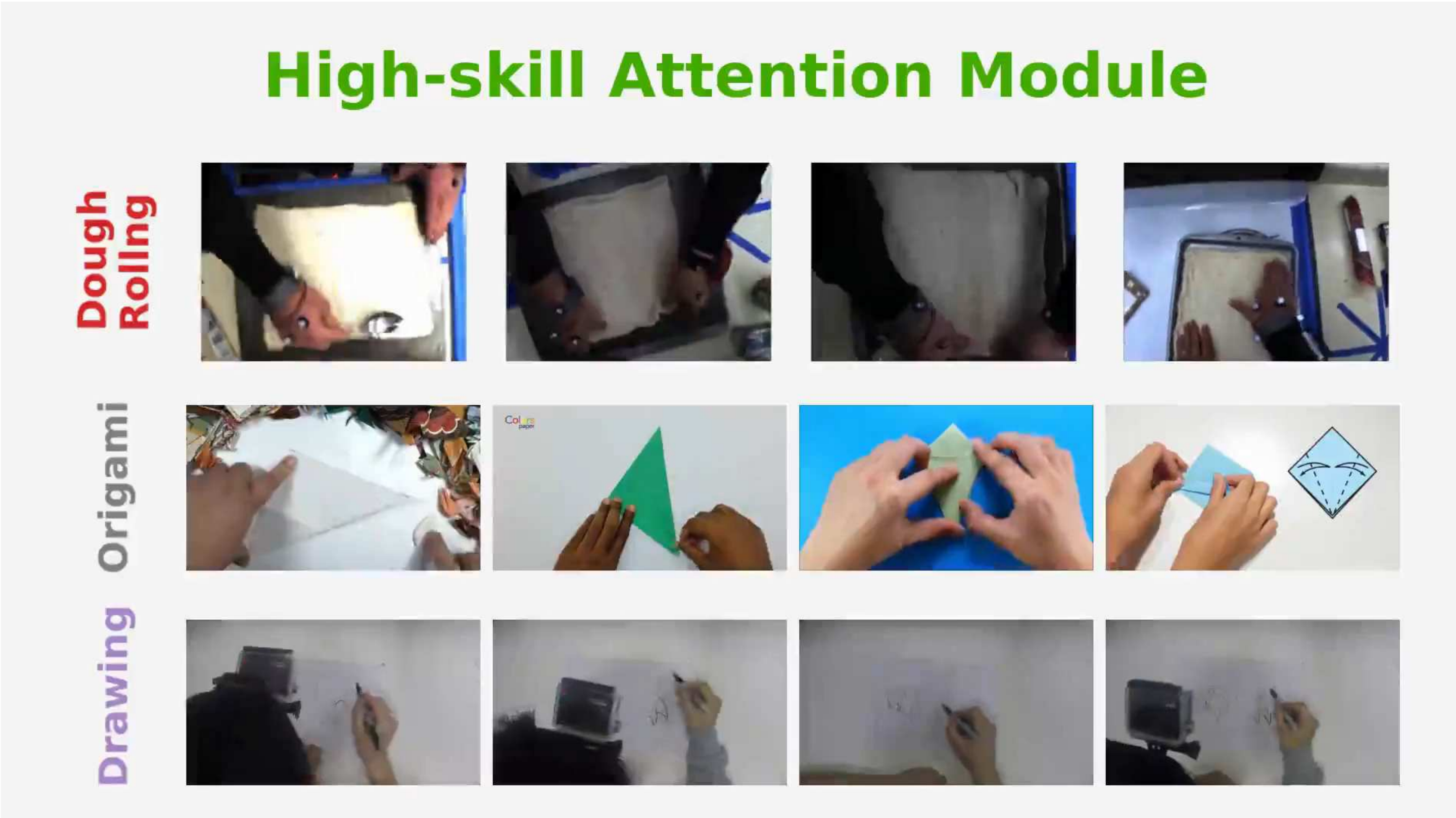
The Pros and Cons: Rank-Aware Temporal Attention

with: Hazel Doughty
Walterio Mayol-Cuevas



The Pros and Cons: Rank-Aware Temporal Attention

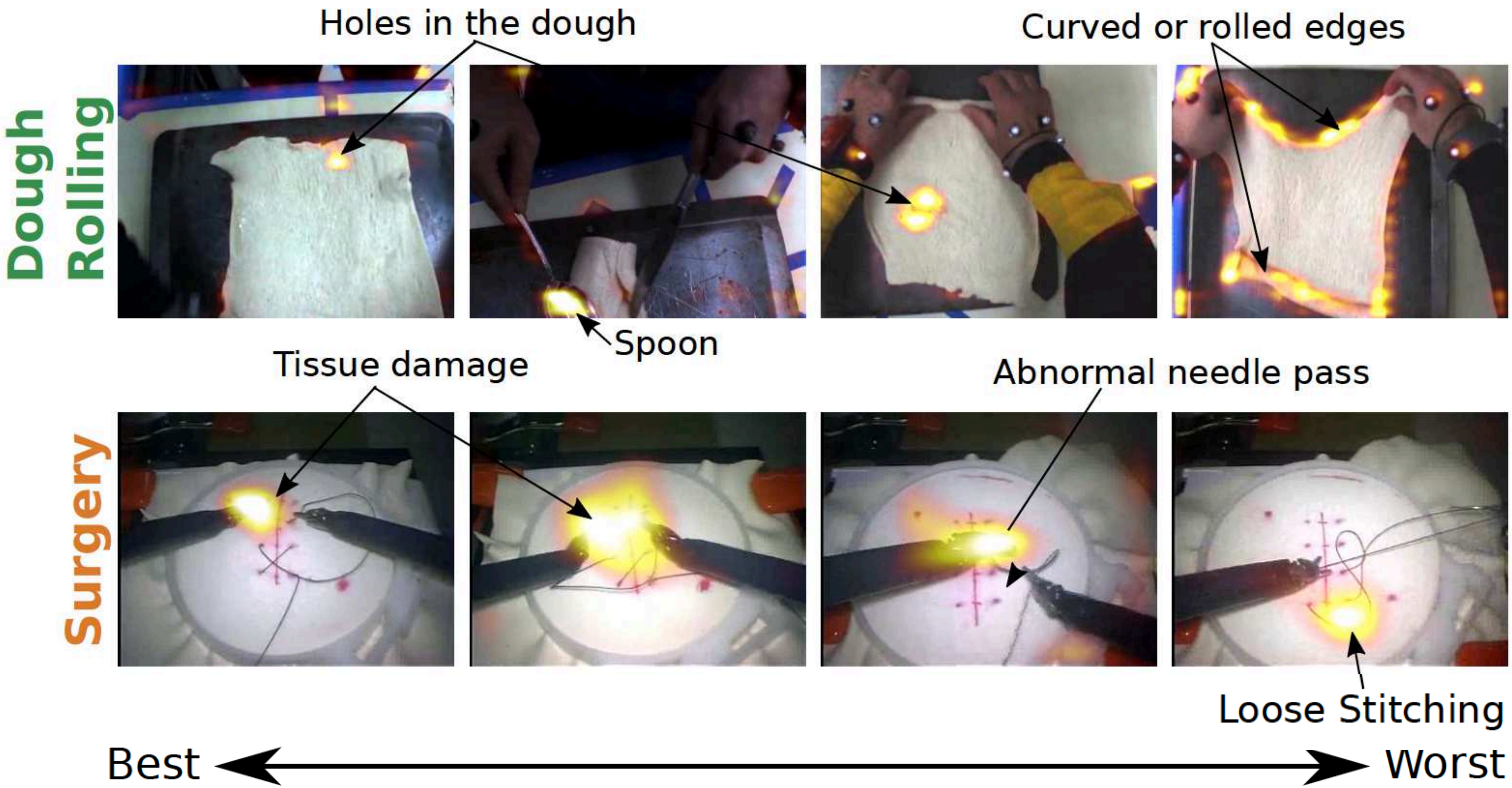
with: Hazel Doughty
Walterio Mayol-Cuevas



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, D Damen, W Mayol-Cuevas (2018). Who's Better? Who's Best? Pairwise Deep Ranking for Skill Determination. CVPR

The Pros and Cons: Rank-Aware Temporal Attention

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

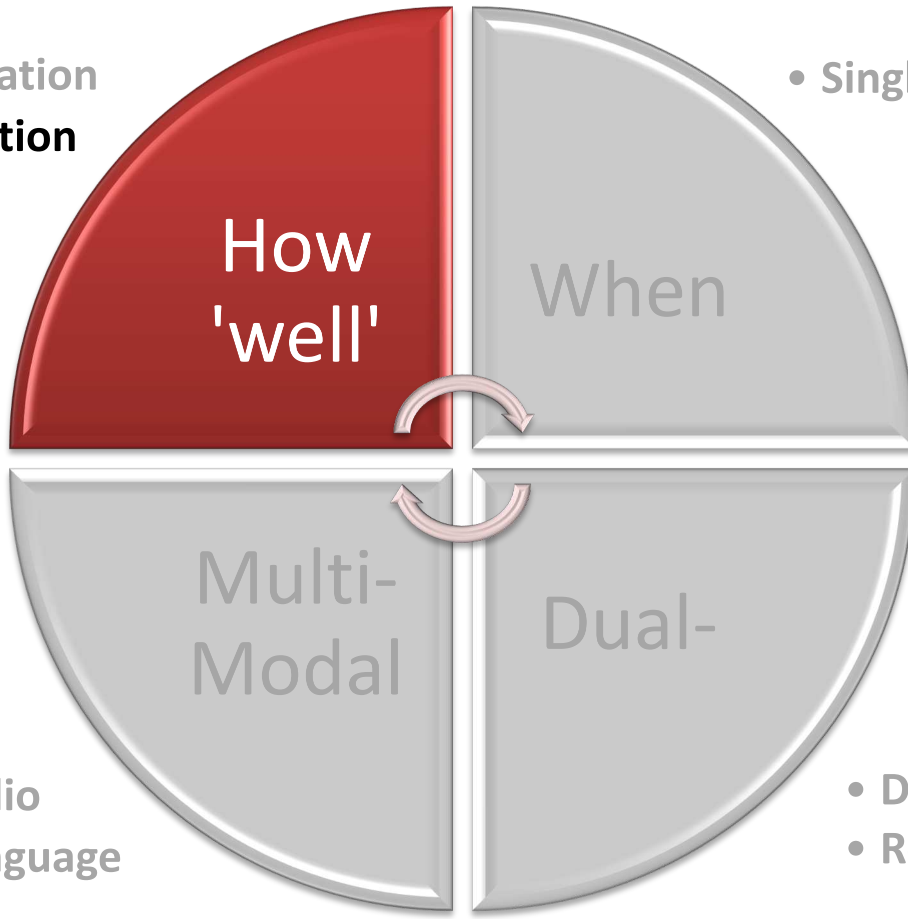
Downloads

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

Fine-Grained Object Interactions

- Skill Determination
- **Action Completion**

- Single-timestamp



- Vision+Audio
- Vision+Language

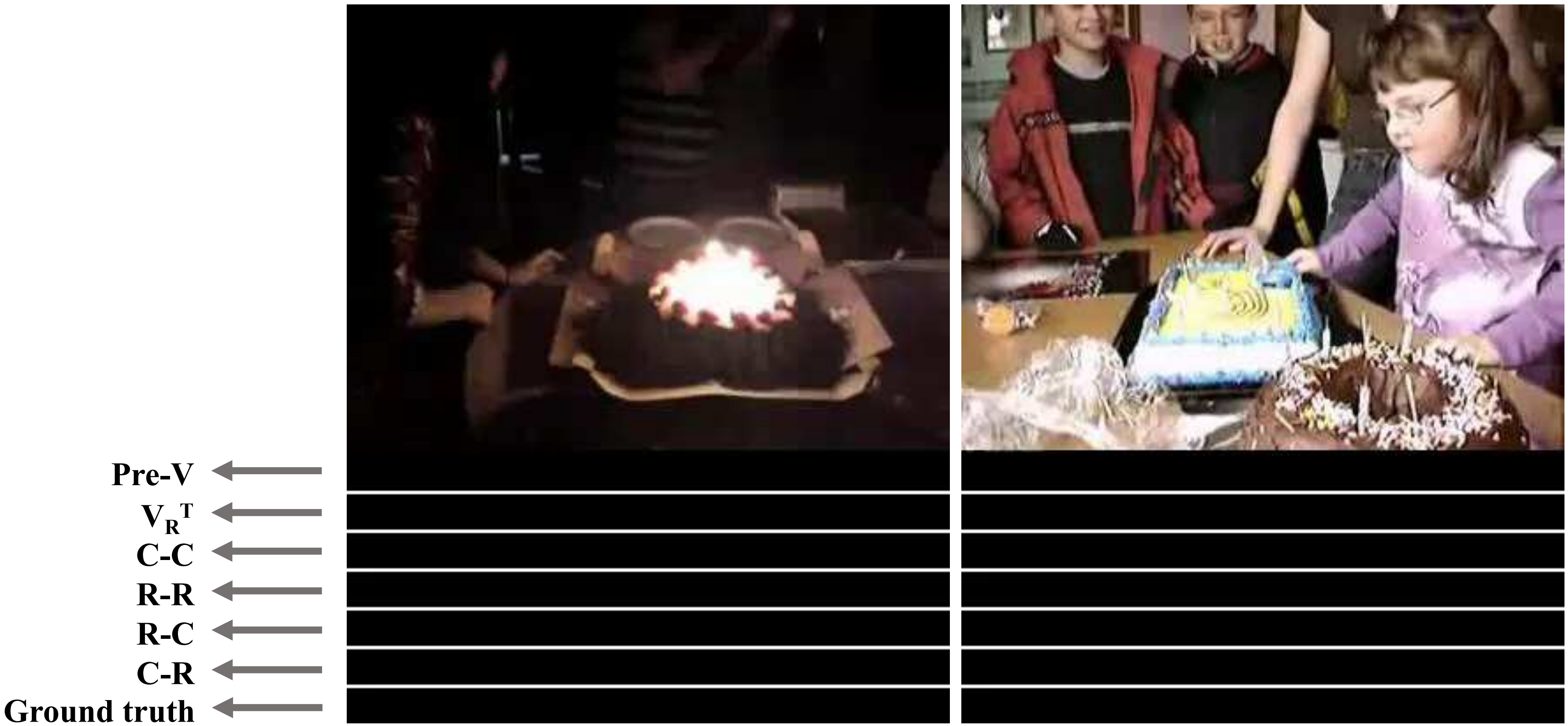
- DDLSTM
- Retro-Actions

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

with: Farnoosh Heidarivincheh
Majid Mirmehdi

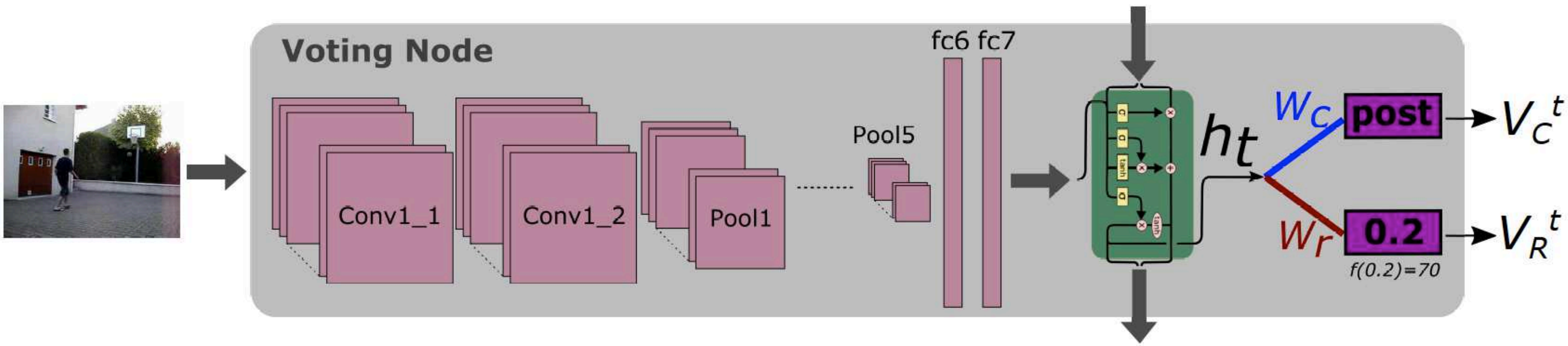
Action Completion Detection



with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

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

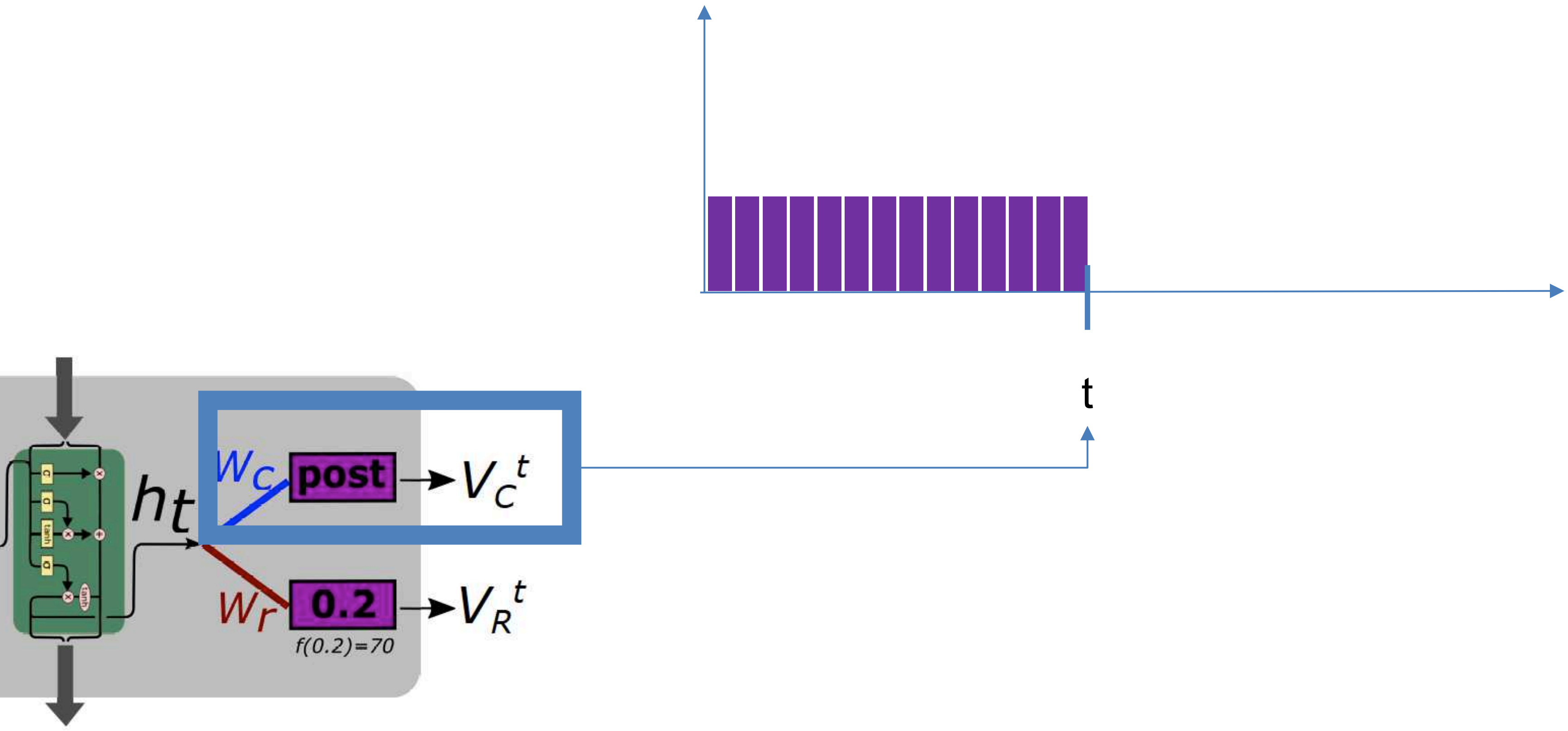


F Heidarivincheh, M Mirmehdi, D Damen (2018). Action Completion: A Temporal Model for Moment Detection. BMVC

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

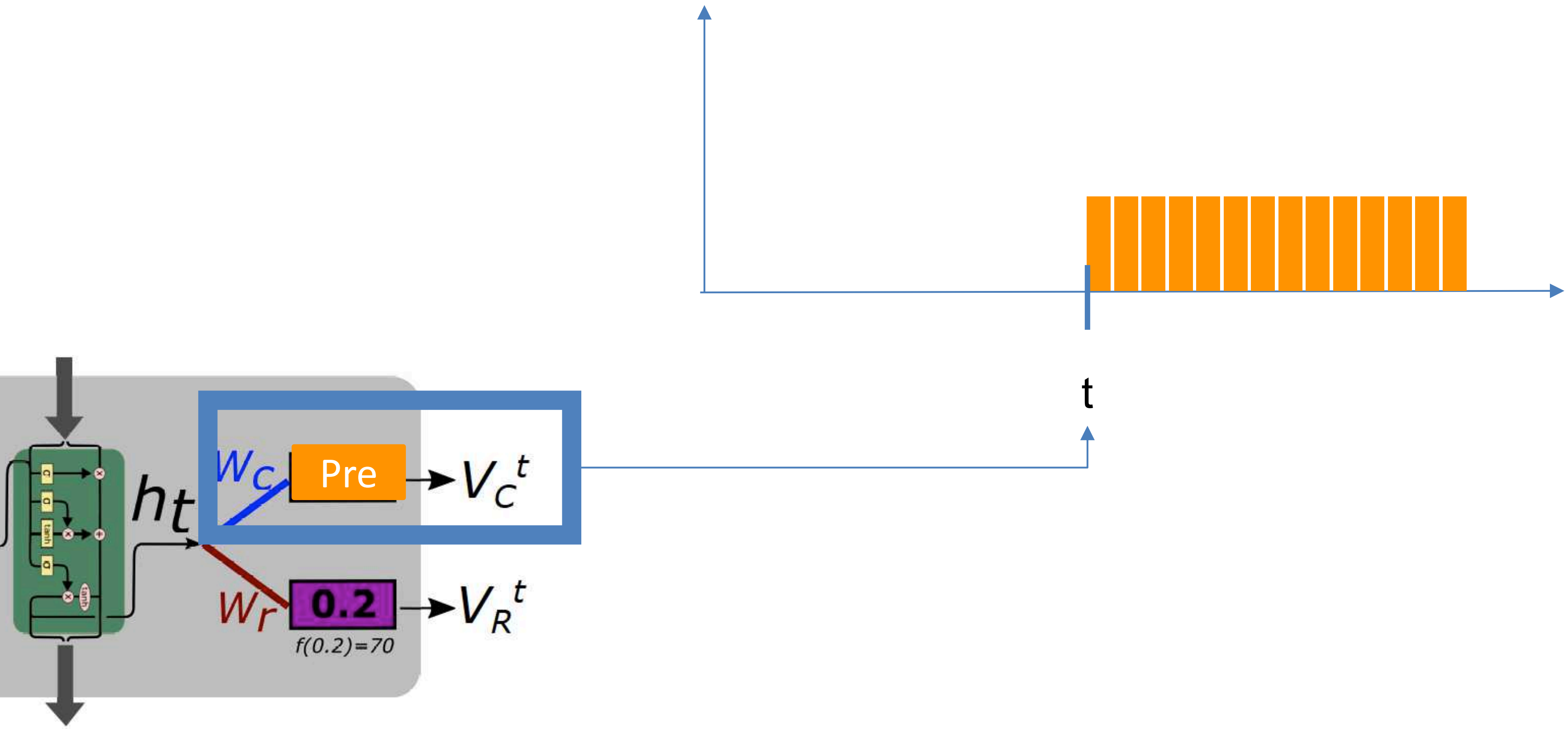
1. Classification-Based Voting



with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

1. Classification-Based Voting

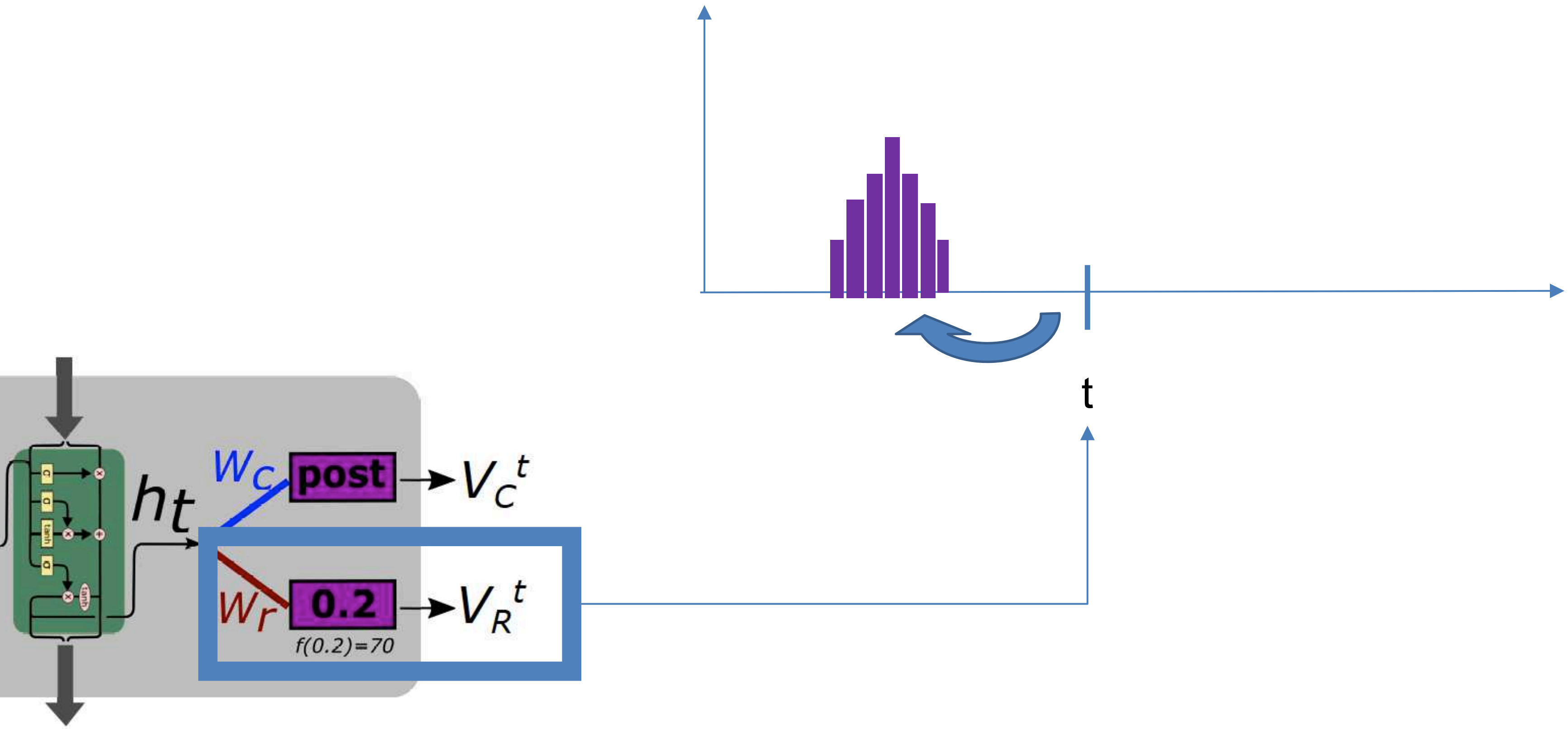


F Heidarivincheh, M Mirmehdi, D Damen (2018). Action Completion: A Temporal Model for Moment Detection. BMVC

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

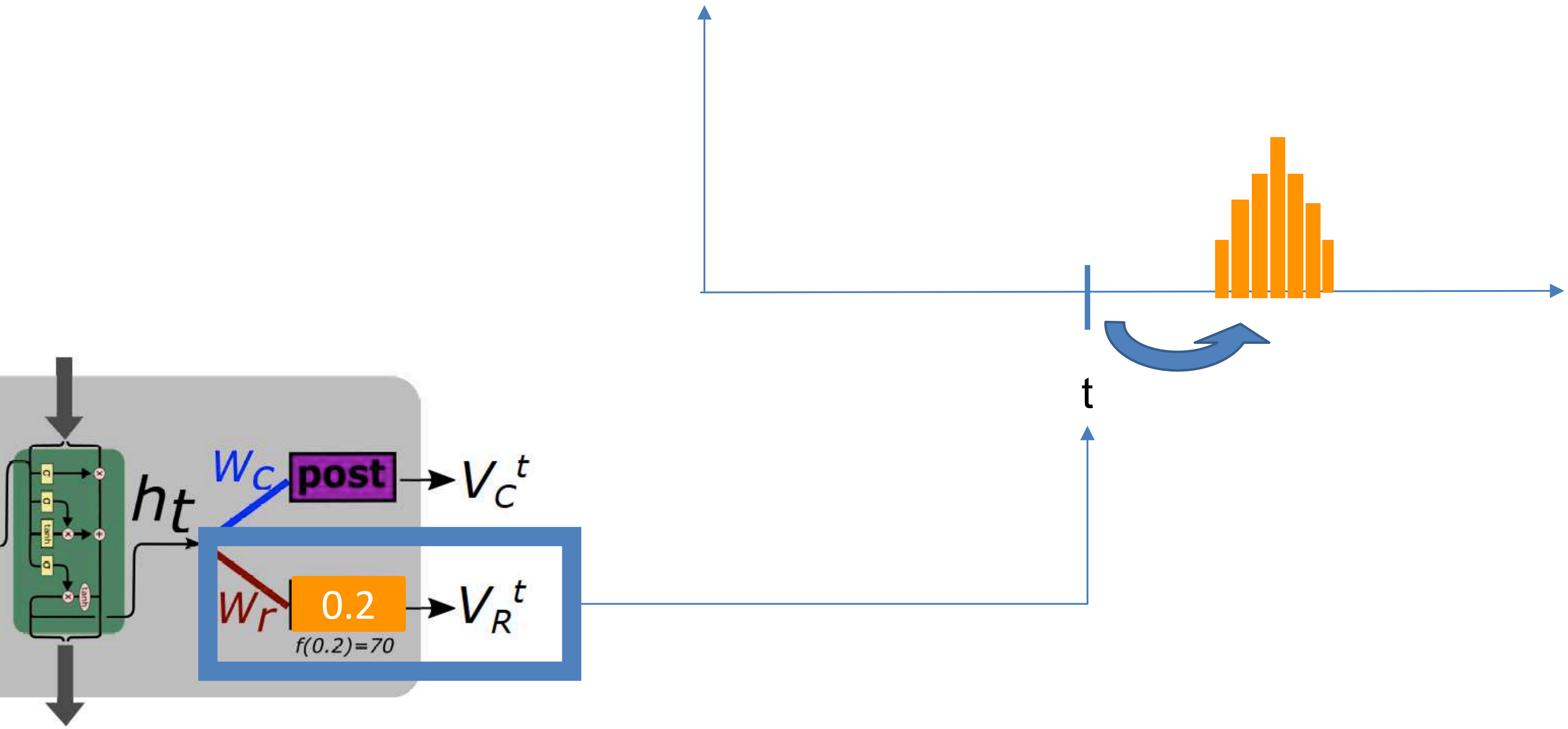
2. Regression-Based Voting



with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

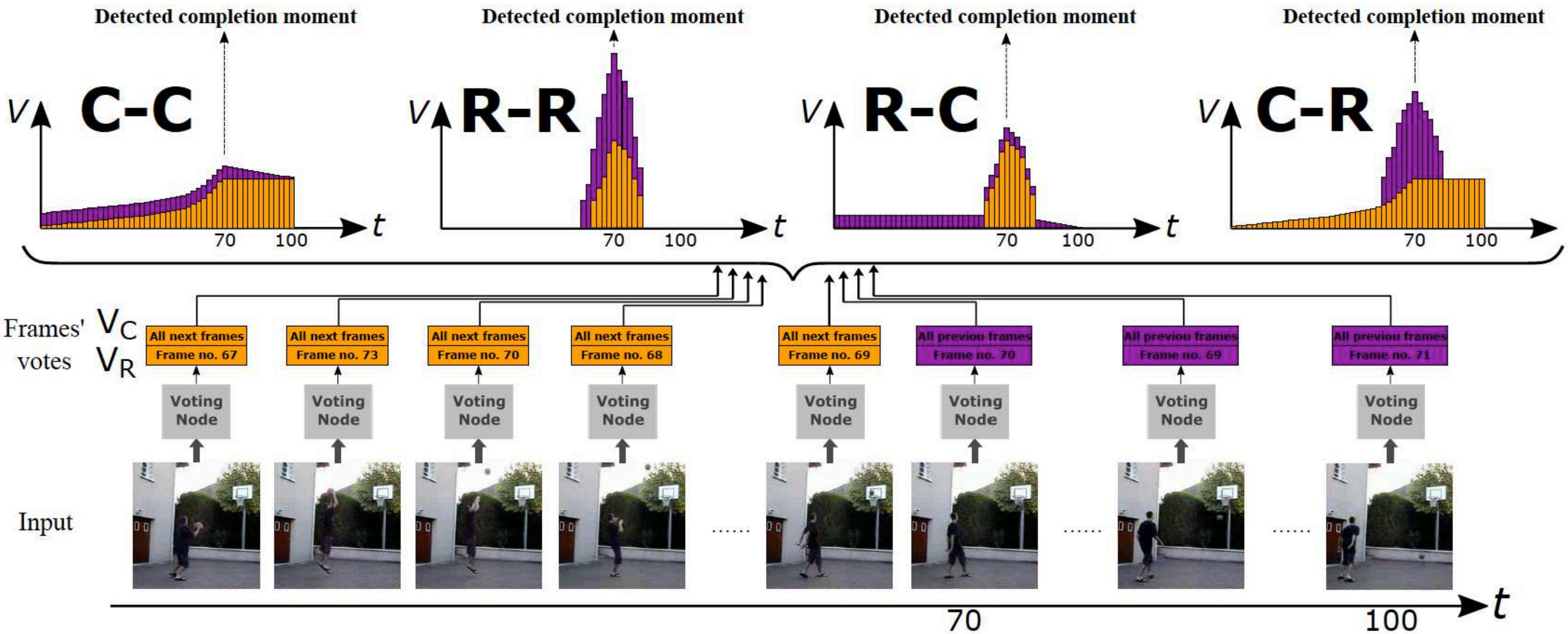
2. Regression-Based Voting



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

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection



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

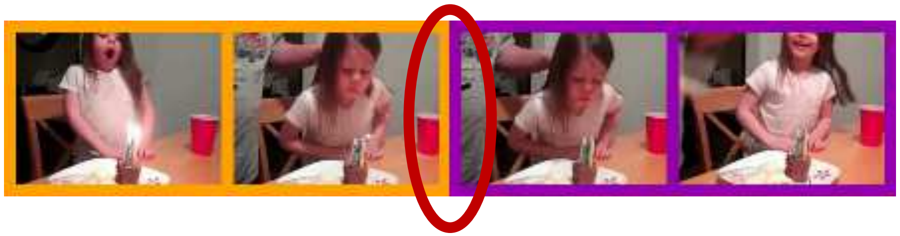
with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

Frame-level labels**:** annotations are expensive, subjective and noisy.



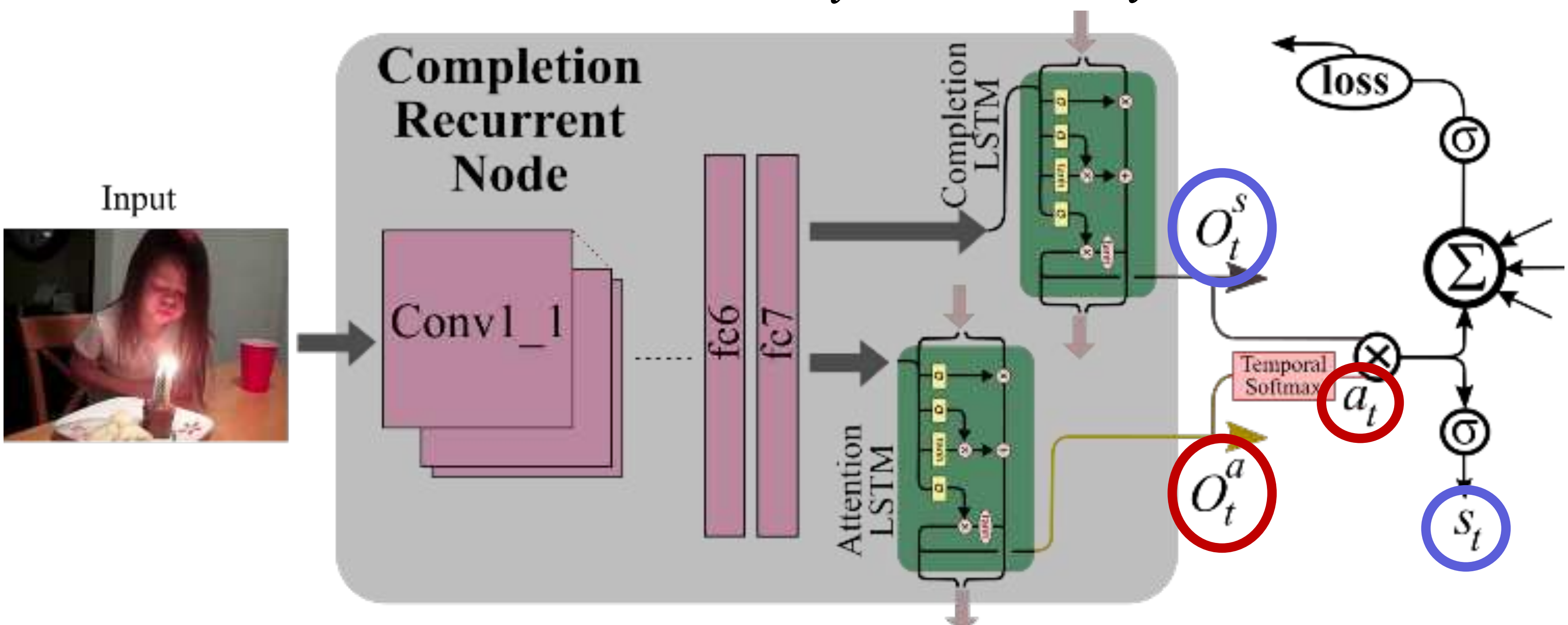
We detect completion using only **weak labels** during training.

sequence-level ***complete*** and ***incomplete*** labels

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

$$\begin{cases} \text{Attention LSTM} \rightarrow o_t^a \xrightarrow{\text{Softmax}} a_t \\ \text{Completion LSTM} \rightarrow o_t^s \xrightarrow{\sigma(o_t^s \times a_t)} s_t \end{cases}$$



F Heidarivincheh, M Mirmehdi, D Damen (2019). Weakly-Supervised Completion Moment Detection using Temporal Attention. ICCV Workshop on Human Behaviour Understanding

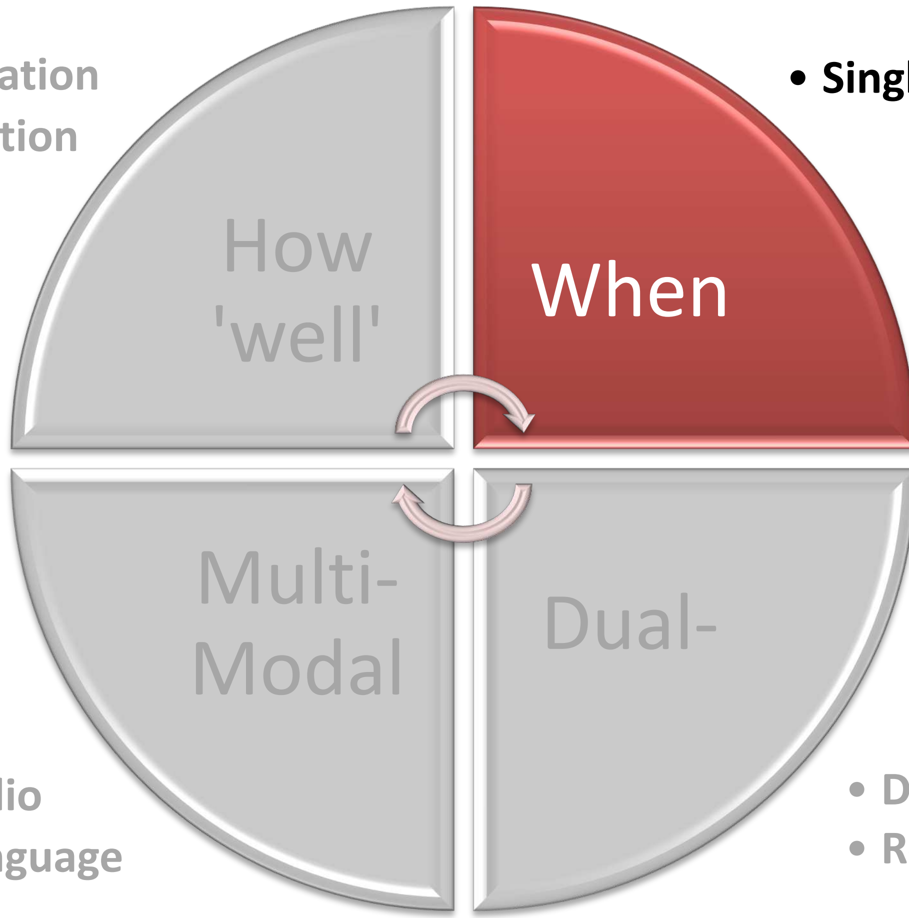
with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

Fine-Grained Object Interactions

- Skill Determination
- Action Completion

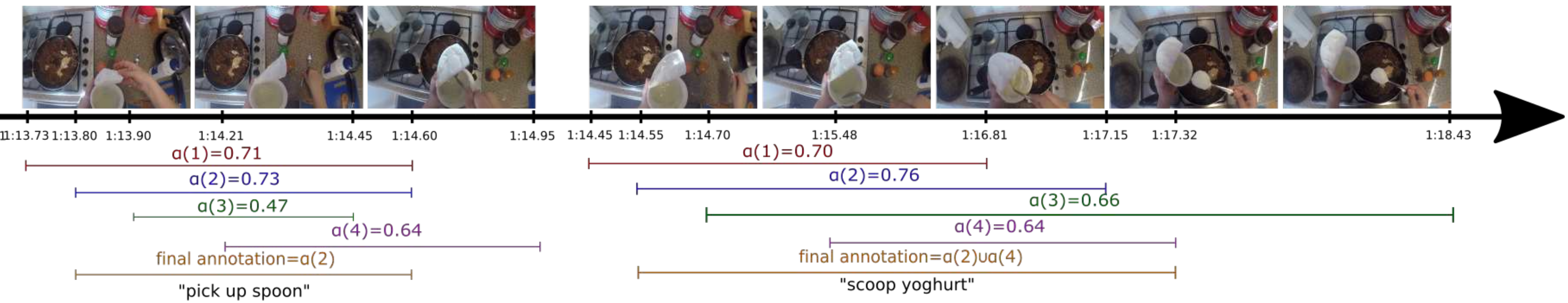
- Single-timestamp



- Vision+Audio
- Vision+Language

- DDLSTM
- Retro-Actions

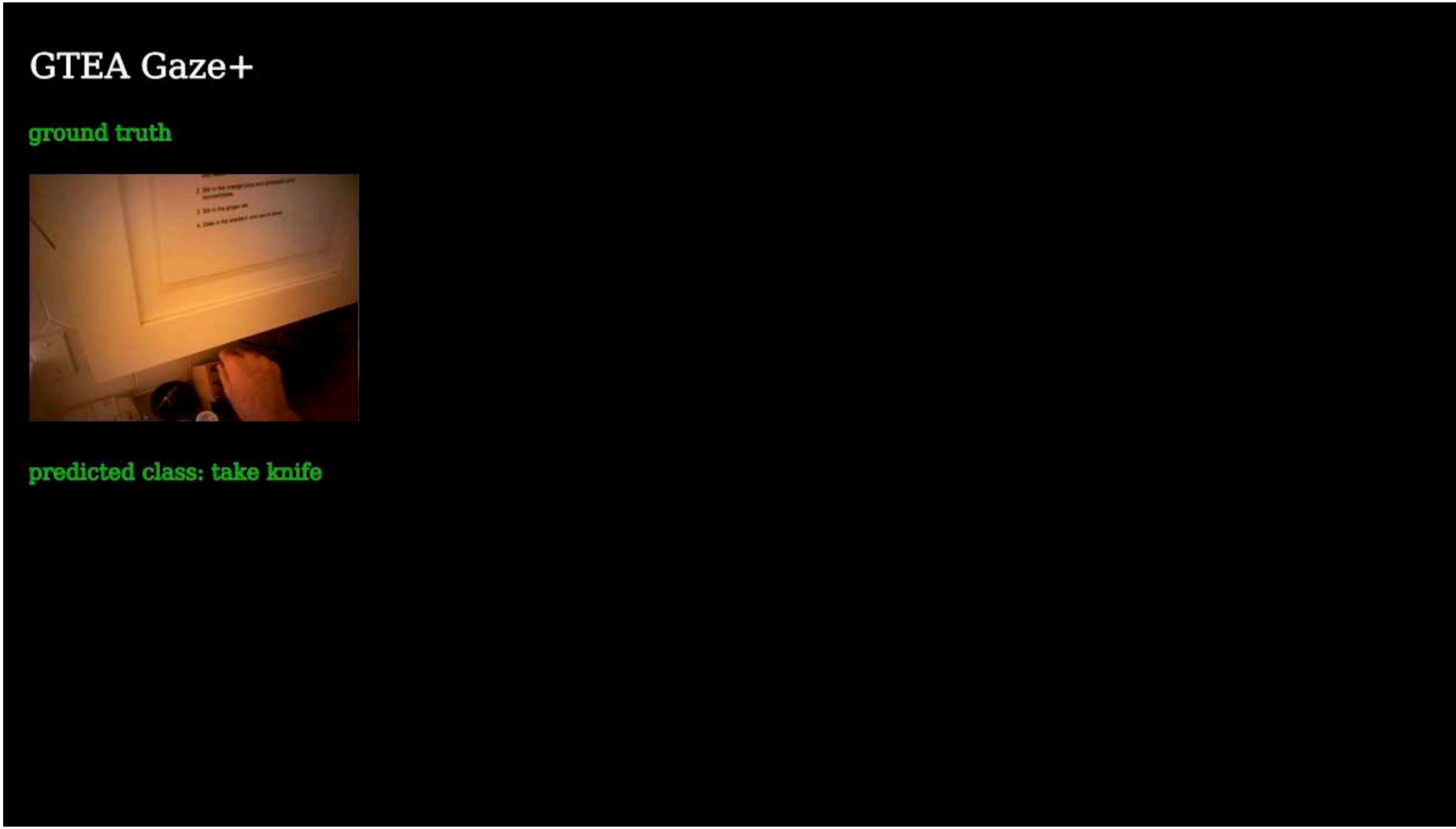
Temporal Boundaries for Object Interactions



- How robust are current state-of-the-art approaches to annotated boundaries in test segments?
- Modify test segment boundaries, maintaining significant overlap of segments IoU > 0.5
- **Correct in Green** – **Incorrect in Red**

with: Davide Moltisanti
Michael Wray

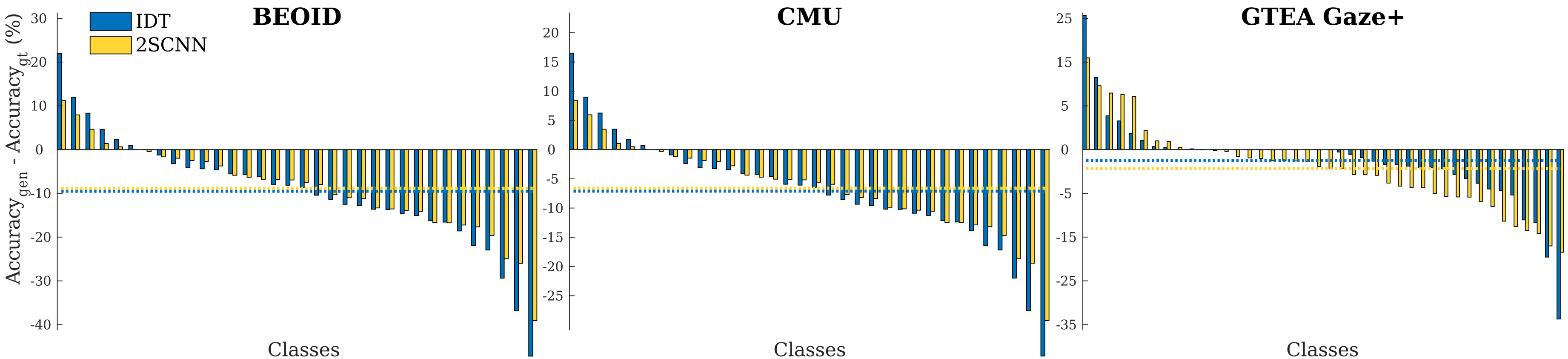
Trespassing the Boundaries



D Moltisanti et al (2017). Trespassing the Boundaries: Labeling Temporal Bounds for Object Interactions in Egocentric Video. *International Conference on Computer Vision (ICCV)*

with: Davide Moltisanti
Michael Wray

Trespassing the Boundaries



D Moltisanti et al (2017). Trespassing the Boundaries: Labeling Temporal Bounds for Object Interactions in Egocentric Video. *International Conference on Computer Vision (ICCV)*

Action Recognition from a Single Timestamp

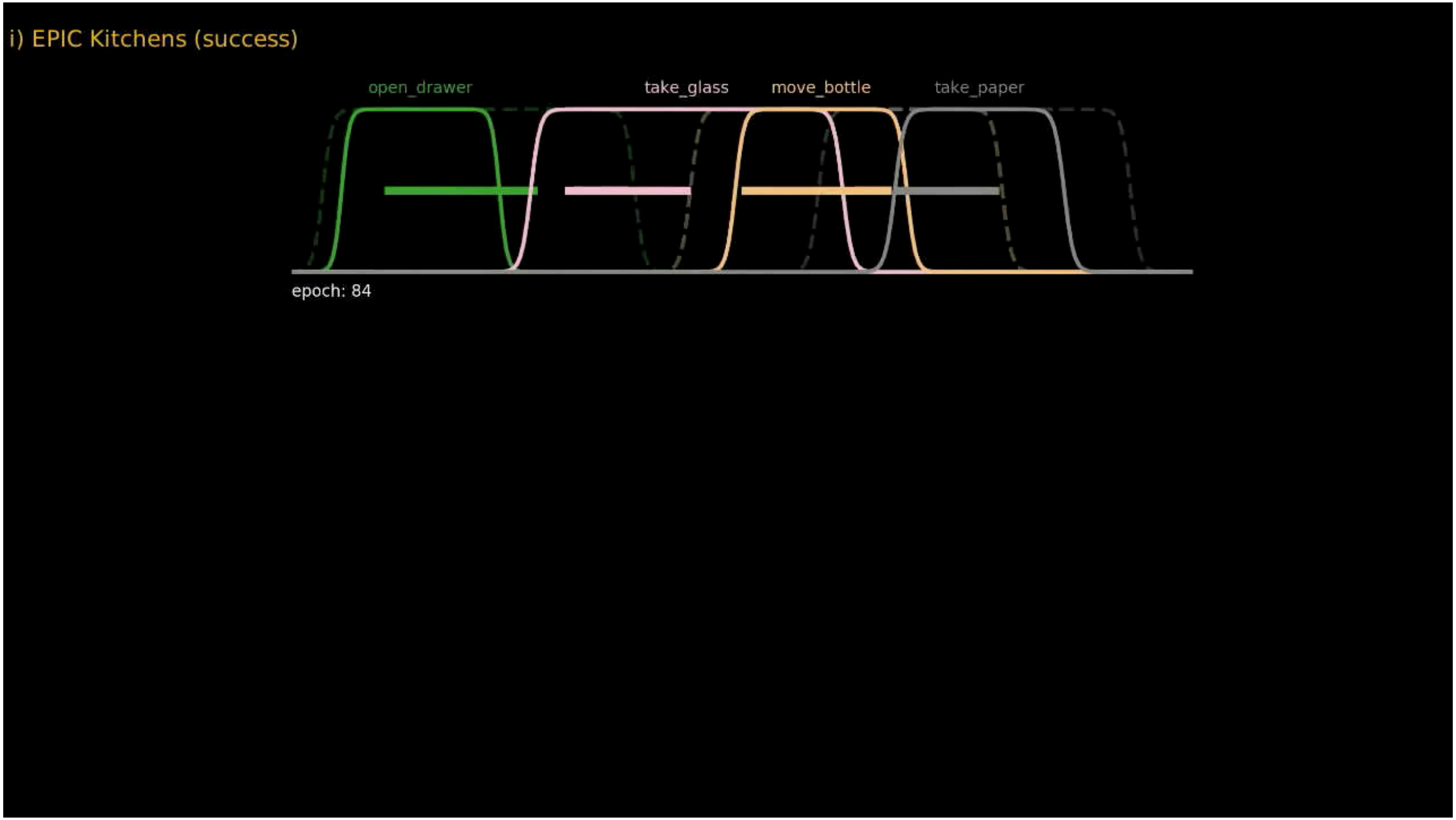
with: Davide Moltisanti
Sanja Fidler

- Learning from Single timestamps

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

Action Recognition from a Single Timestamp

with: Davide Moltisanti
Sanja Fidler

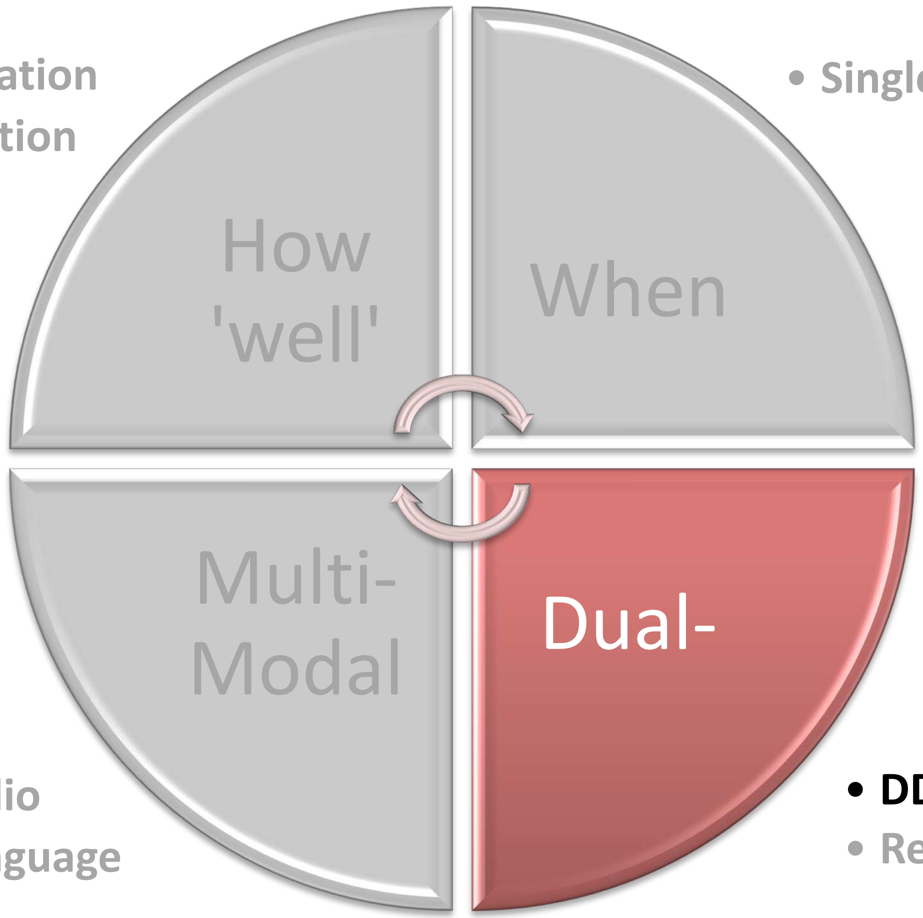


D Moltisanti et al (2019). Action Recognition from Single Timestamp Supervision in Untrimmed Videos. *Computer Vision and Pattern Recognition (CVPR)*

Fine-Grained Object Interactions

- Skill Determination
- Action Completion

- Single-timestamp



- Vision+Audio
- Vision+Language

- **DDLSTM**
- Retro-Actions

Dual-Domain LSTM for Cross-Dataset Action Recognition

with: Toby Perrett

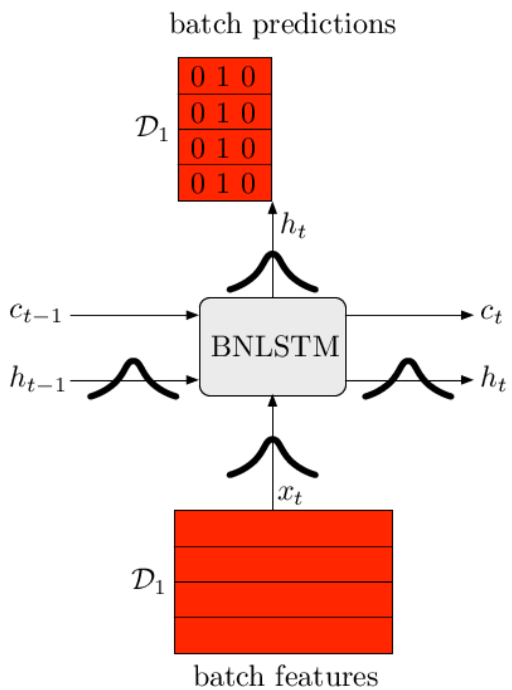
T Perrett, D Damen (2019). DDLSTM: Dual-Domain LSTM for Cross-Dataset Action Recognition. CVPR

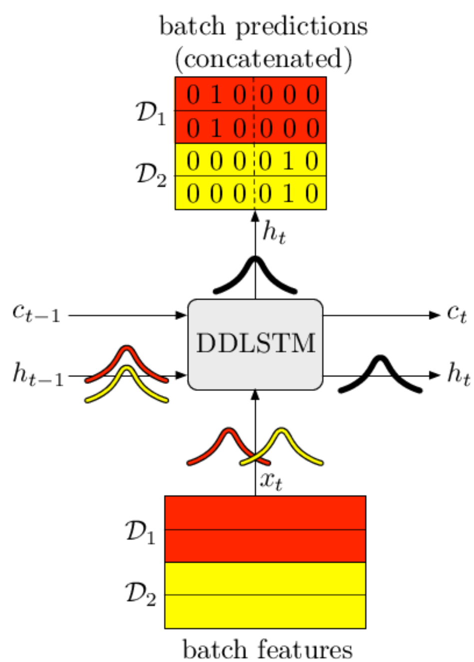
Dual-Domain LSTM for Cross-Dataset Action Recognition

with: Toby Perrett

BNLSTM
1 dataset

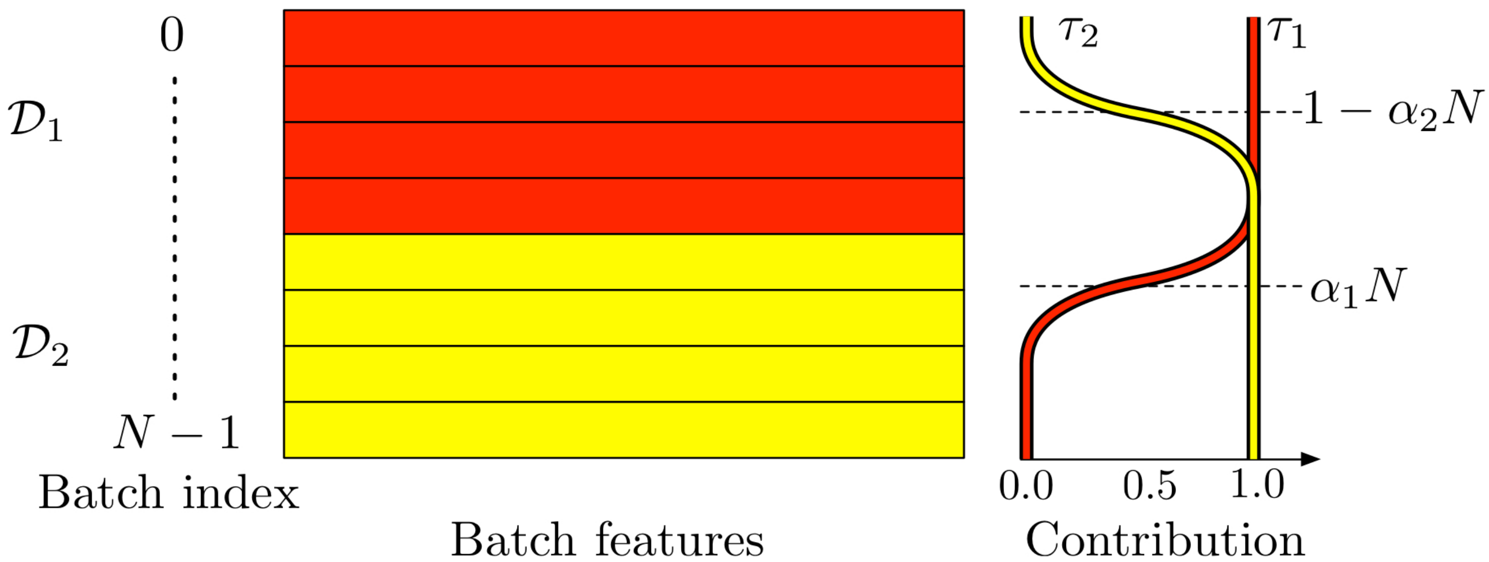
DDLSTM
2 datasets





Dual-Domain LSTM for Cross-Dataset Action Recognition

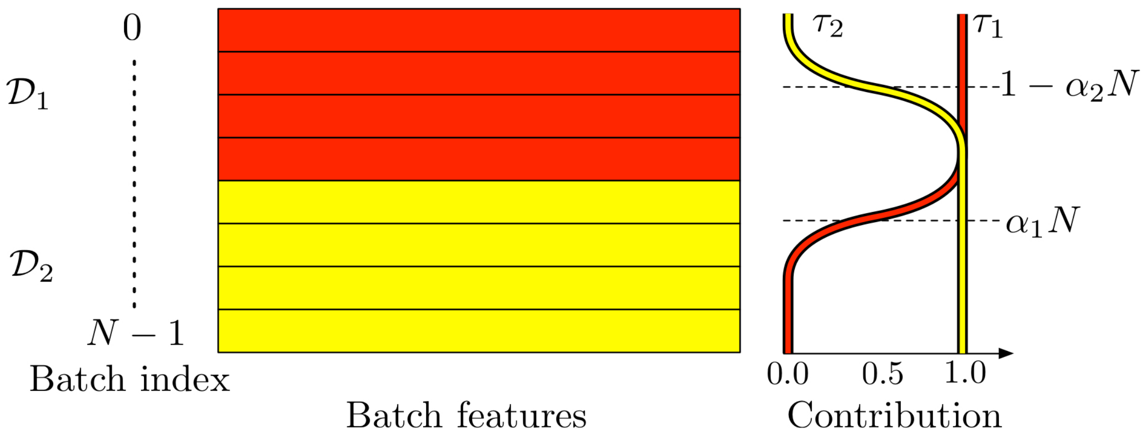
with: Toby Perrett



T Perrett, D Damen (2019). DDLSTM: Dual-Domain LSTM for Cross-Dataset Action Recognition. CVPR

Dual-Domain LSTM for Cross-Dataset Action Recognition

with: Toby Perrett

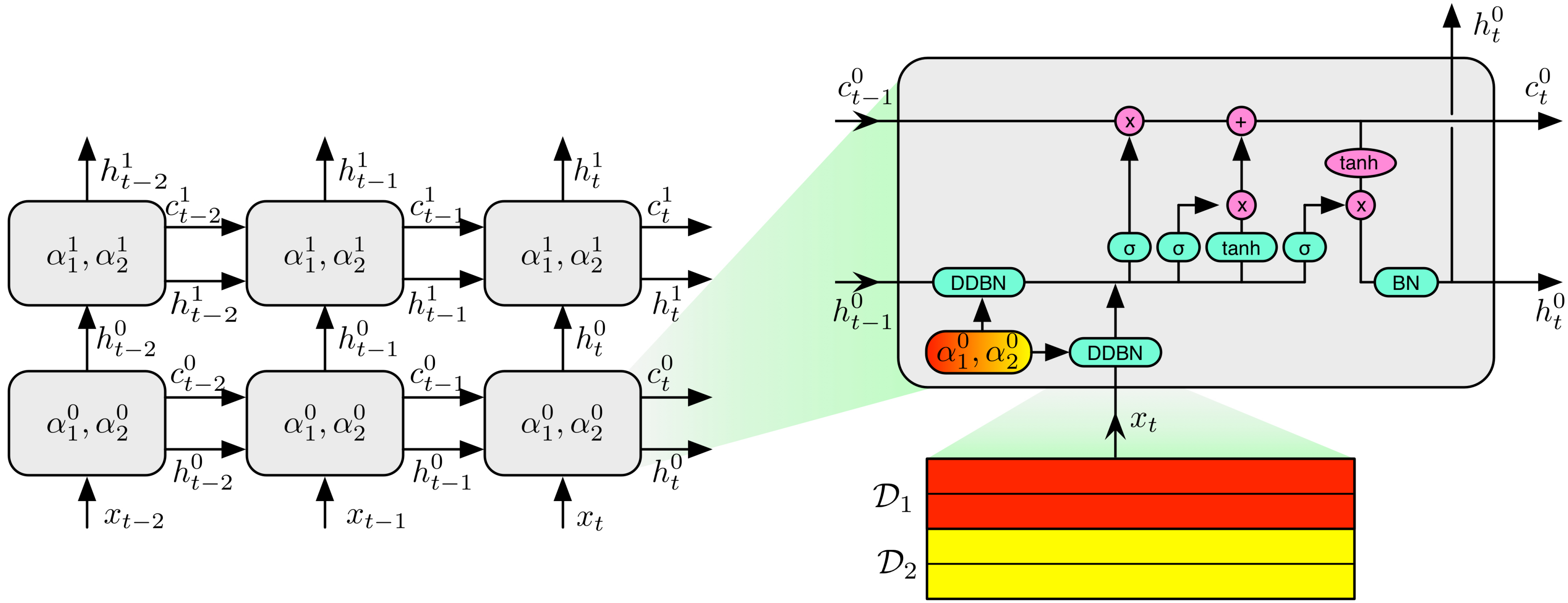


$$\tau_1(\alpha_1, j) = \frac{1 - \tanh(j - \alpha_1 N)}{2}$$

$$\tau_2(\alpha_2, j) = \frac{1 + \tanh(j - \alpha_2 N)}{2}$$

Dual-Domain LSTM for Cross-Dataset Action Recognition

with: Toby Perrett

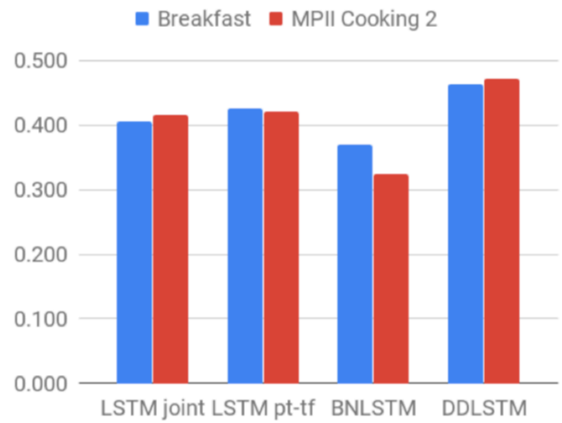


T Perrett, D Damen (2019). DDLSTM: Dual-Domain LSTM for Cross-Dataset Action Recognition. CVPR

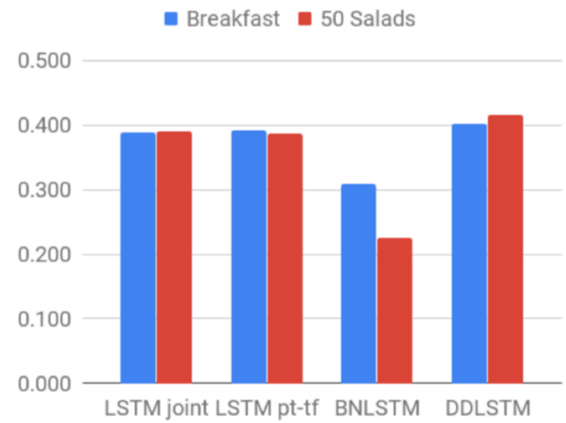
Dual-Domain LSTM for Cross-Dataset Action Recognition

with: Toby Perrett

(a) Breakfast



(b) 50 Salads



(c) MPII Cooking 2

T Perrett, D Damen (2019). DDLSTM: Dual-Domain LSTM for Cross-Dataset Action Recognition. CVPR

Dual-Domain LSTM for Cross-Dataset Action Recognition

with: Toby Perrett

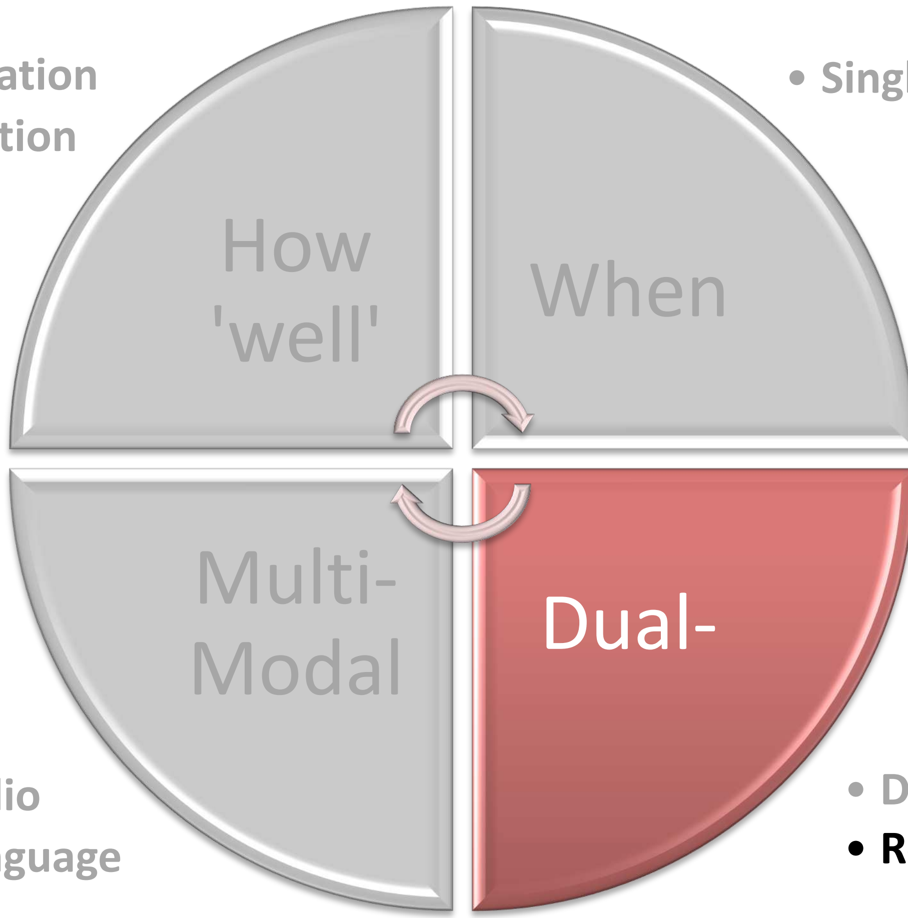
D1	D2	Training	LSTM Type	D1 Acc	D2 Acc
ActivityNet	50 Salads	Pt/ft	LSTM	44.4	42.1
ActivityNet	50 Salads	Joint	DDLSTM	44.3	42.2
Thumos	50 Salads	Pt/ft	LSTM	65.9	42.0
Thumos	50 Salads	Joint	DDLSTM	66.1	42.3
EPIC	50 Salads	Pt/ft	LSTM	31.5	**44.9**
EPIC	50 Salads	Joint	DDLSTM	33.1	**48.9**

T Perrett, D Damen (2019). DDLSTM: Dual-Domain LSTM for Cross-Dataset Action Recognition. CVPR

Fine-Grained Object Interactions

- Skill Determination
- Action Completion

- Single-timestamp



- Vision+Audio
- Vision+Language

- DDLSTM
- **Retro-Actions**

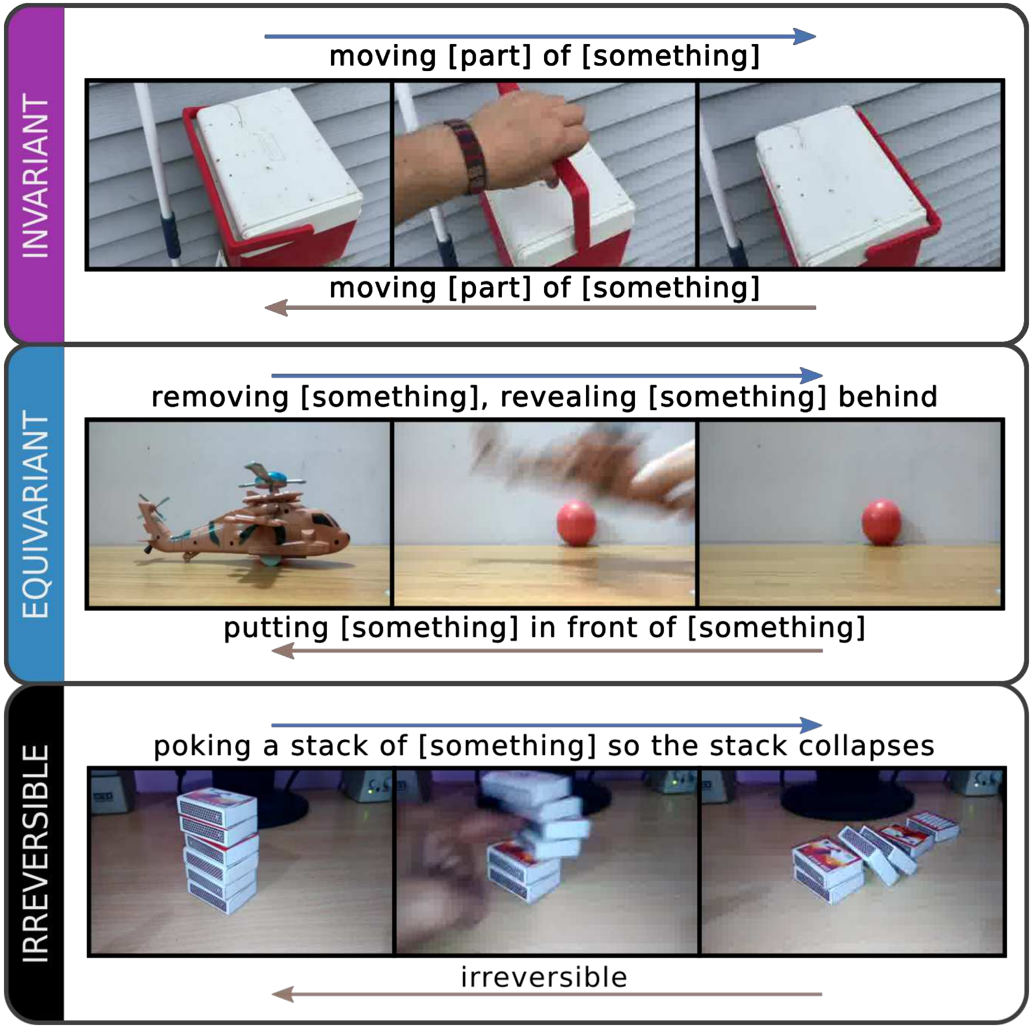
with: Will Price

Retro-actions

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

with: Will Price

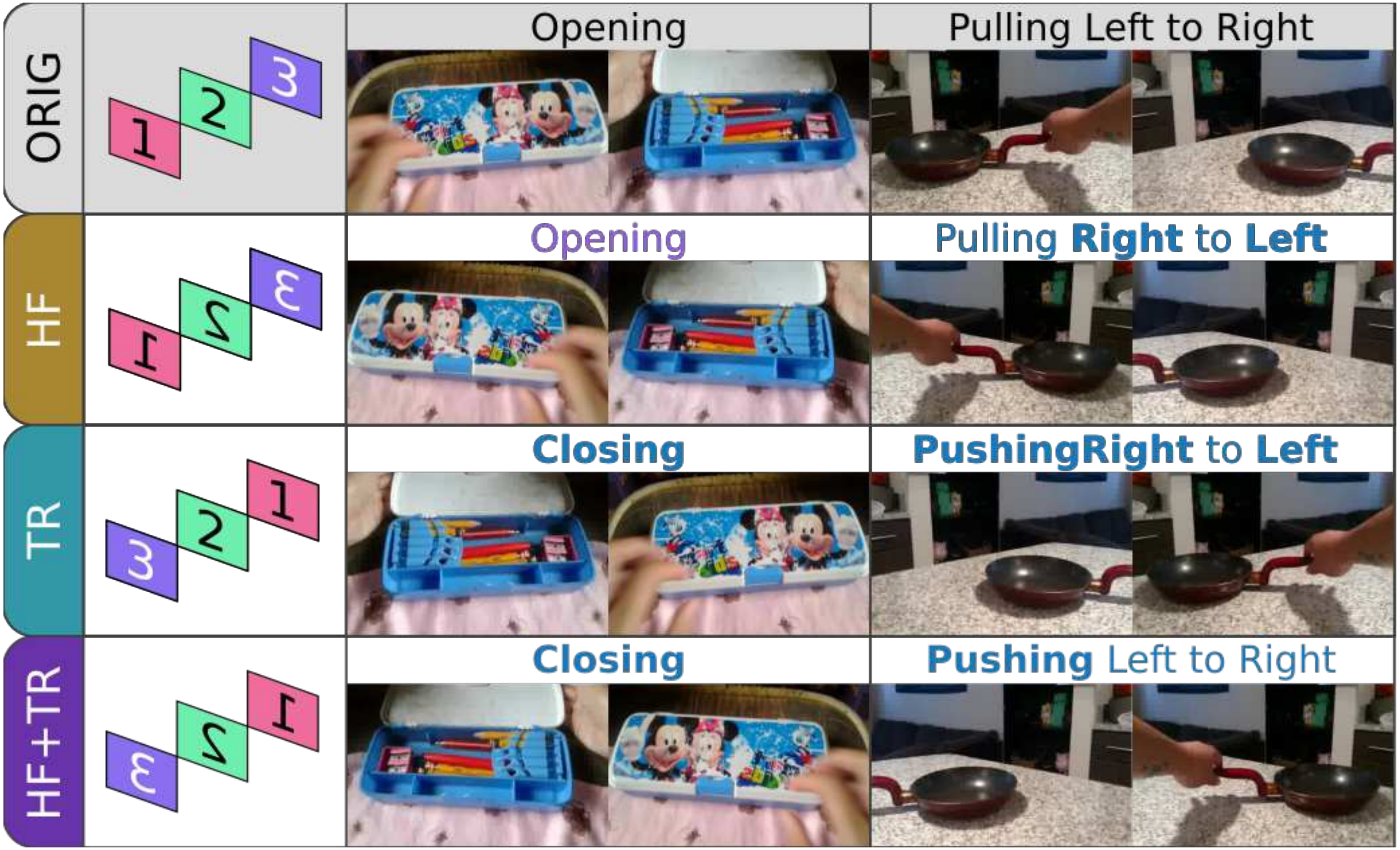
Retro-actions



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

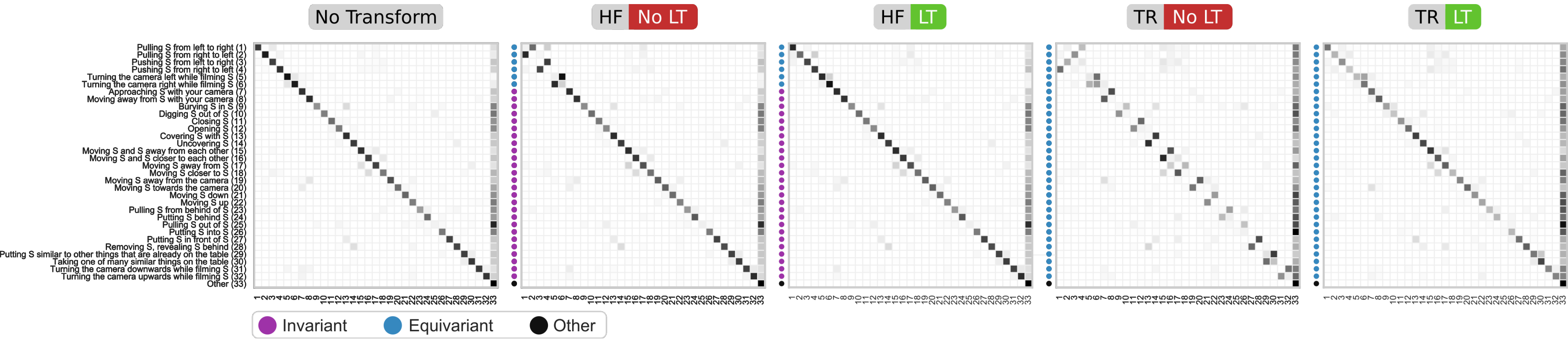
with: Will Price

Retro-actions



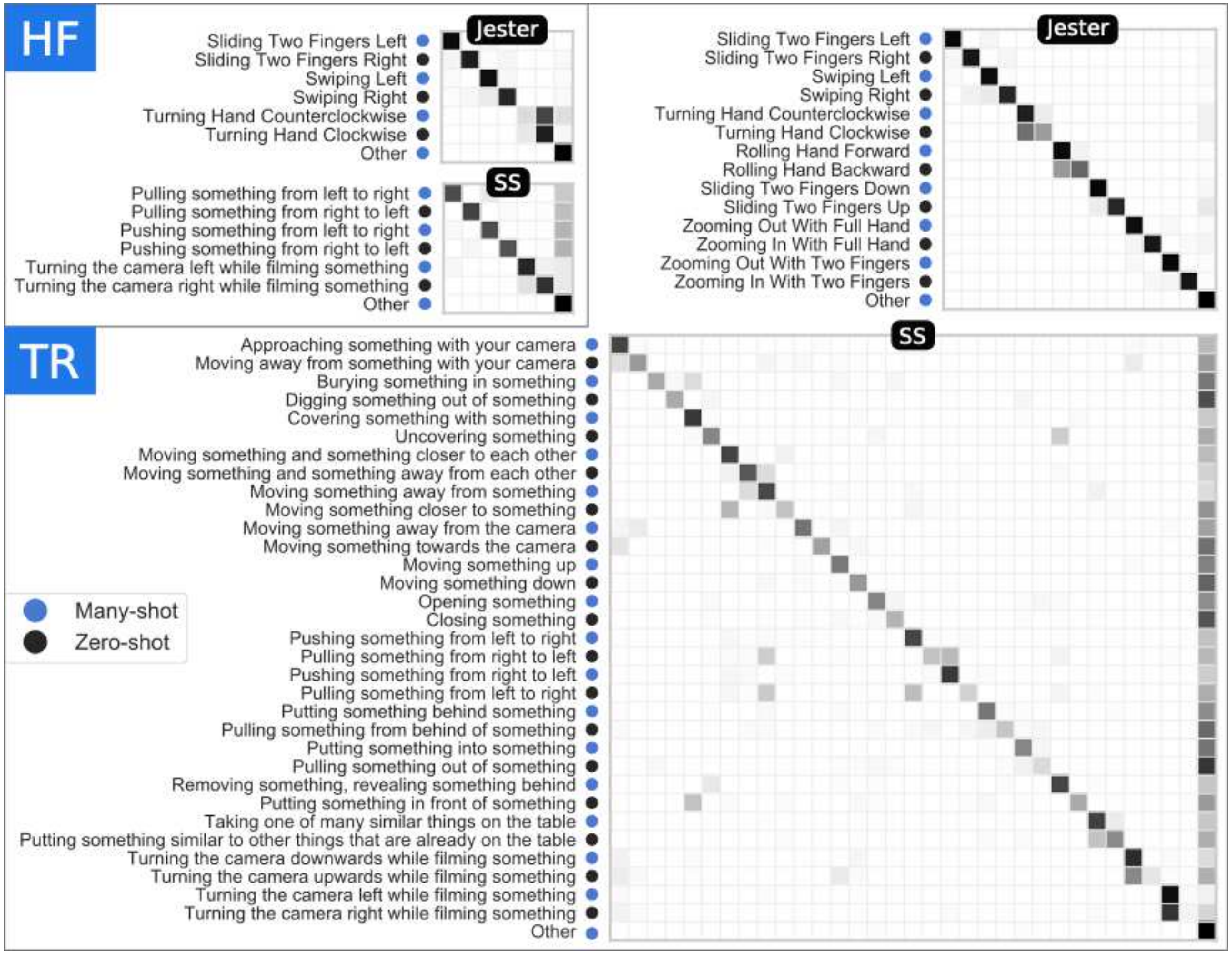
with: Will Price

Retro-actions



with: Will Price

Retro-actions – Zero-Shot Learning

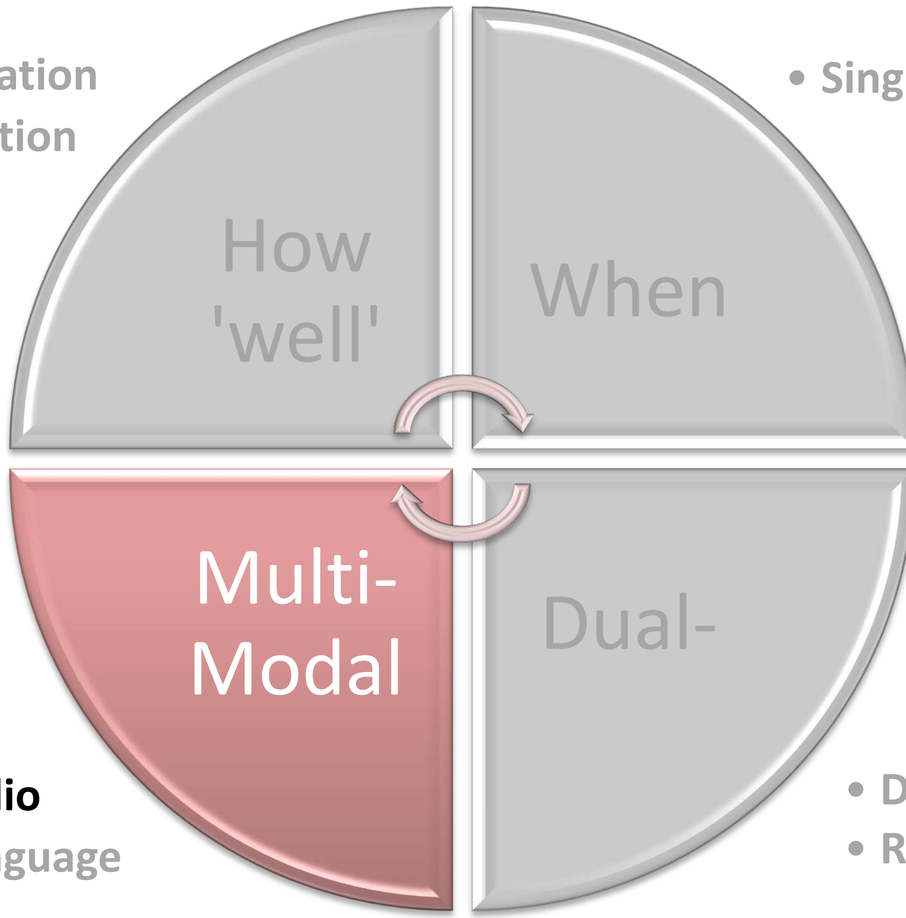


W Price, D Damen (2019). Retro-Actions: Learning 'Close' by Time-Reversing 'Open' Videos. ICCV MDALC Workshop

Fine-Grained Object Interactions

- Skill Determination
- Action Completion

- Single-timestamp

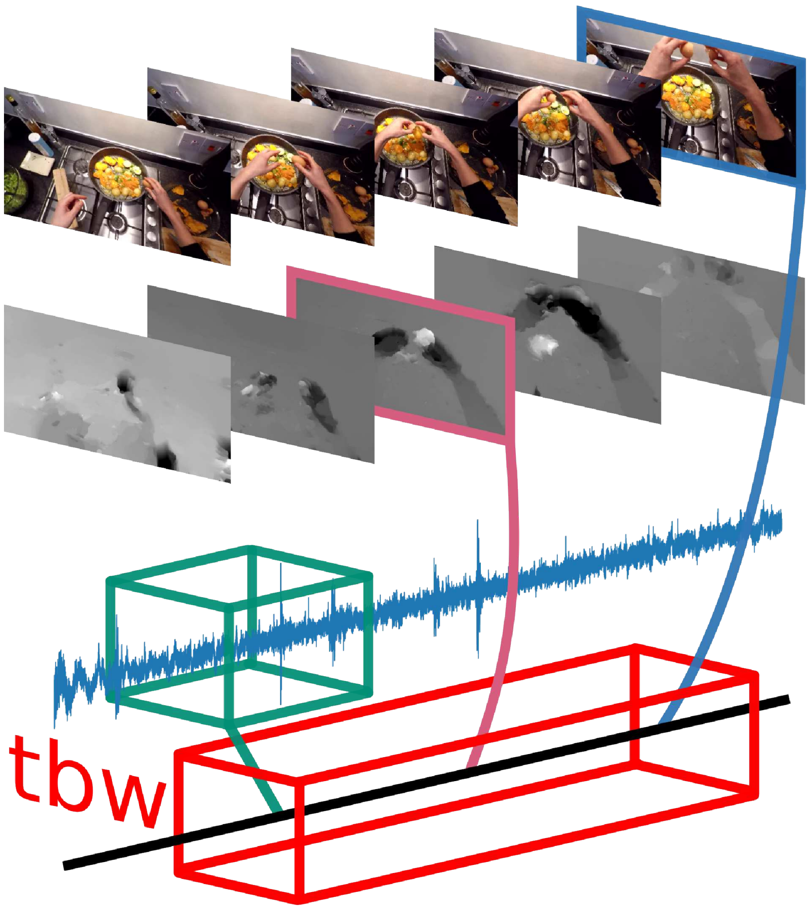


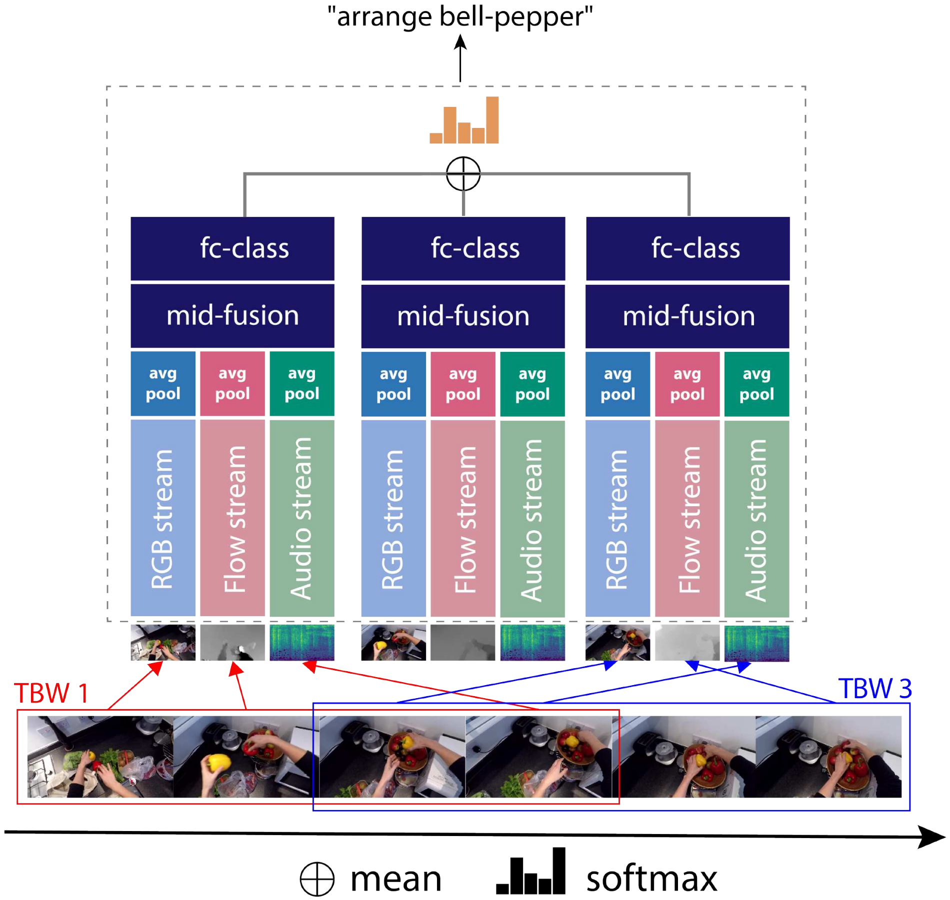
- **Vision+Audio**
- Vision+Language

- DDLSTM
- Retro-Actions

Audio-Visual Temporal Binding for Egocentric Action Recognition

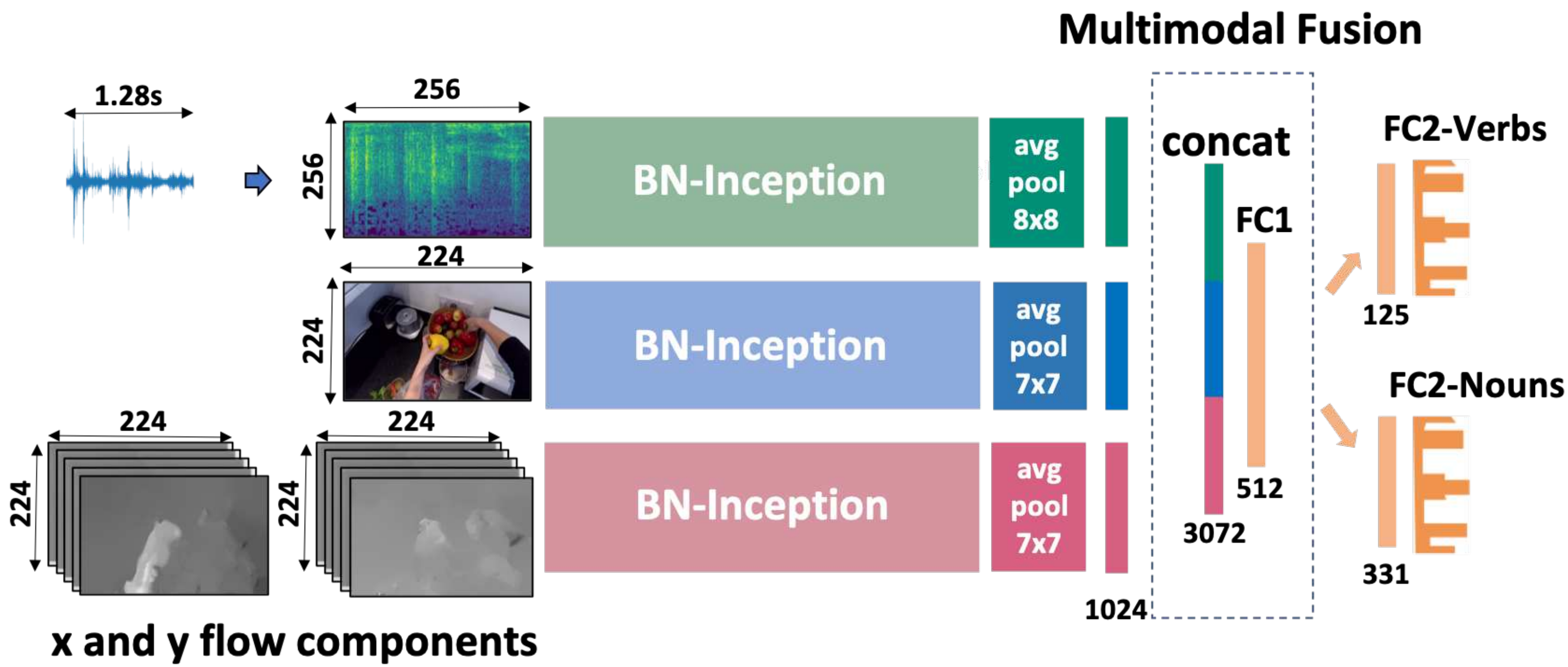
with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman





Audio-Visual Temporal Binding for Egocentric Action Recognition

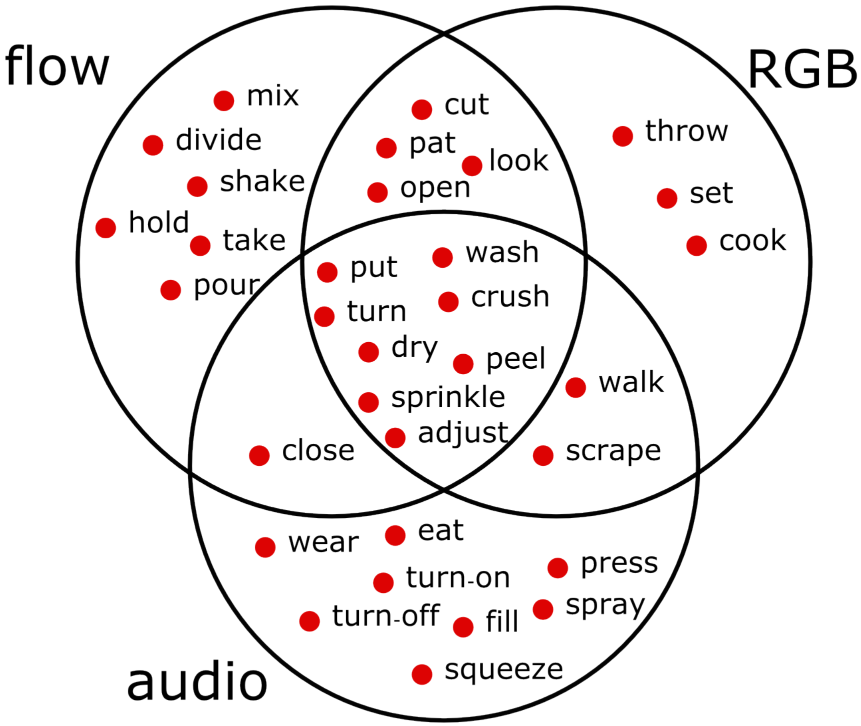
with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



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

Audio-Visual Temporal Binding for Egocentric Action Recognition

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



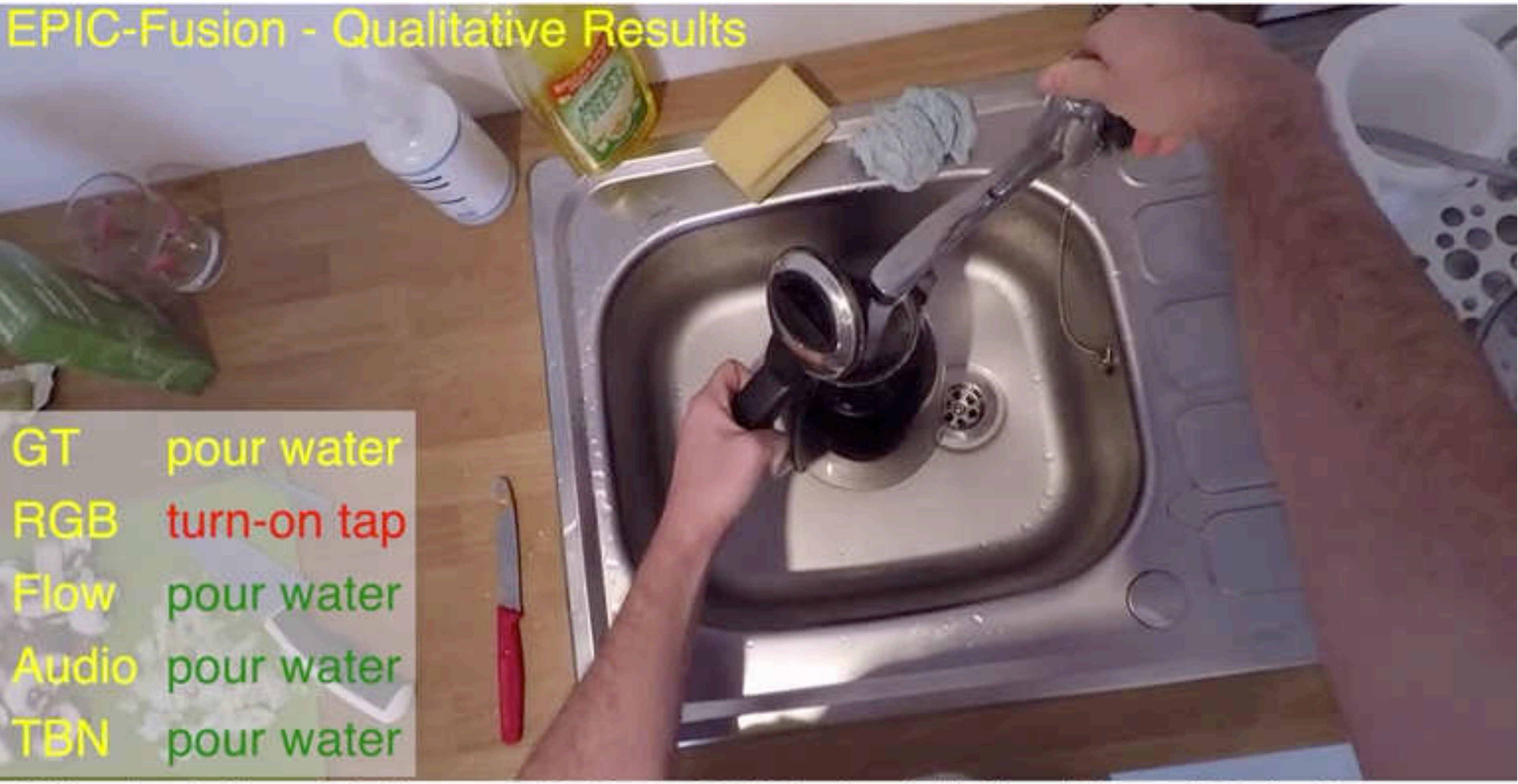
Audio-Visual Temporal Binding for Egocentric Action Recognition

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**

Audio-Visual Temporal Binding for Egocentric Action Recognition

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



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

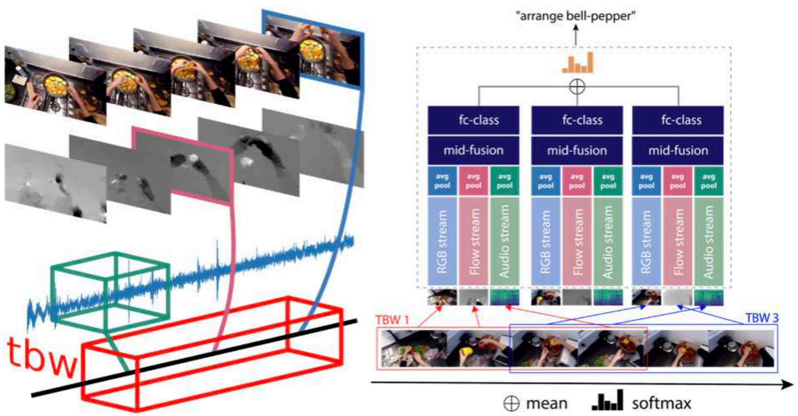
Audio-Visual Temporal Binding for Egocentric Action Recognition

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman

EPIC-Fusion: Audio-Visual Temporal Binding for Egocentric Action Recognition

Evangelos Kazakos[1], Arsha Nagrani[2], Andrew Zisserman[2] and Dima Damen[1]

[1]University of Bristol, VIL, [2]University of Oxford, VGG



Abstract

We focus on multi-modal fusion for egocentric action recognition, and propose a novel architecture for multi-modal temporal-binding, i.e. the combination of modalities within a range of temporal offsets. We train the architecture with three modalities – RGB, Flow and Audio – and combine them with mid-level fusion alongside sparse temporal sampling of fused representations. In contrast with previous works, modalities are fused before temporal aggregation, with shared modality and fusion weights over time. Our proposed architecture is trained end-to-end, outperforming individual modalities as well as late-fusion of modalities.

We demonstrate the importance of audio in egocentric vision, on per-class basis, for identifying actions as well as interacting objects. Our method achieves state of the art results on both the seen and unseen test sets of the largest egocentric dataset: EPIC-Kitchens, on all metrics using the public leaderboard.

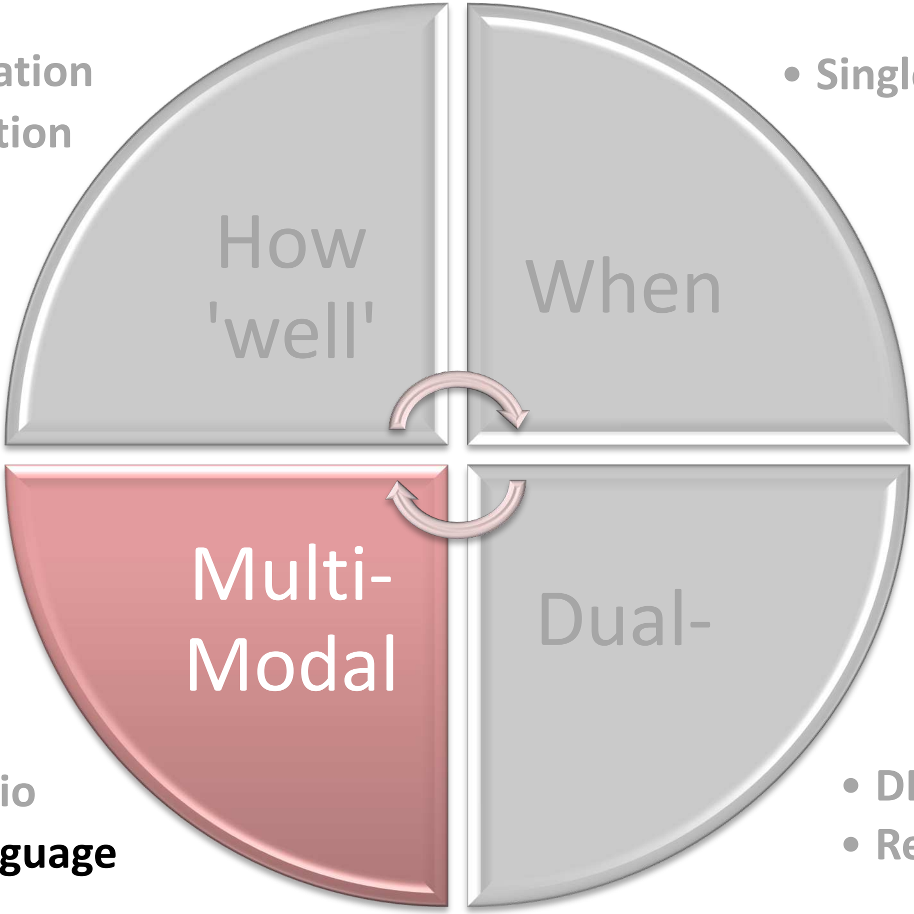
Downloads

- Paper [ArXiv]
- Code and models [GitHub]

Fine-Grained Object Interactions

- Skill Determination
- Action Completion

- Single-timestamp



- Vision+Audio
- **Vision+Language**

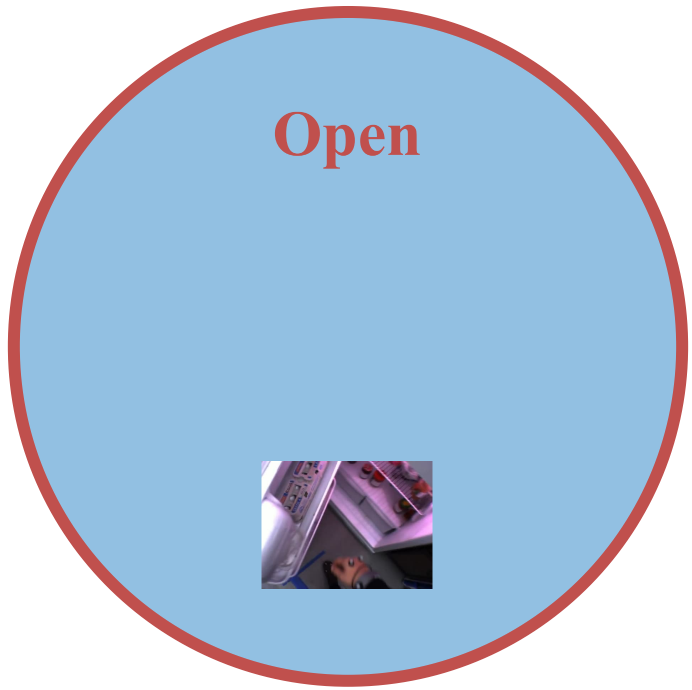
- DDLSTM
- Retro-Actions

with: Michael Wray

The *Verbs* Dilemma

with: Michael Wray

The *Verbs* Dilemma



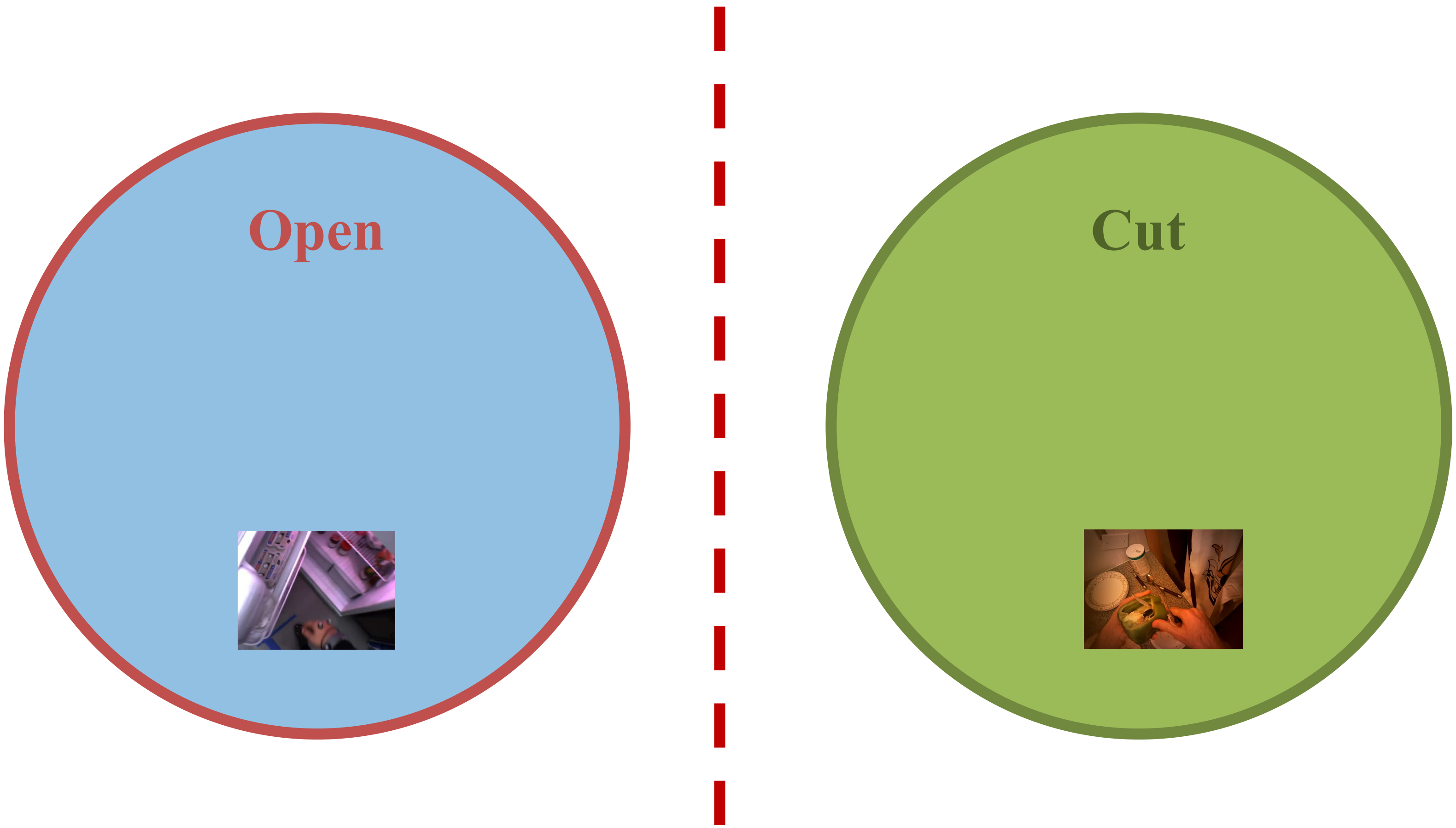
with: Michael Wray

The *Verbs* Dilemma

M Wray and D Damen (2019). Learning Visual Actions Using Multiple Verb-Only Labels. BMVC

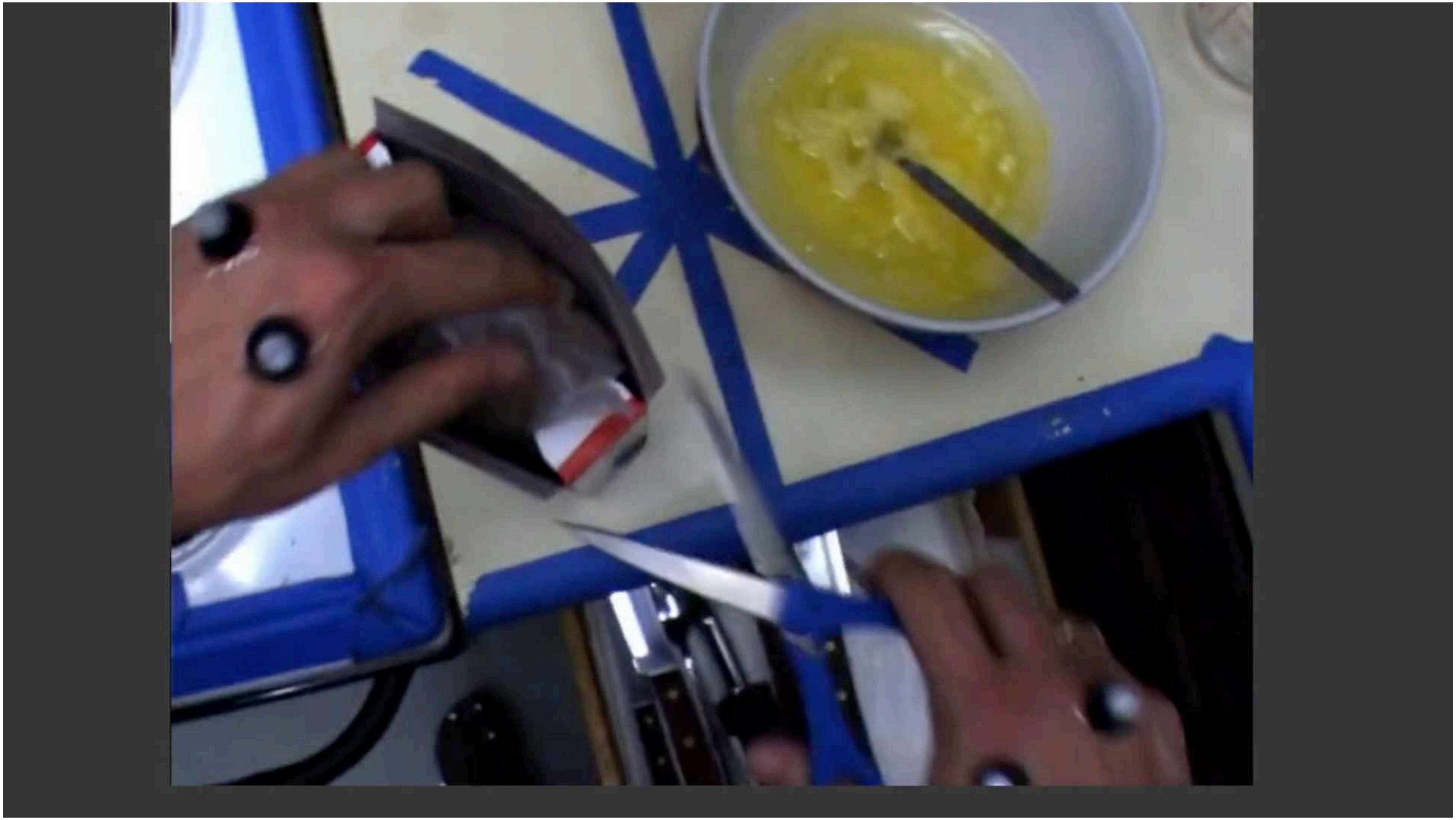
with: Michael Wray

The *Verbs* Dilemma



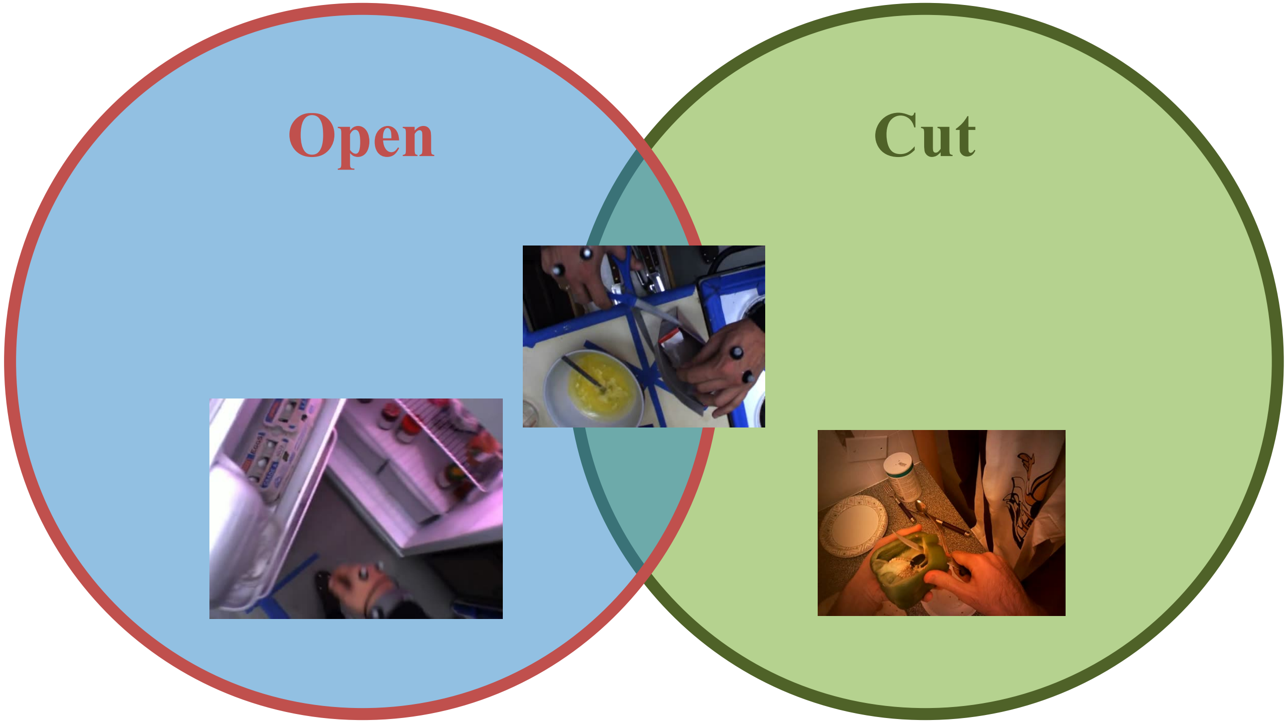
with: Michael Wray

The *Verbs* Dilemma



M Wray and D Damen (2019). Learning Visual Actions Using Multiple Verb-Only Labels. BMVC

The *Verbs* Dilemma



M Wray and D Damen (2019). Learning Visual Actions Using Multiple Verb-Only Labels. BMVC

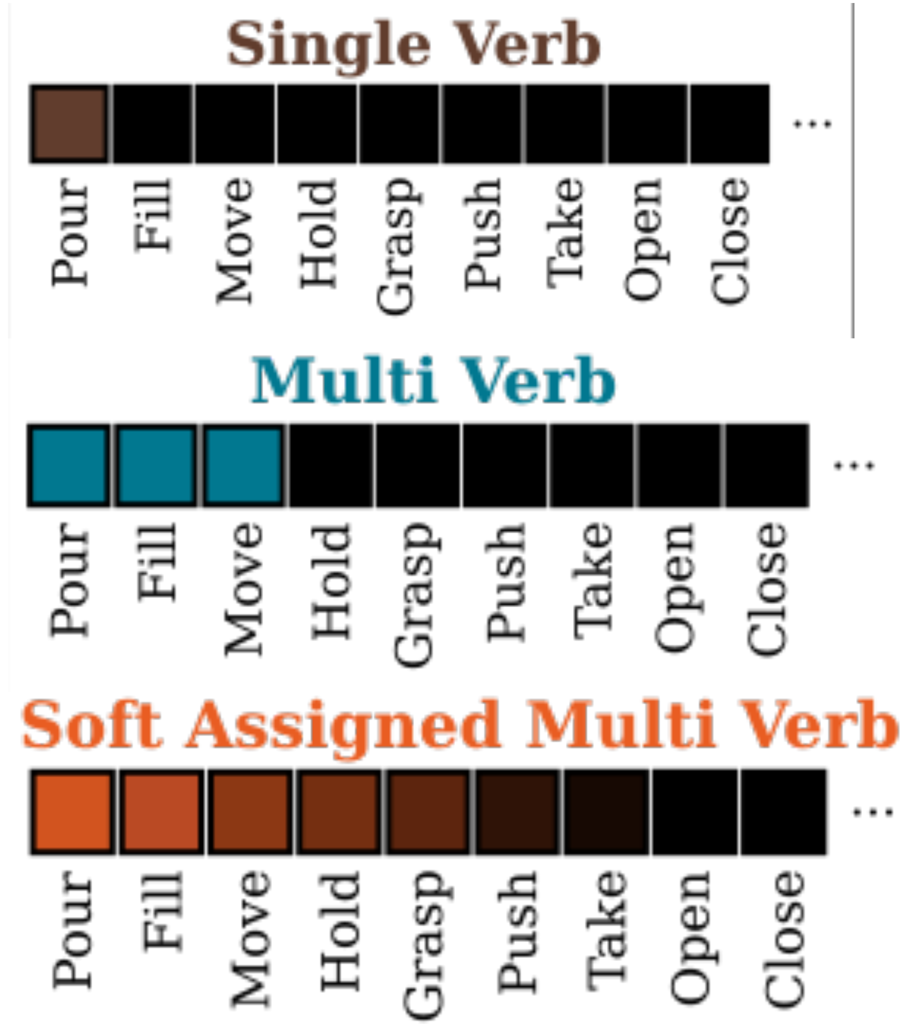
The *Verbs* Dilemma

- Action representations using a single verb is highly-ambiguous
  - Solution1: pre-selected non-overlapping verbs (SL)
    - run, walk, open, close
  - Solution2: Using nouns to disambiguate actions (V-N)
    - open-drawer, open-bottle, open-fridge
    - actions constrained to known nouns
  - Solution3: Multi-verb labels (ML, SAML)
    - open, hold, pull

M Wray and D Damen (2019). Learning Visual Actions Using Multiple Verb-Only Labels. BMVC

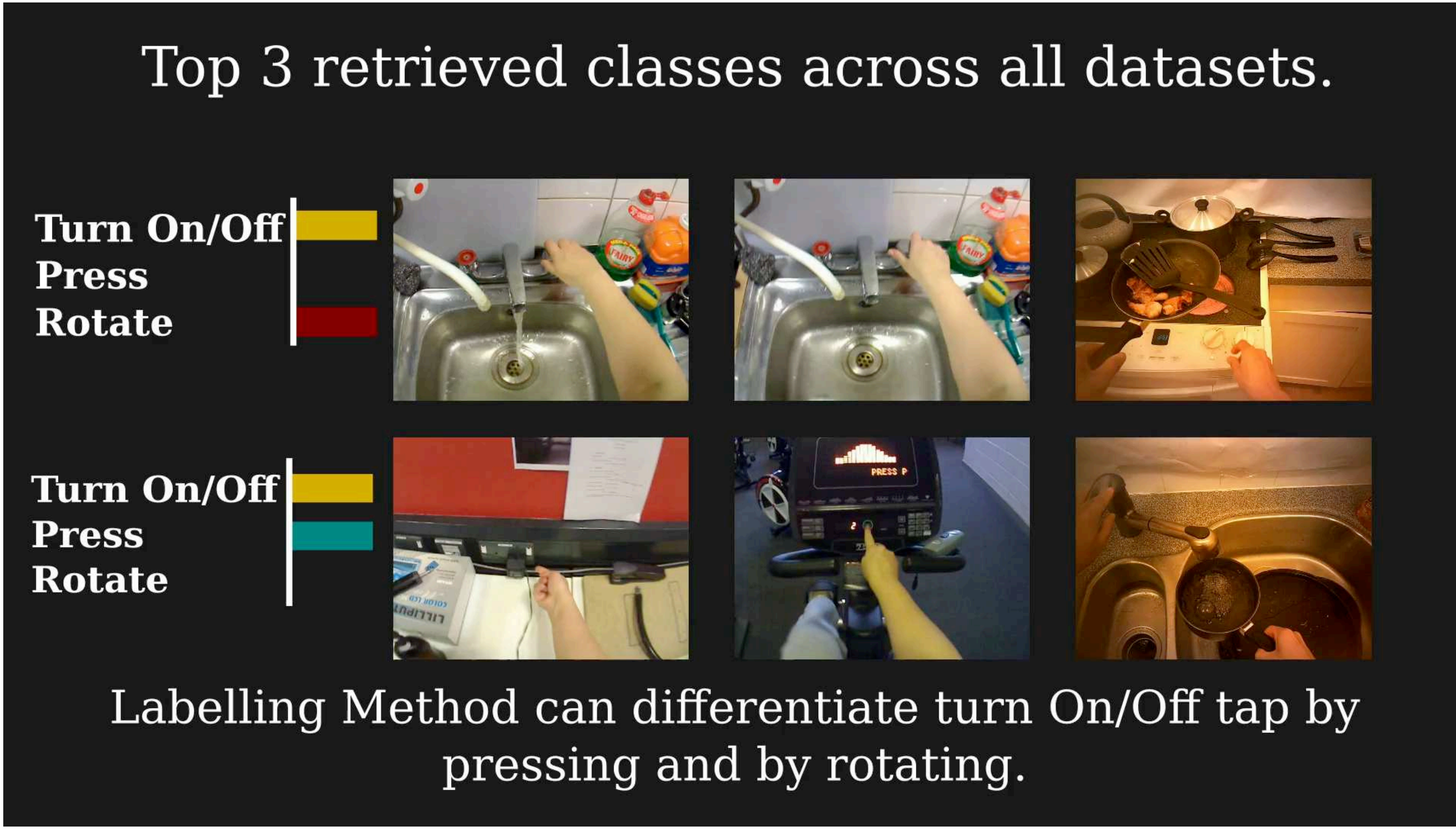
with: Michael Wray

The *Verbs* Dilemma



with: Michael Wray

The *Verbs* Dilemma



University of BRISTOL

M Wray and D Damen (2019). Learning Visual Actions Using Multiple Verb-Only Labels. BMVC

Fine-Grained Action Retrieval

with: Michael Wray
Gabriela Csurka
Diane Larlus

In this work we focus on
Fine-Grained Action Retrieval

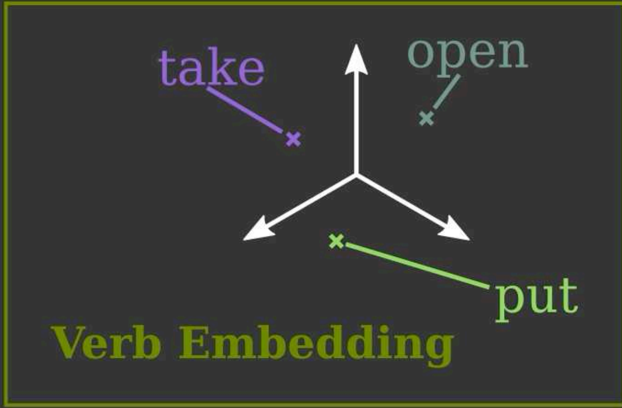
I put meat on a ball of dough

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

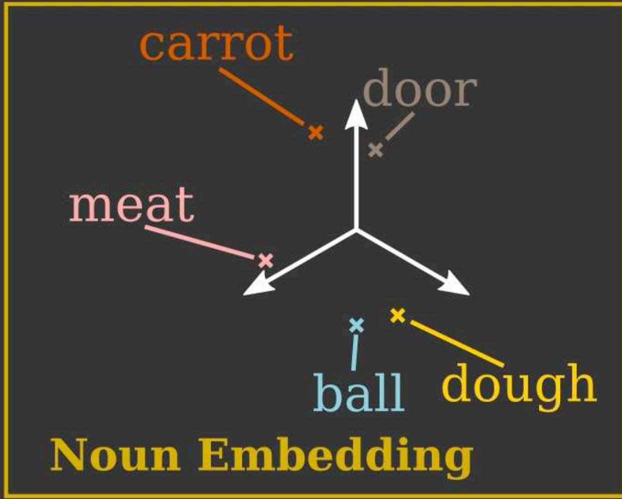
Fine-Grained Action Retrieval

with: Michael Wray
Gabriela Csurka
Diane Larlus

We embed the video and representations



[put]

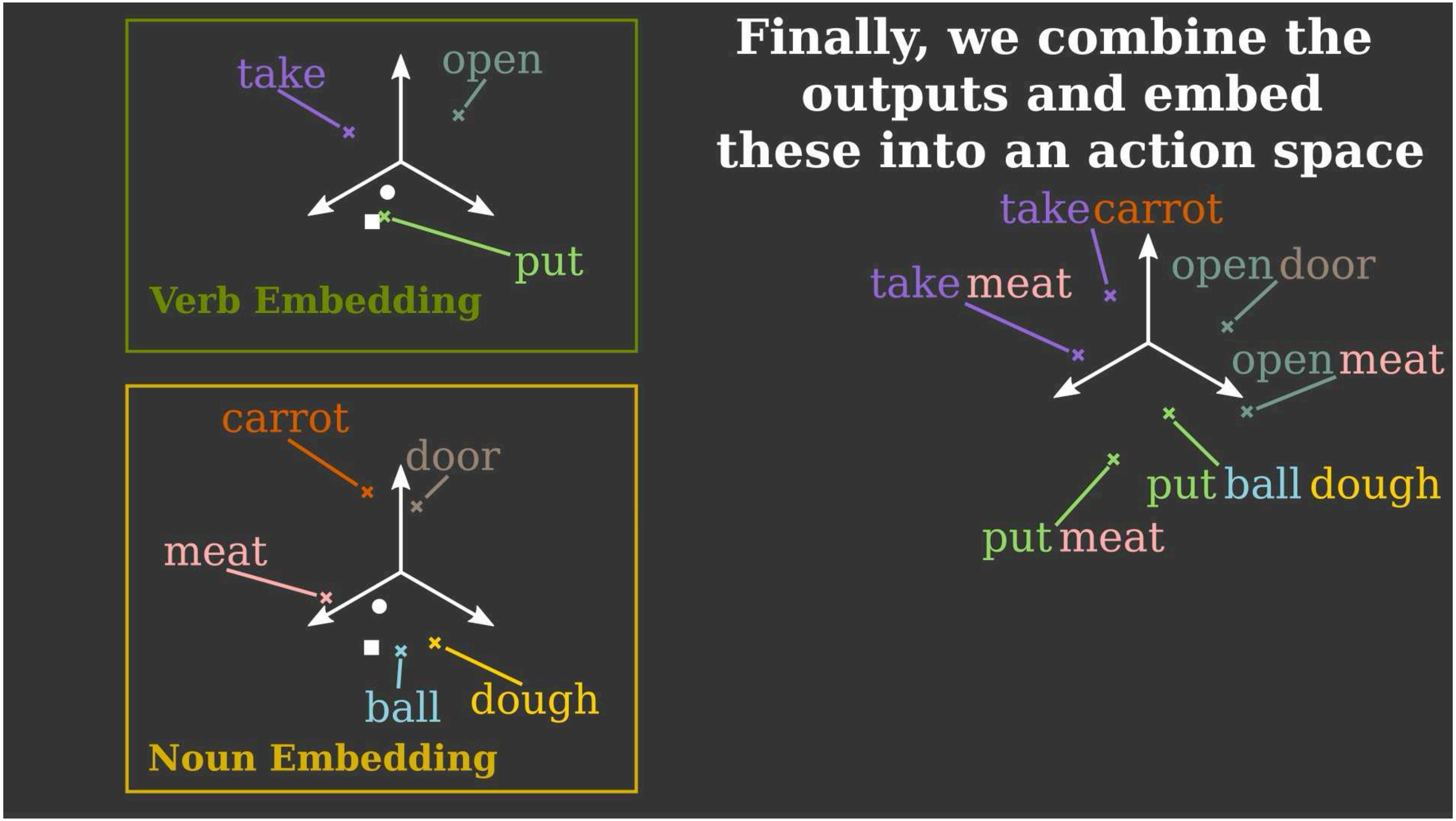


[meat, ball, dough]

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

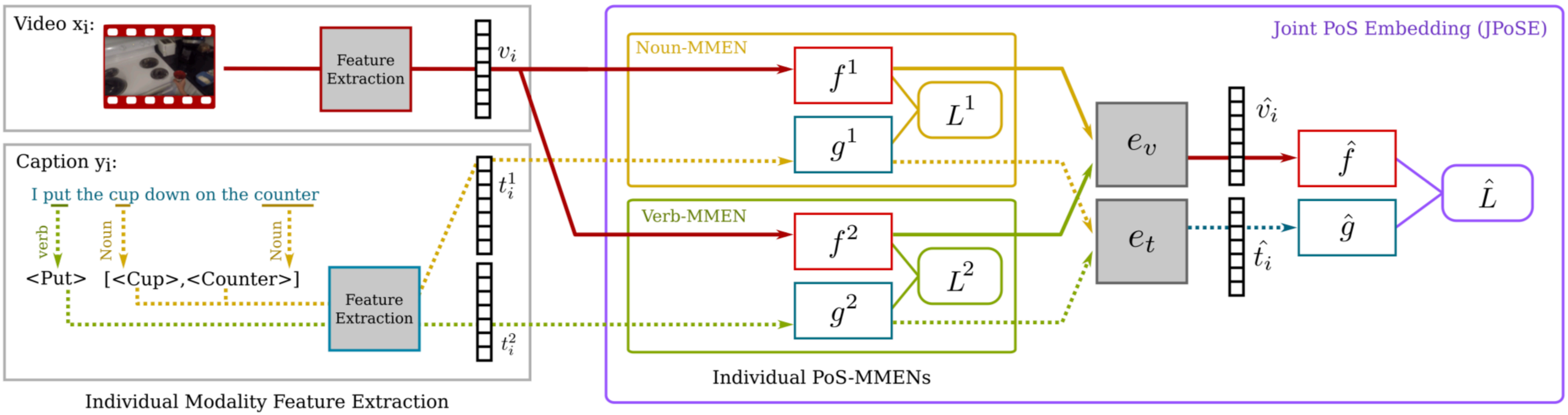
Fine-Grained Action Retrieval

with: Michael Wray
Gabriela Csurka
Diane Larlus



with: Michael Wray
Gabriela Csurka
Diane Larlus

Fine-Grained Action Retrieval



Fine-Grained Action Retrieval

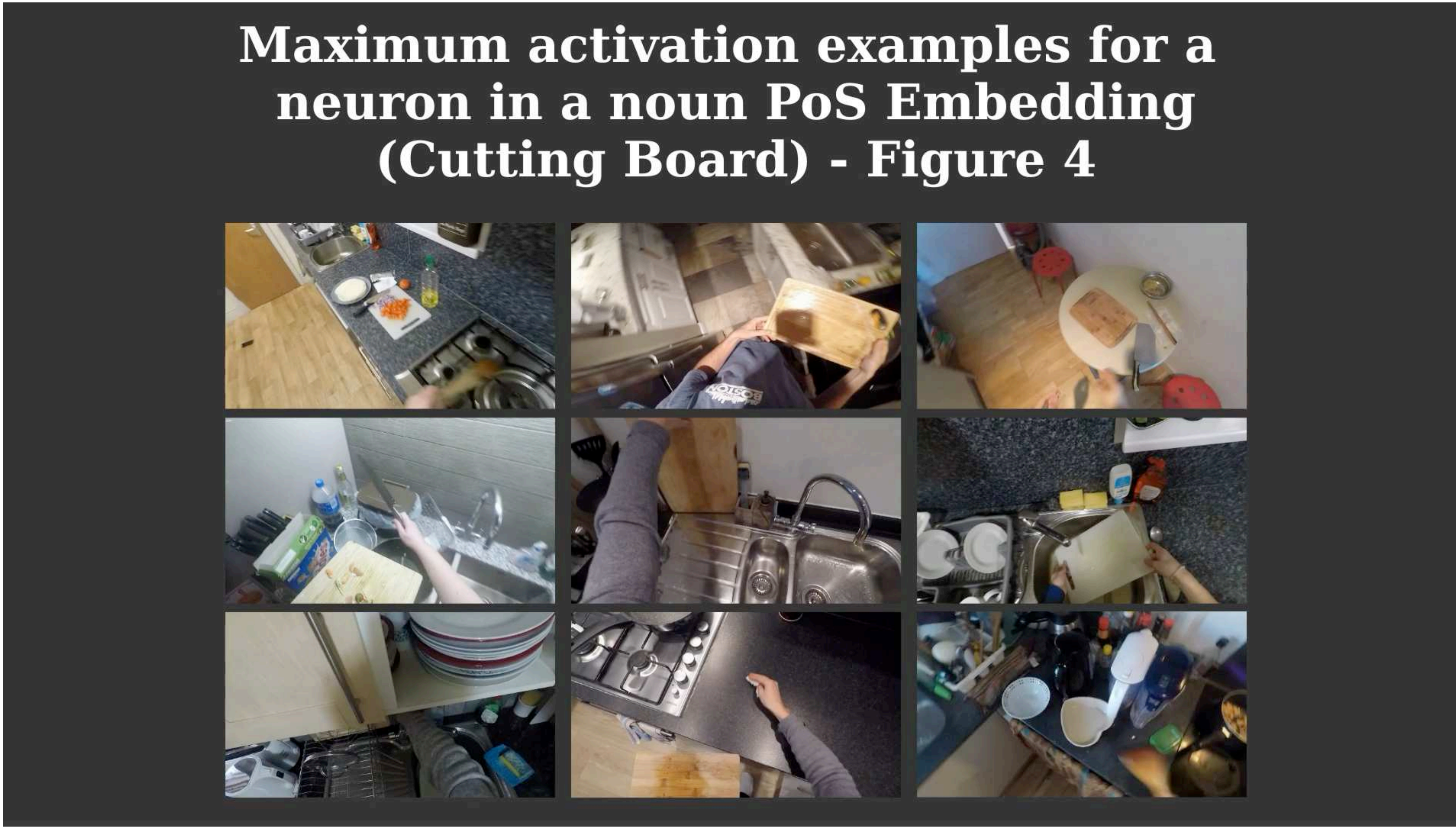
with: Michael Wray
Gabriela Csurka
Diane Larlus

EPIC	SEEN		UNSEEN	
	vt	tv	vt	tv
Random Baseline	0.6	0.6	0.9	0.9
CCA Baseline	20.6	7.3	14.3	3.7
MMEN (Verb)	3.6	4.0	3.9	4.2
MMEN (Noun)	9.9	9.2	7.9	6.1
MMEN (Caption)	14.0	11.2	10.1	7.7
MMEN ([Verb, Noun])	18.7	13.6	13.3	9.5
JPoSE (Verb, Noun)	**23.2**	**15.8**	**14.6**	**10.2**

Table 2. Cross-modal action retrieval on EPIC.

Fine-Grained Action Retrieval

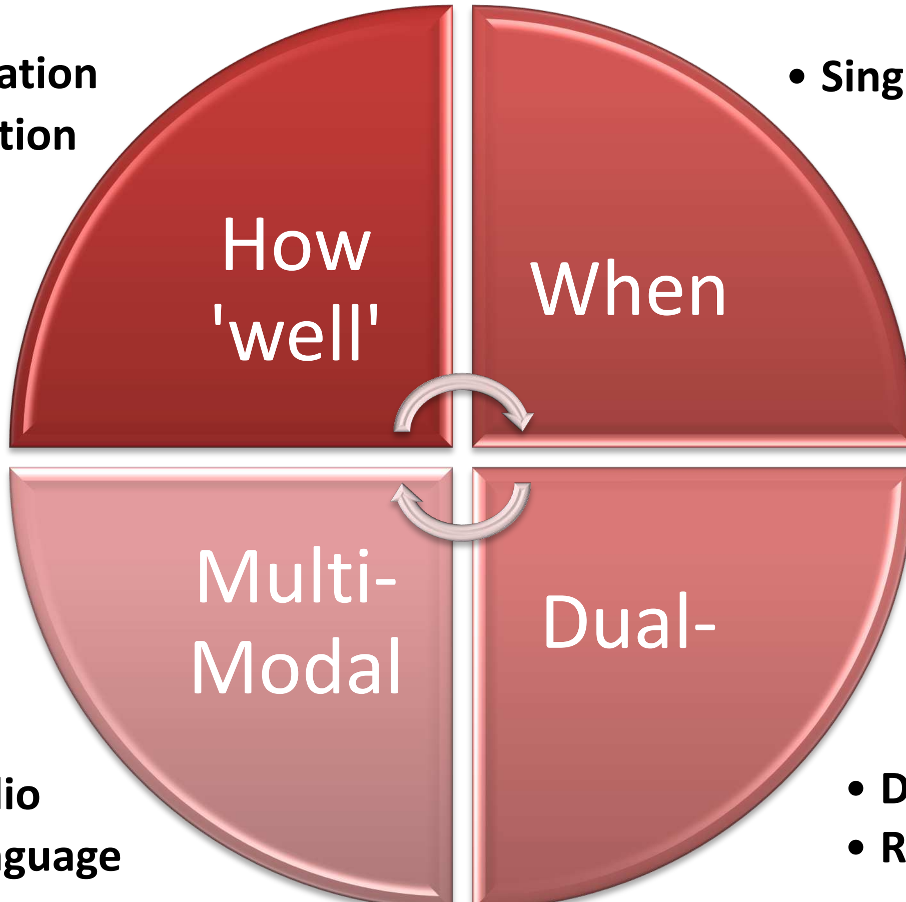
with: Michael Wray
Gabriela Csurka
Diane Larlus



Fine-Grained Object Interactions

- Skill Determination
- Action Completion

- Single-timestamp

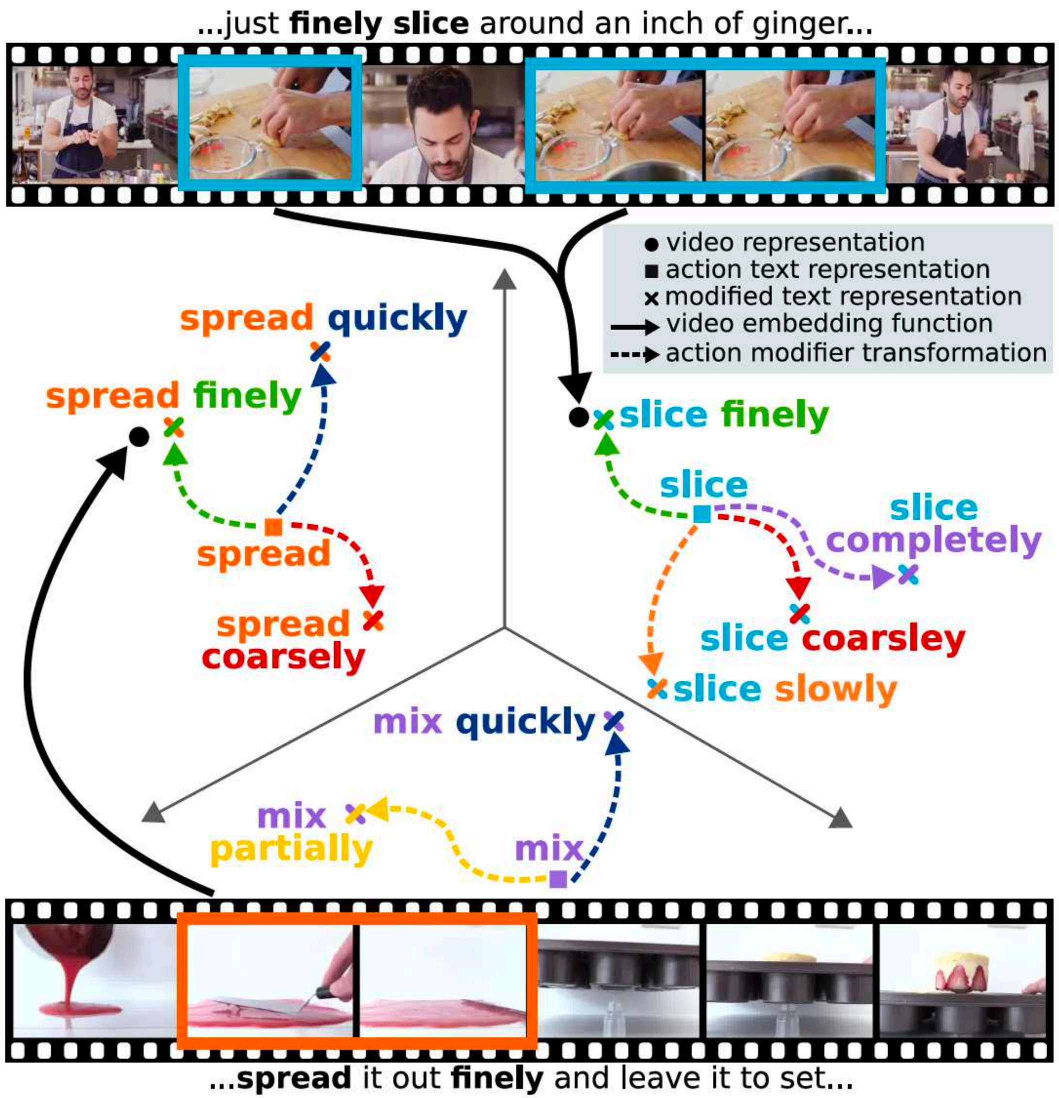


- Vision+Audio
- Vision+Language

- DDLSTM
- Retro-Actions

Now on Arxiv…

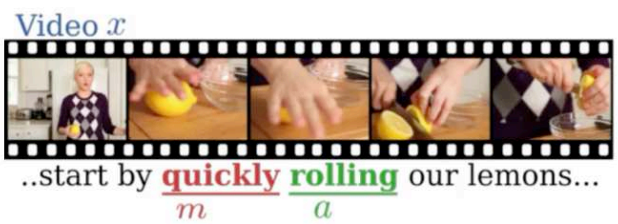
with: Hazel Doughty
Ivan Laptev
Walterio Mayol-Cuevas



Now on Arxiv…

with: Hazel Doughty
Ivan Laptev
Walterio Mayol-Cuevas

Method



April/May 2020….

with: Hazel Doughty
Jian Ma
Giovanni Maria Farinella
Antonino Furnari
Evangelos Kazkos
Davide Moltisanti
Jonathan Munro
Toby Perrett
Will Price
Michael Wray

Bigger

Better

Denser

100 hours

The Team

Thank you…

For further info, datasets, code, publications…

http://dimadamen.github.io

@dimadamen

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

Join epic-community@bristol.ac.uk

- send an email to: sympa@sympa.bristol.ac.uk
- with the subject: subscribe epic-community
- and blank message content