A Fine(r)-Grained Perspective onto Object Interactions

University of BRISTOL

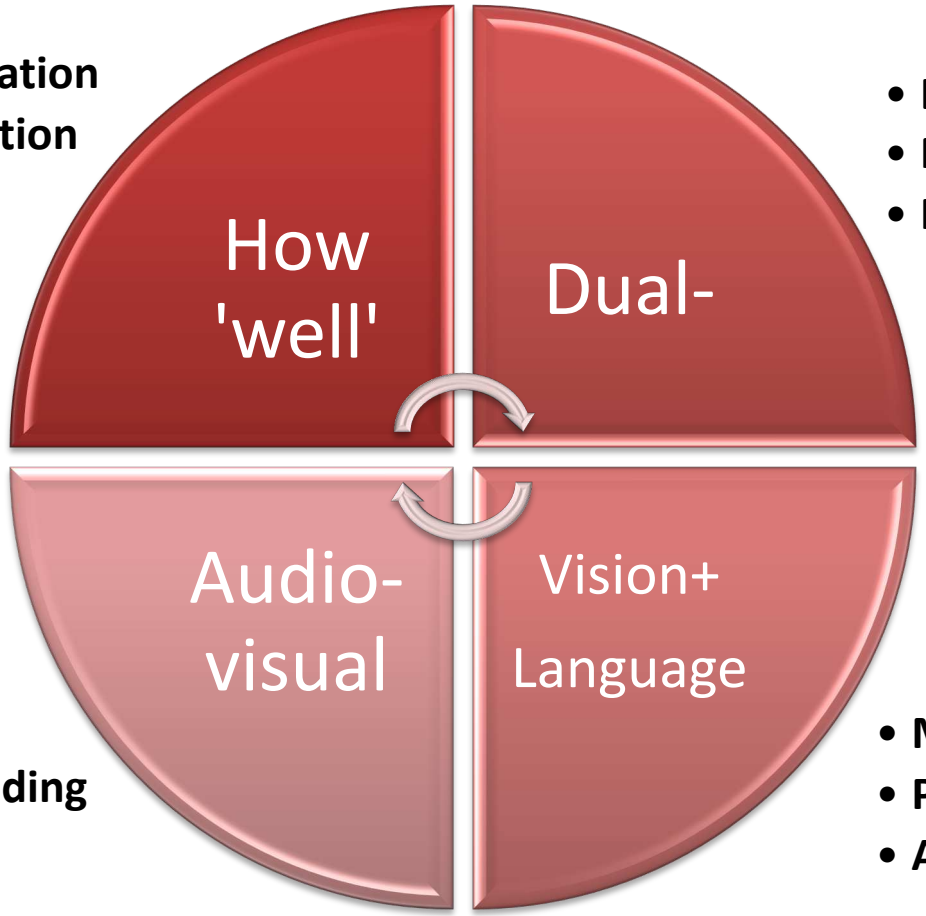
Dima Damen

28 April 2020

Natural Interactions

Fine-Grained Object Interactions

- Skill Determination
- Action Completion



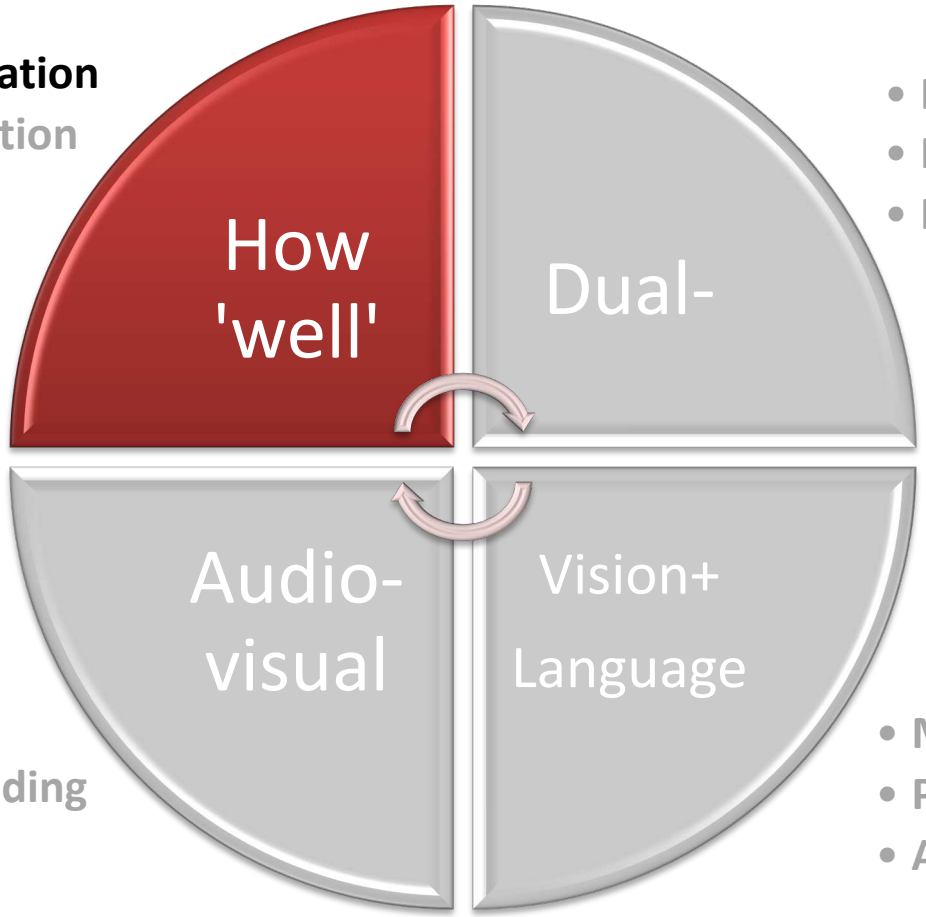
- DDLSTM
- Multi-modal UDA
- Retro-Actions

- Temporal Binding

- Multi-Verb Labels
- Part-of-Speech
- Adverbs

Fine-Grained Object Interactions

- **Skill Determination**
- Action Completion



- DDLSTM
- Multi-modal UDA
- Retro-Actions

- Temporal Binding

- Multi-Verb Labels
- Part-of-Speech
- Adverbs

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

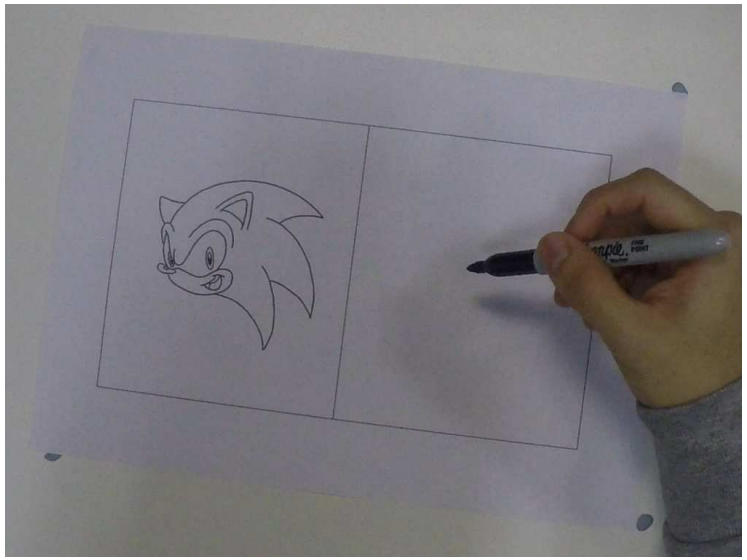
with: Hazel Doughty
Walterio Mayol-Cuevas

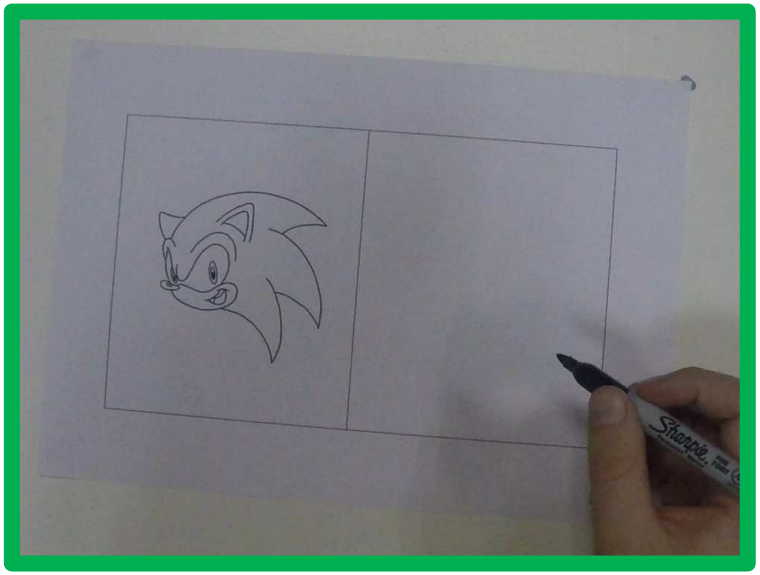
Assess relative skill for a collection of video sequences, applicable to a variety of tasks.

Skill Determination from Video

with: Hazel Doughty
Walterio Mayol-Cuevas

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



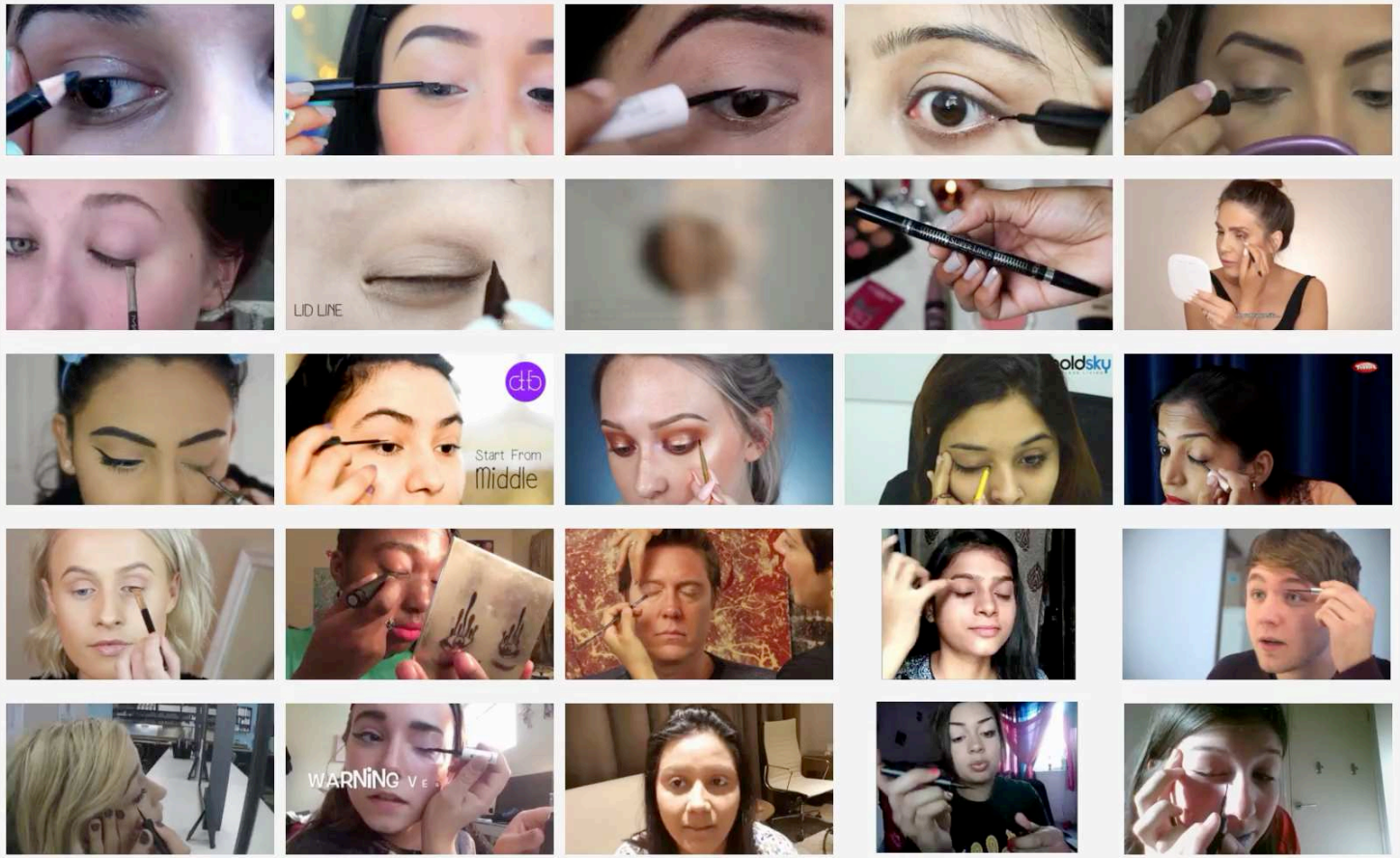


H Doughty, W Mayol-Cuevas, D Damen (2019). The Pros and Cons: Rank-aware Temporal Attention for Skill Determination in Long Videos. *Computer Vision and Pattern Recognition (CVPR)*

Skill Determination in Video

with: Hazel Doughty
Walterio Mayol-Cuevas

Best

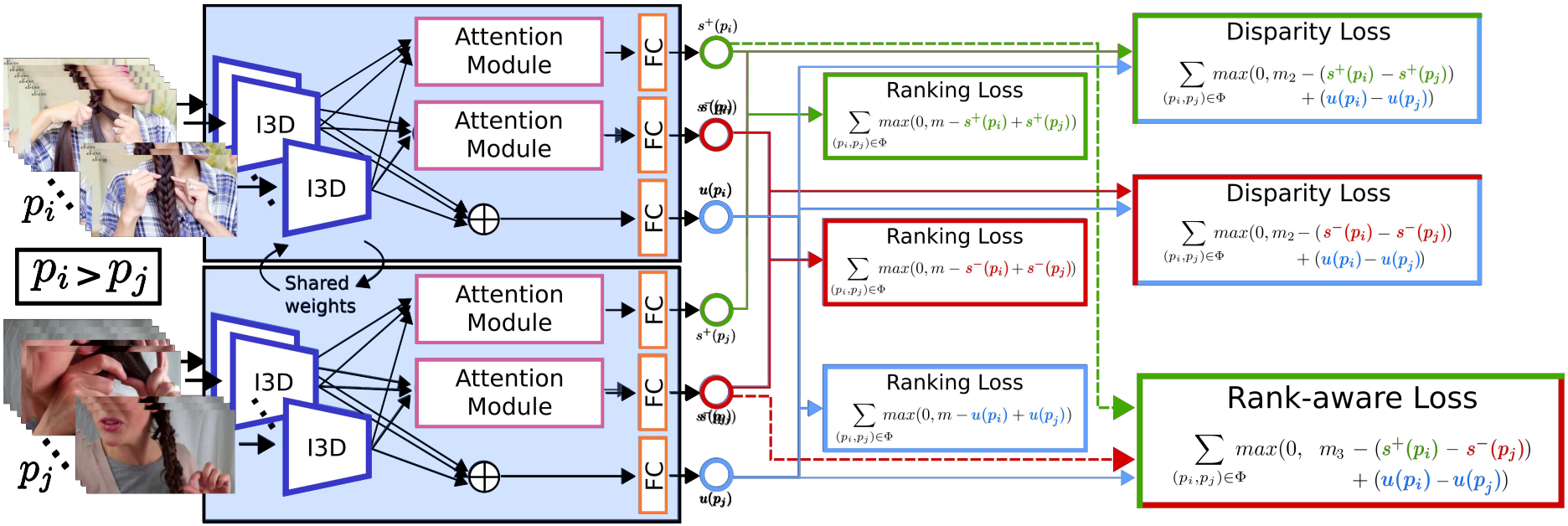


Worst

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

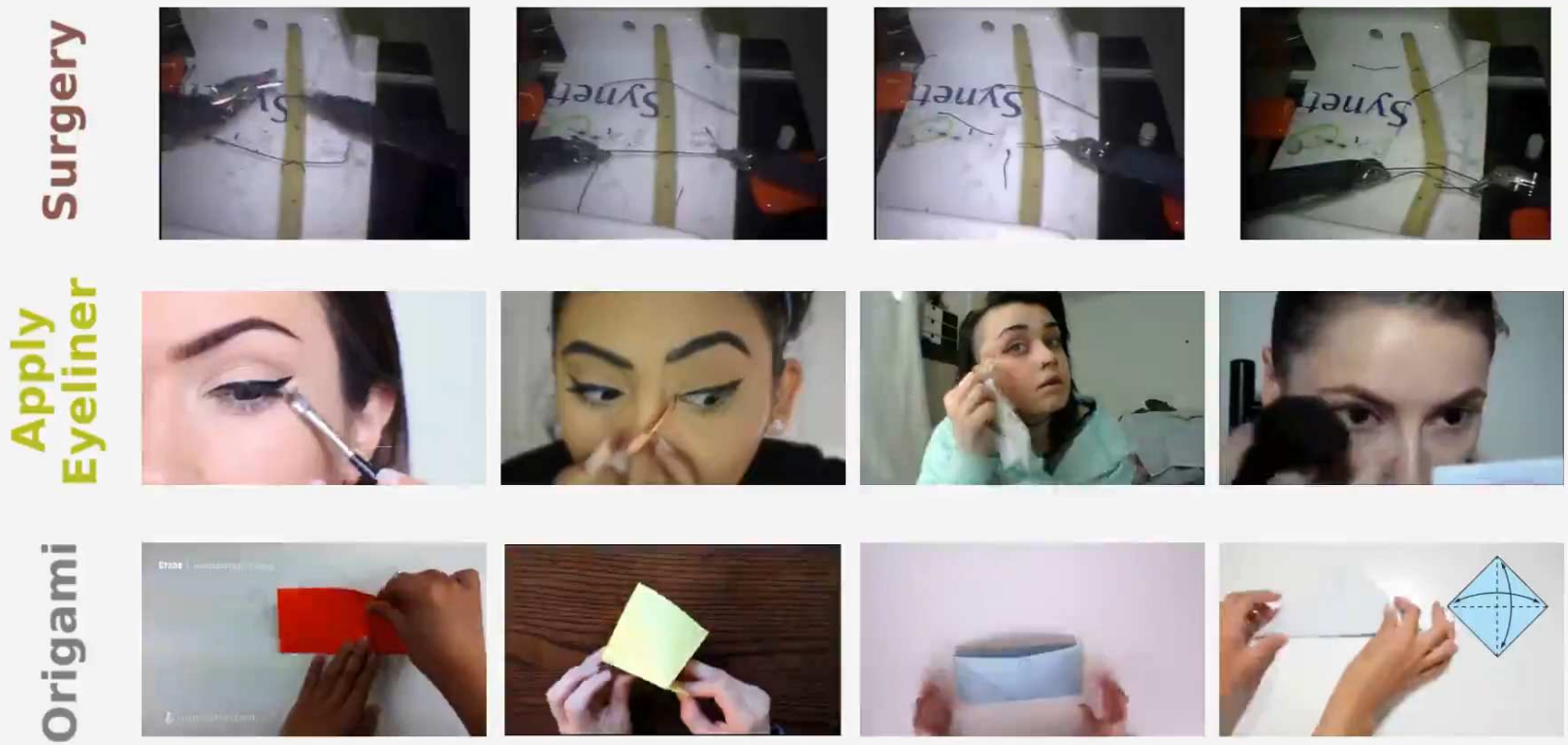
with: Hazel Doughty
Walterio Mayol-Cuevas



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)*

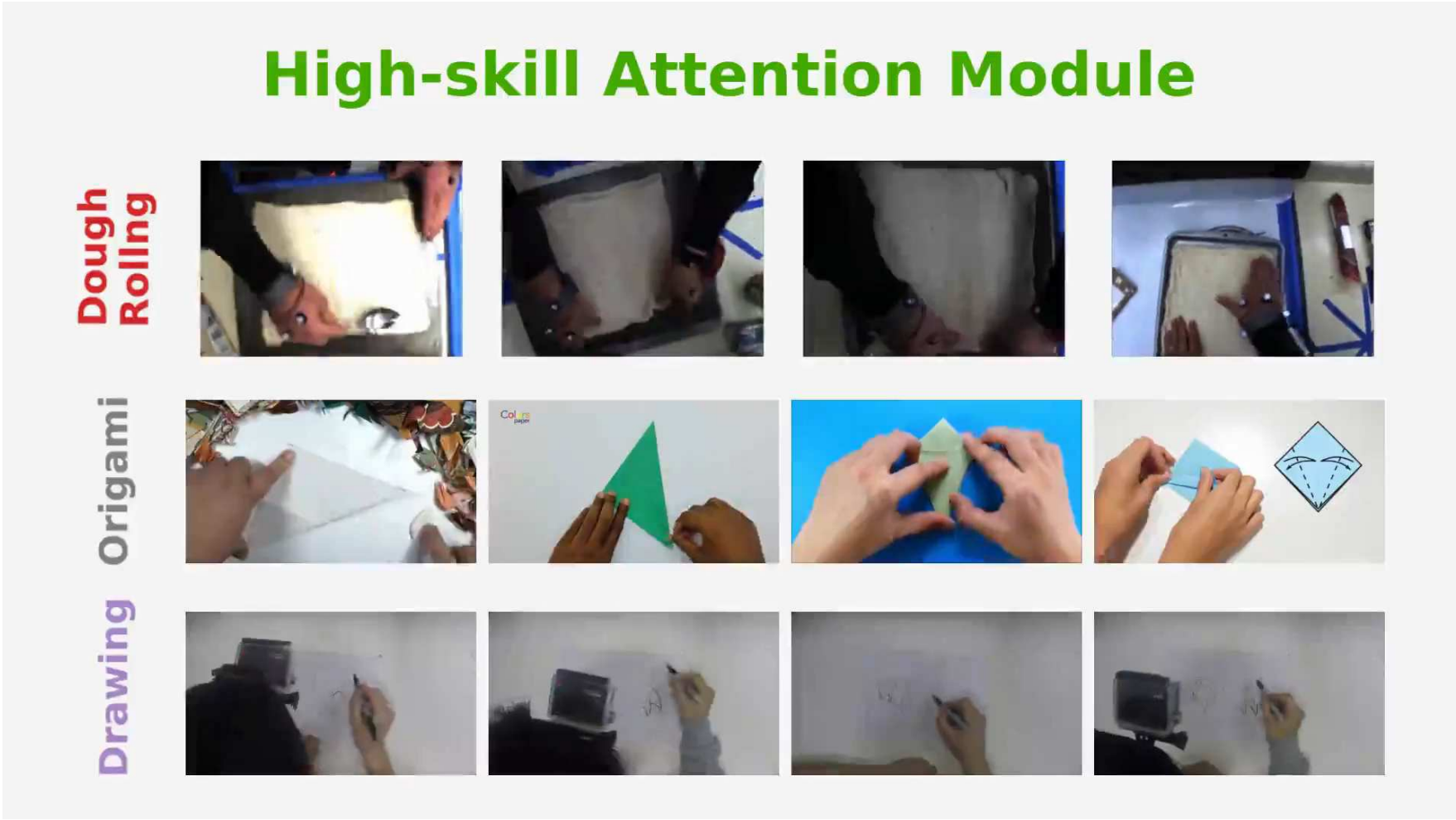
The Pros and Cons: Rank-Aware Temporal Attention

with: Hazel Doughty
Walterio Mayol-Cuevas



The Pros and Cons: Rank-Aware Temporal Attention

with: Hazel Doughty
Walterio Mayol-Cuevas



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

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.

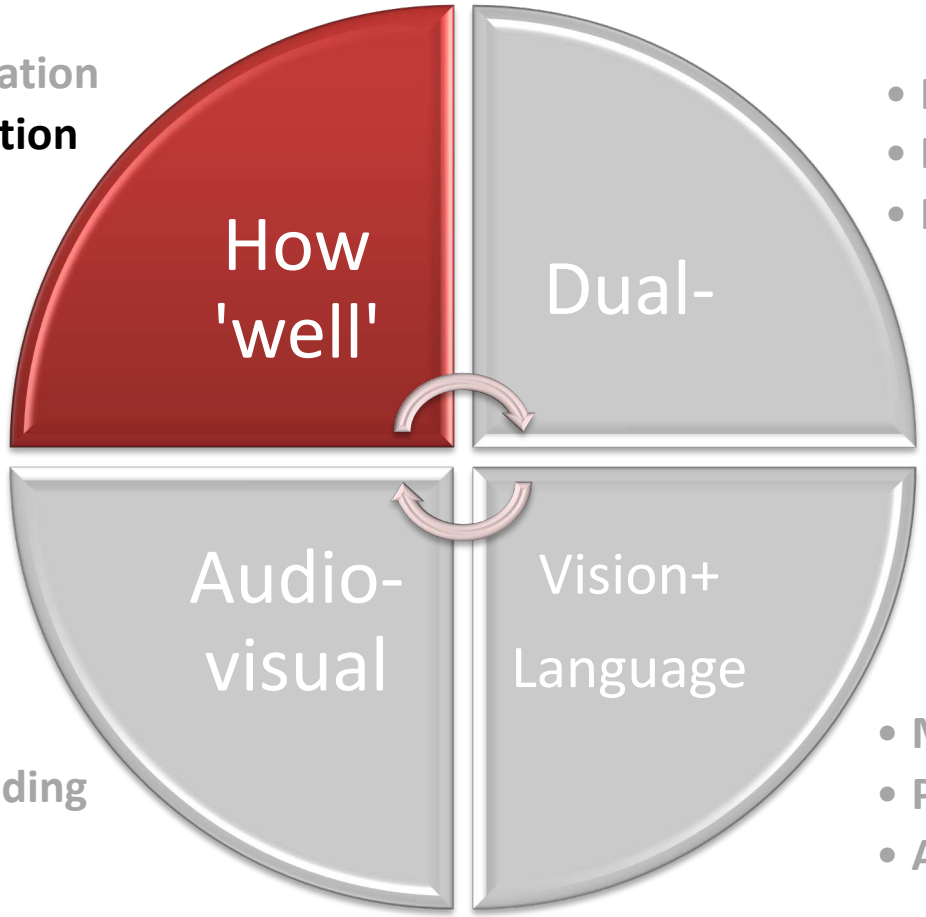
Our approach trains temporal attention modules, learned with only video-level supervision, using a novel rank-aware loss function. In addition to attending to task-relevant video parts, our proposed loss jointly trains two attention modules to separately attend to video parts which are indicative of 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**



- DDLSTM
- Multi-modal UDA
- Retro-Actions

- Temporal Binding

- Multi-Verb Labels
- Part-of-Speech
- Adverbs

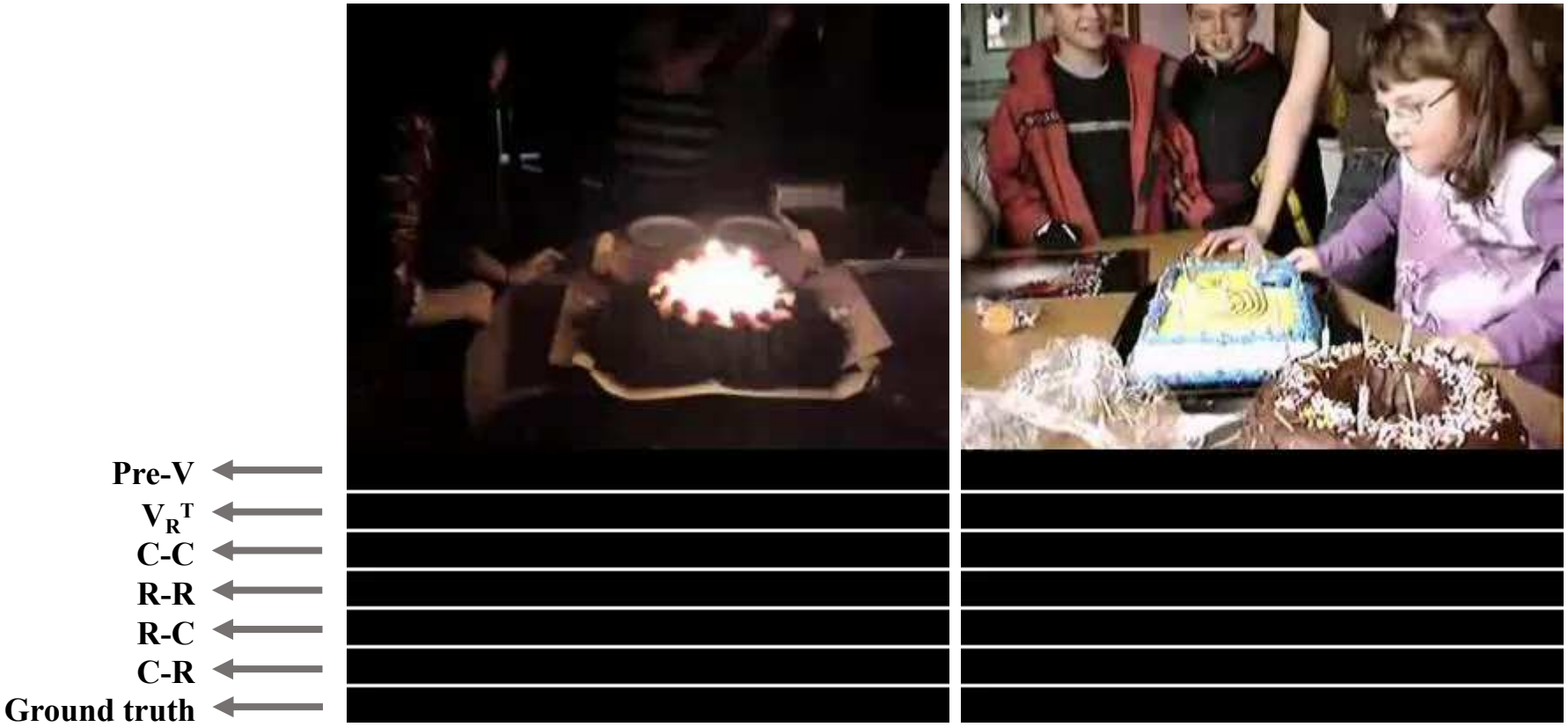
with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

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

with: Farnoosh Heidarivincheh
Majid Mirmehdi

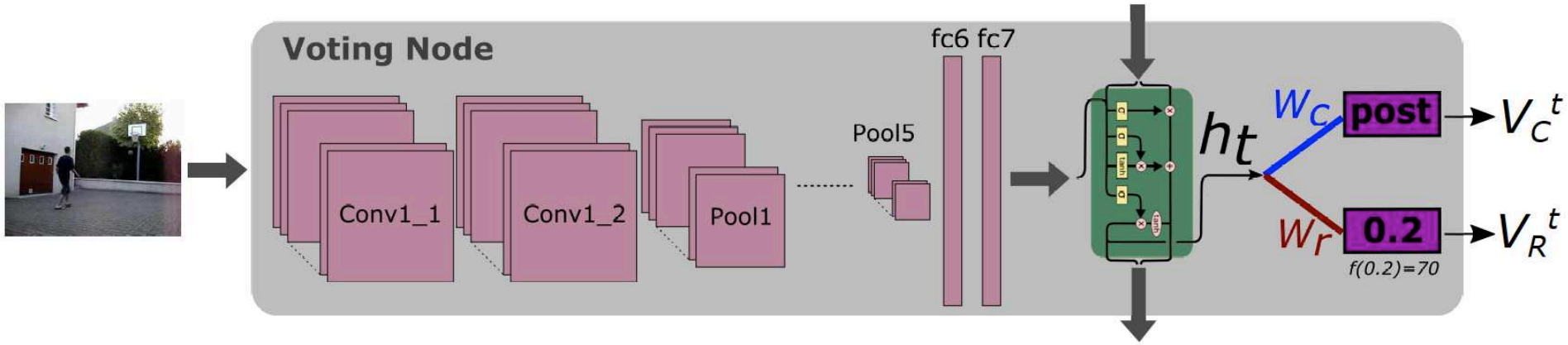
Action Completion Detection



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

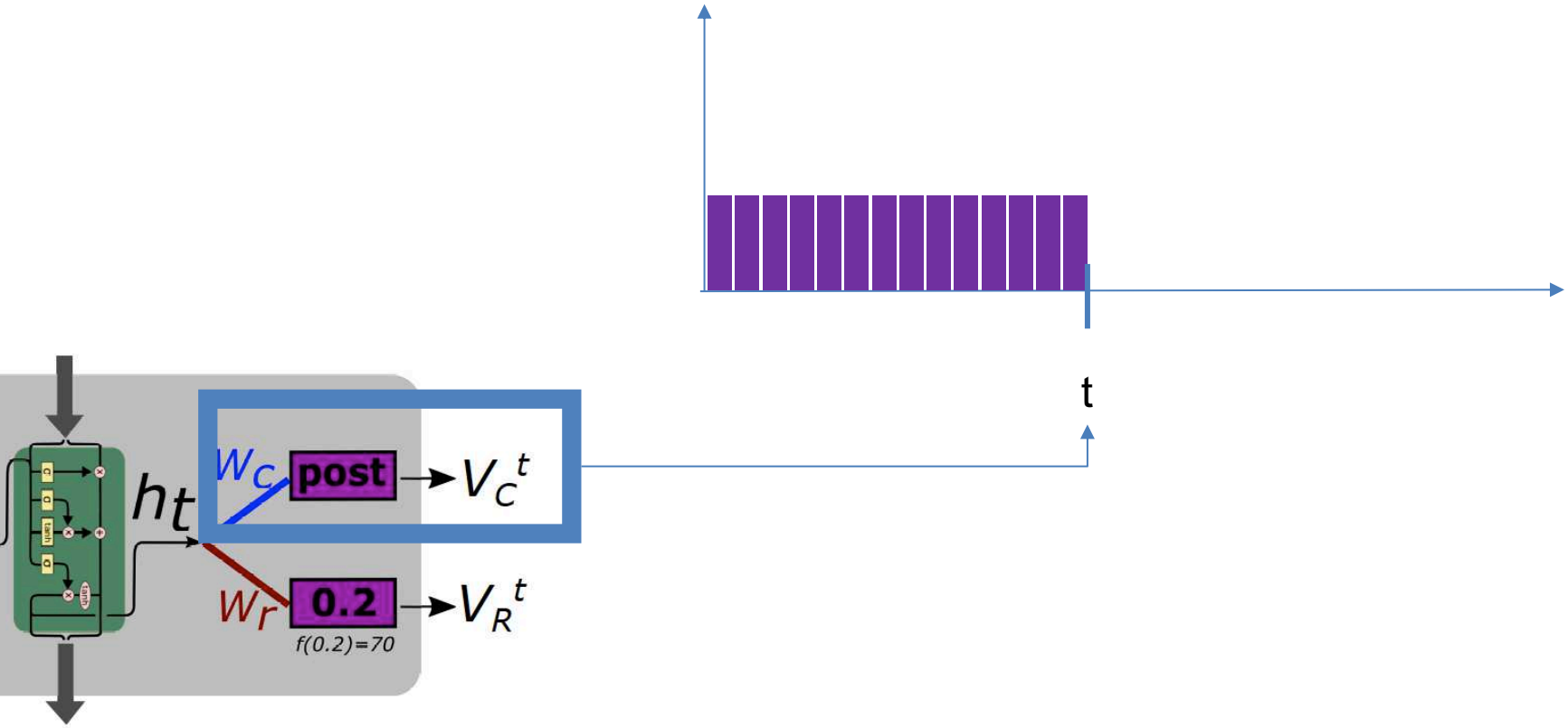
Action Completion Detection

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



Action Completion Detection

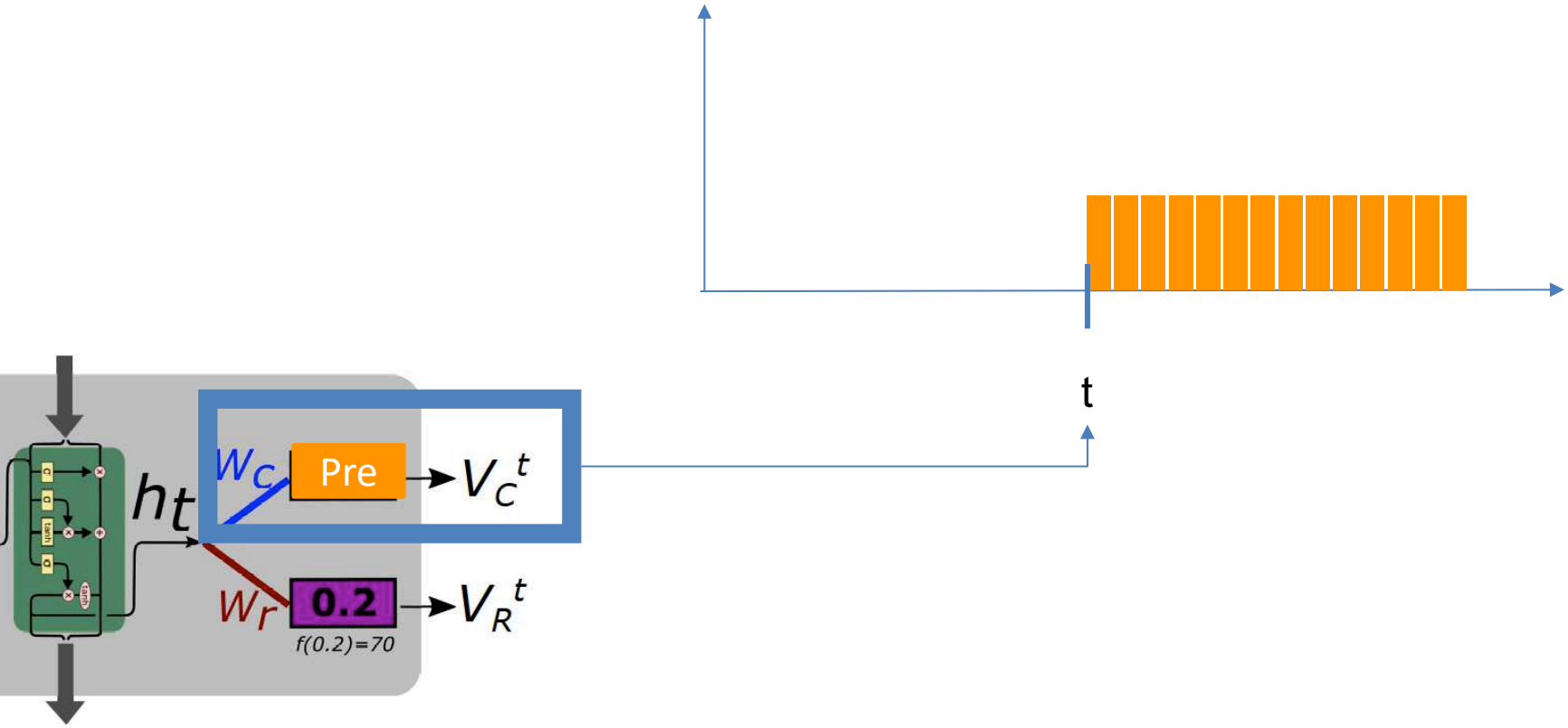
1. Classification-Based Voting



with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

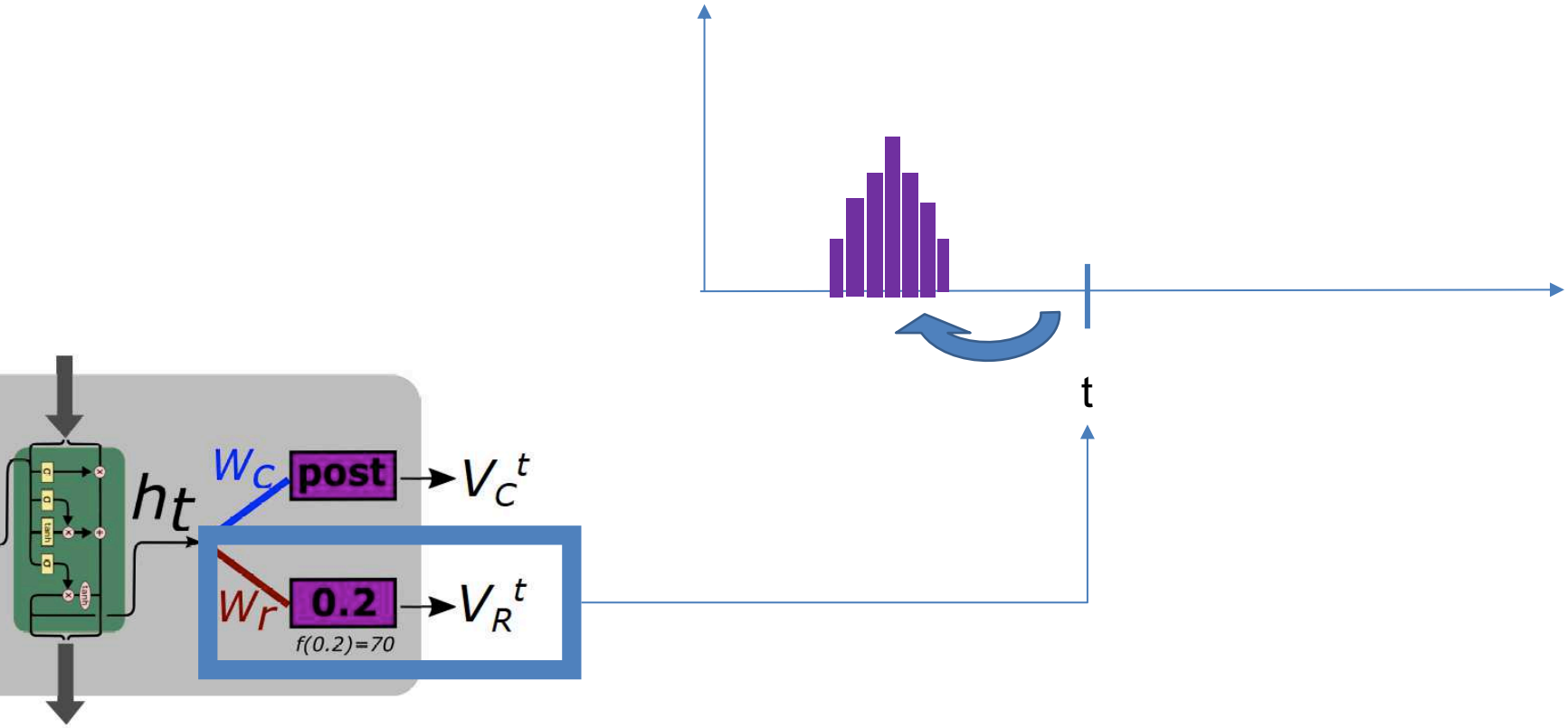
1. Classification-Based Voting



with: Farnoosh Heidarivincheh
Majid Mirmehdi

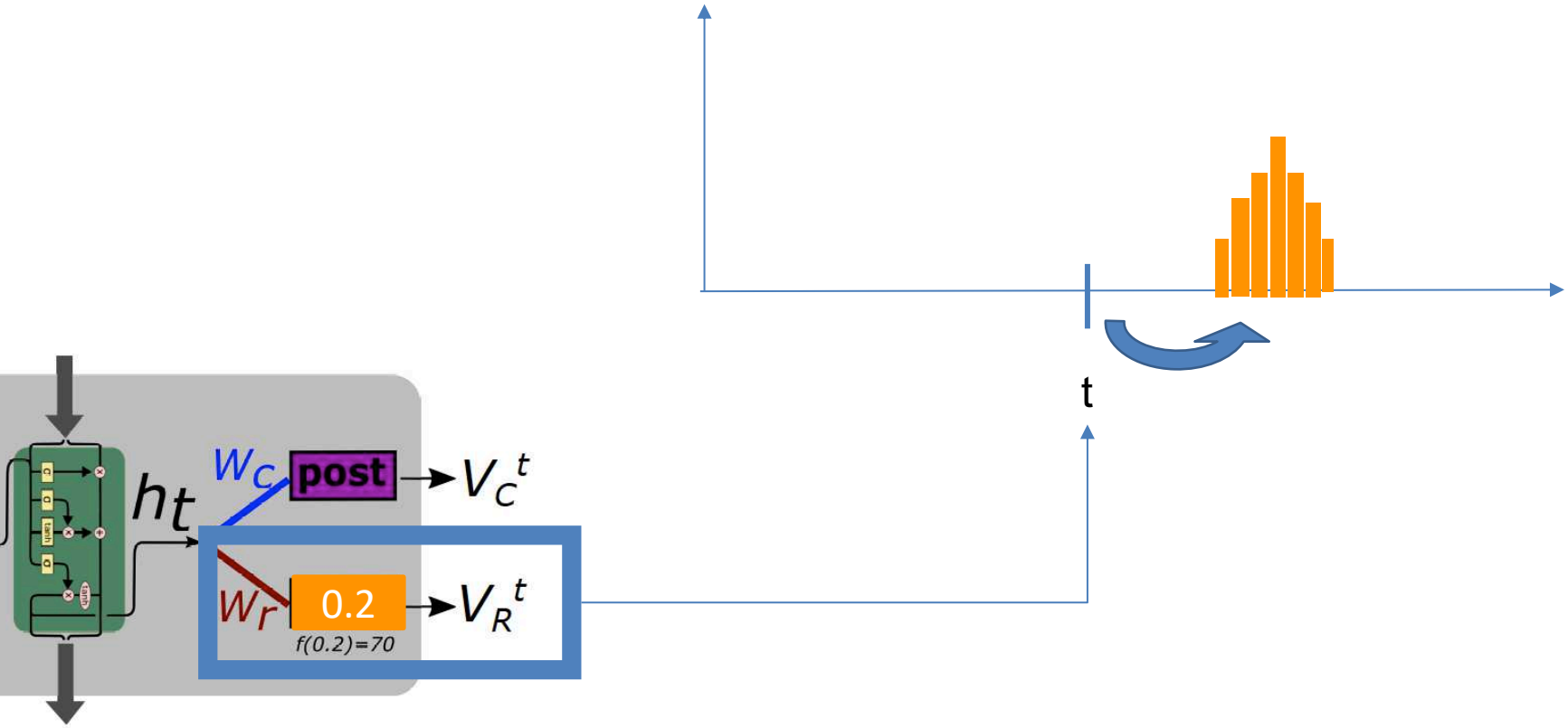
Action Completion Detection

2. Regression-Based Voting



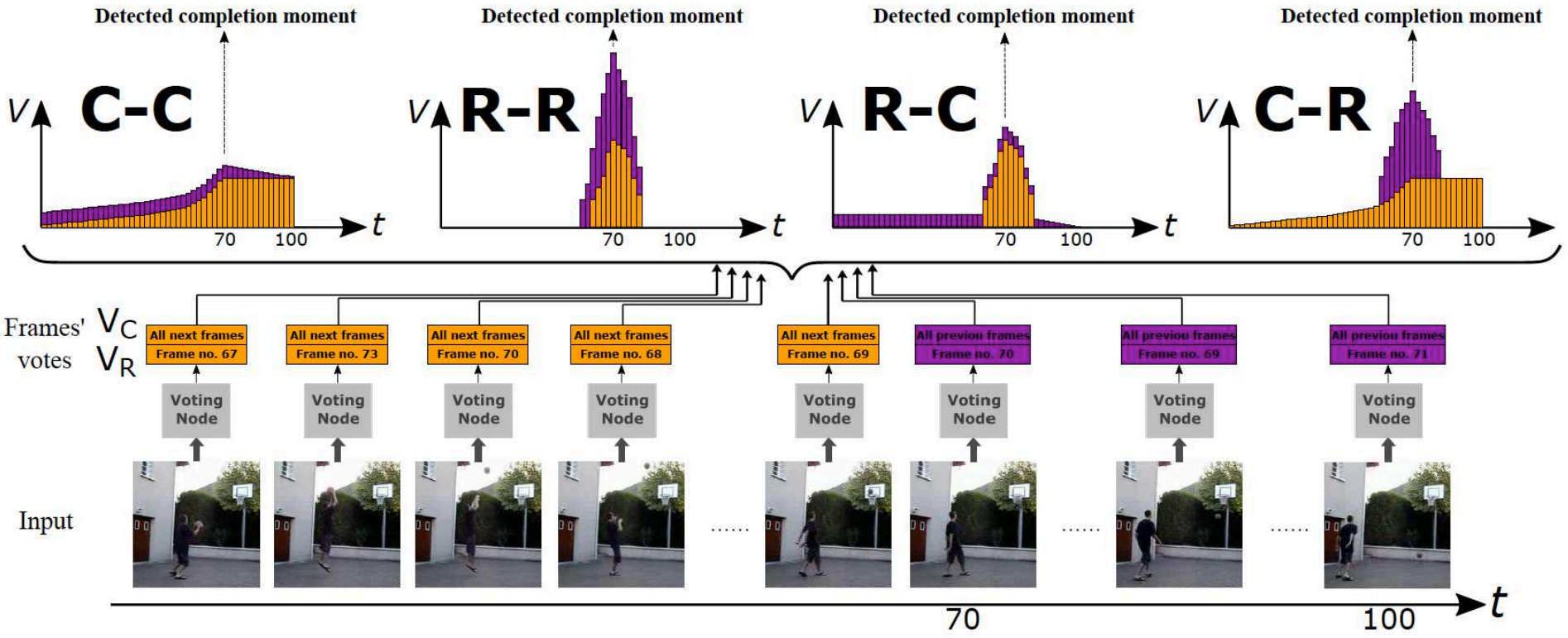
Action Completion Detection

2. Regression-Based Voting



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

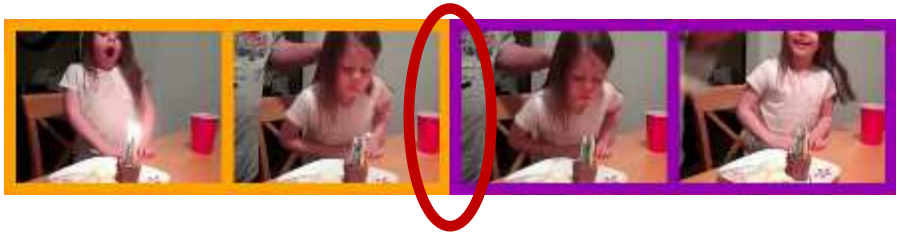
Action Completion Detection

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

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.

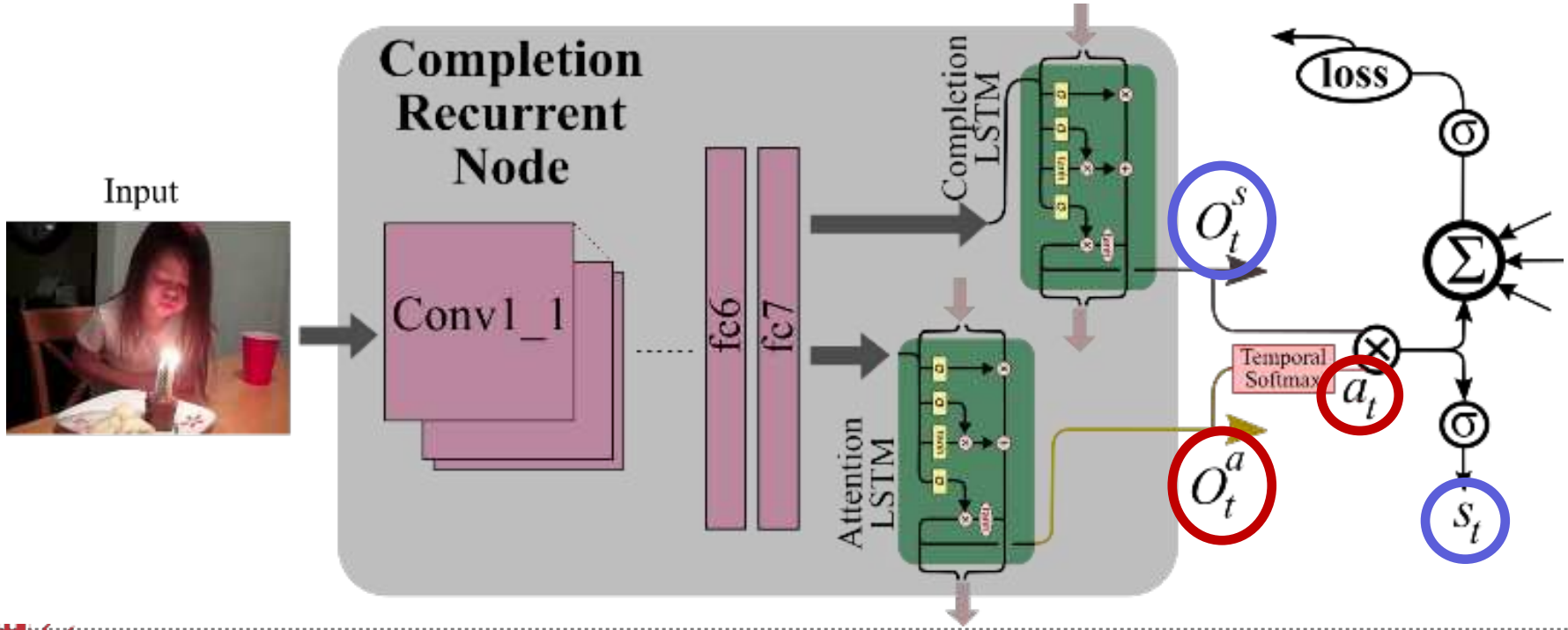
sequence-level *complete* and ***incomplete*** labels

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

Attention LSTM → o_t^a —Softmax→ a_t

Completion LSTM → o_t^s —$\sigma(o_t^s \times a_t)$→ s_t



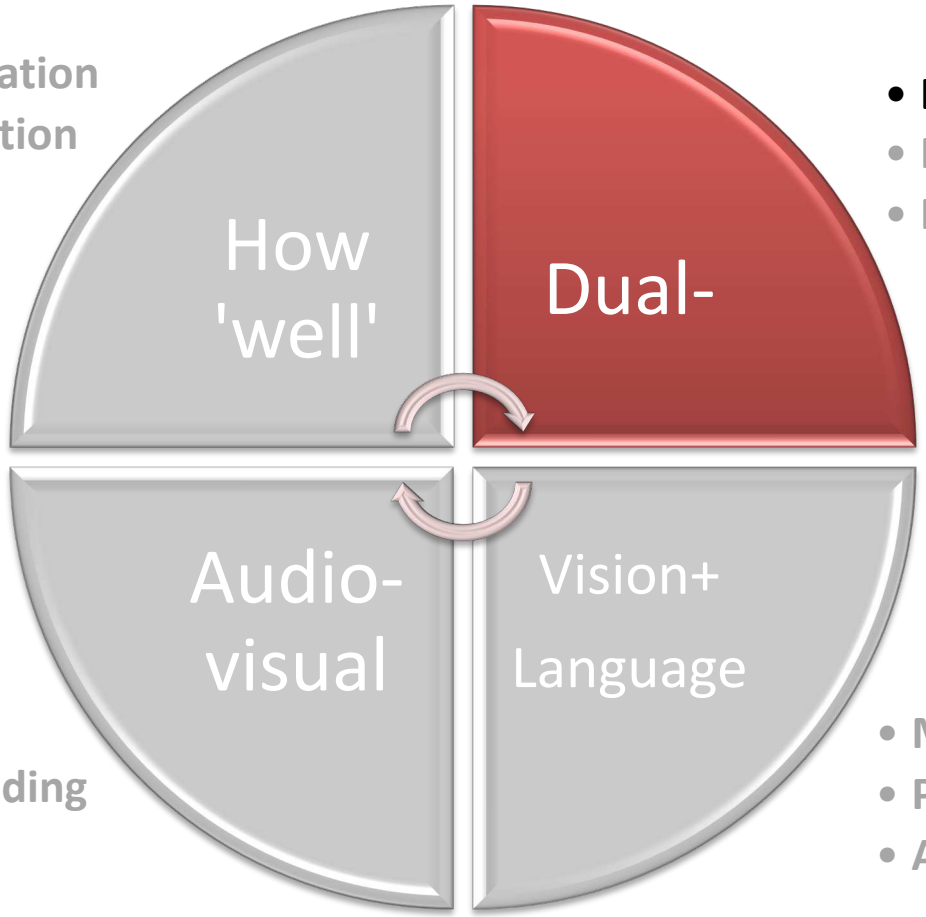
F Heidarivincheh, M Mirmehdi, D Damen (2019). Weakly-Supervised Completion Moment Detection using Temporal Attention. ICCV Workshop on Human Behaviour Understanding

with: Farnoosh Heidarivincheh
Majid Mirmehdi

Action Completion Detection

Fine-Grained Object Interactions

- Skill Determination
- Action Completion



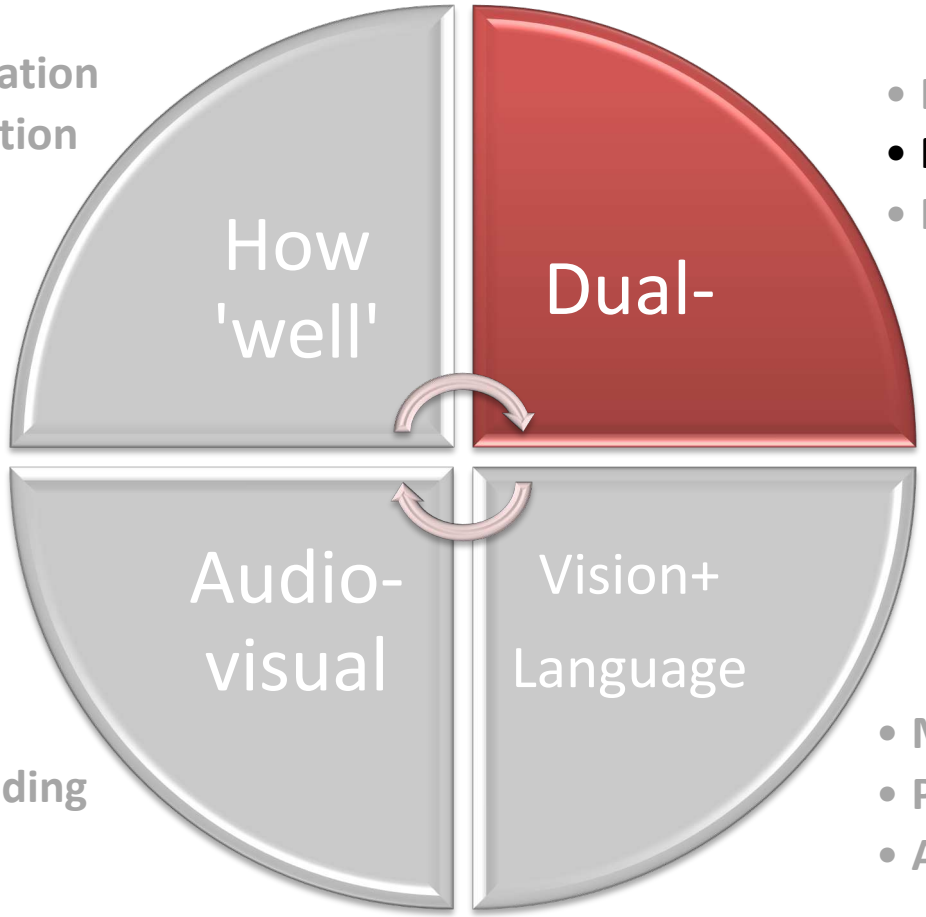
- **DDLSTM**
- Multi-modal UDA
- Retro-Actions

- Temporal Binding

- Multi-Verb Labels
- Part-of-Speech
- Adverbs

Fine-Grained Object Interactions

- Skill Determination
- Action Completion



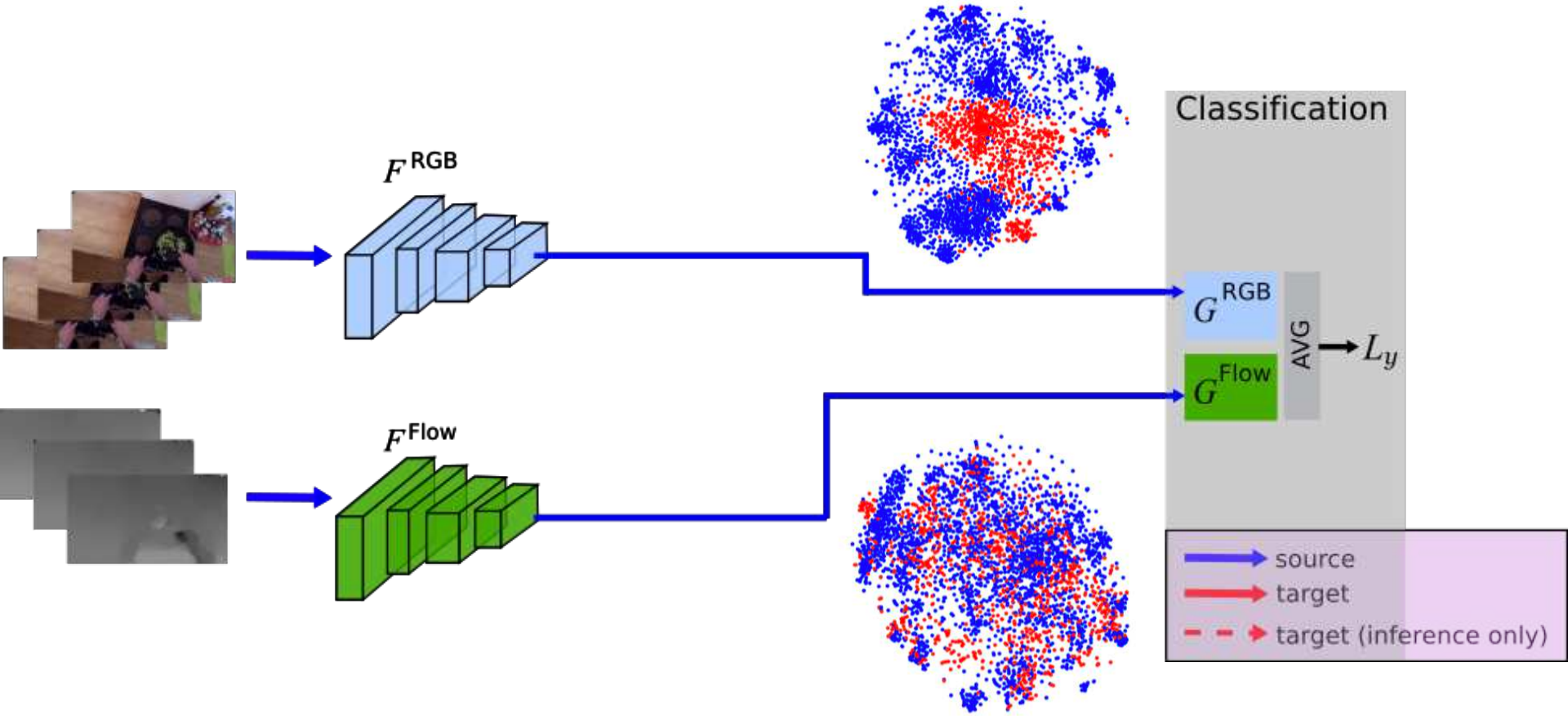
- DDLSTM
- **Multi-modal UDA**
- Retro-Actions

- Temporal Binding

- Multi-Verb Labels
- Part-of-Speech
- Adverbs

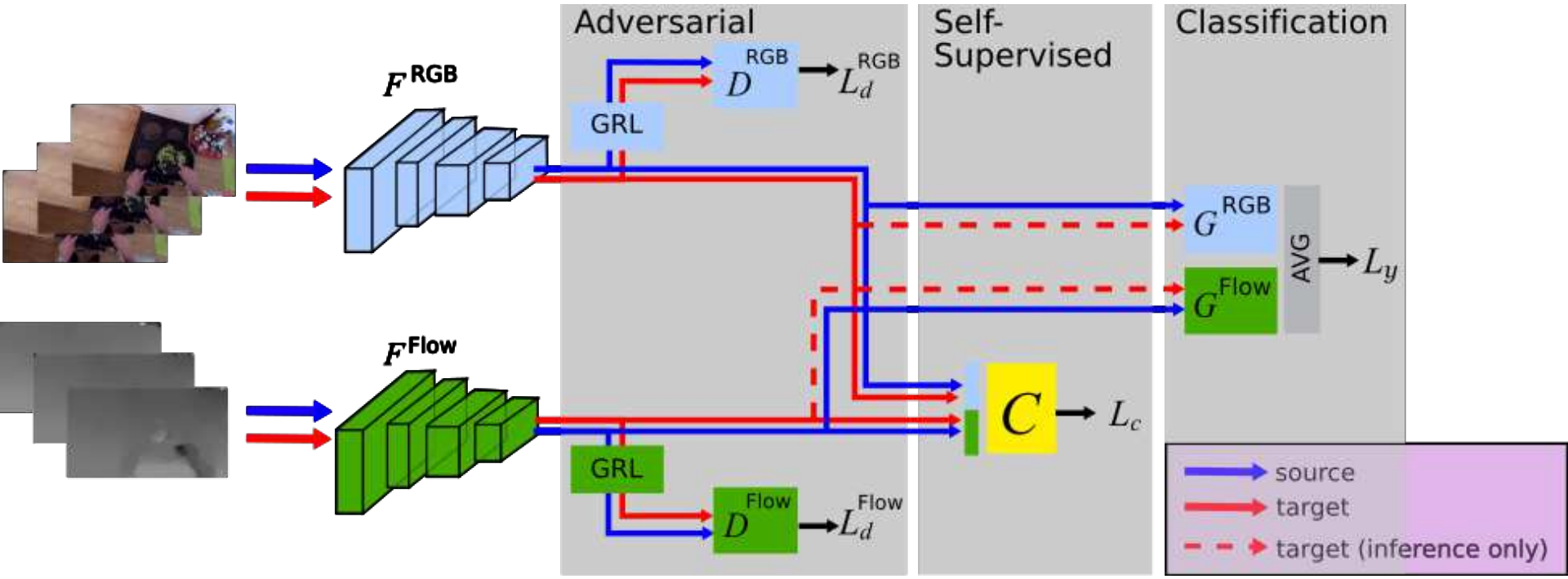
Multi-Modal Domain Adaptation for Fine-Grained Action Recognition

with: Jonathan Munro



Multi-Modal Domain Adaptation for Fine-Grained Action Recognition

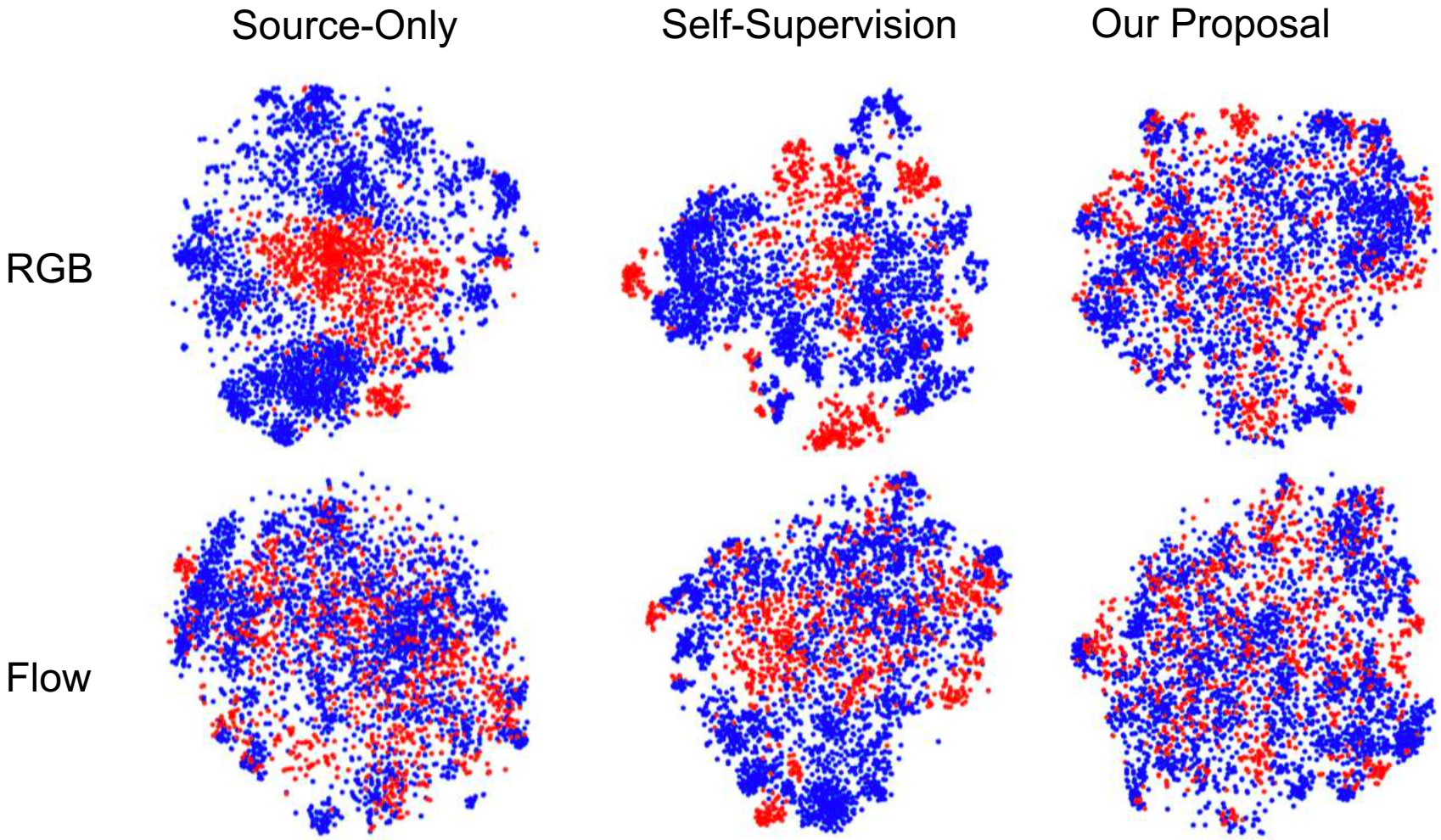
with: Jonathan Munro



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

Multi-Modal Domain Adaptation for Fine-Grained Action Recognition

with: Jonathan Munro



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

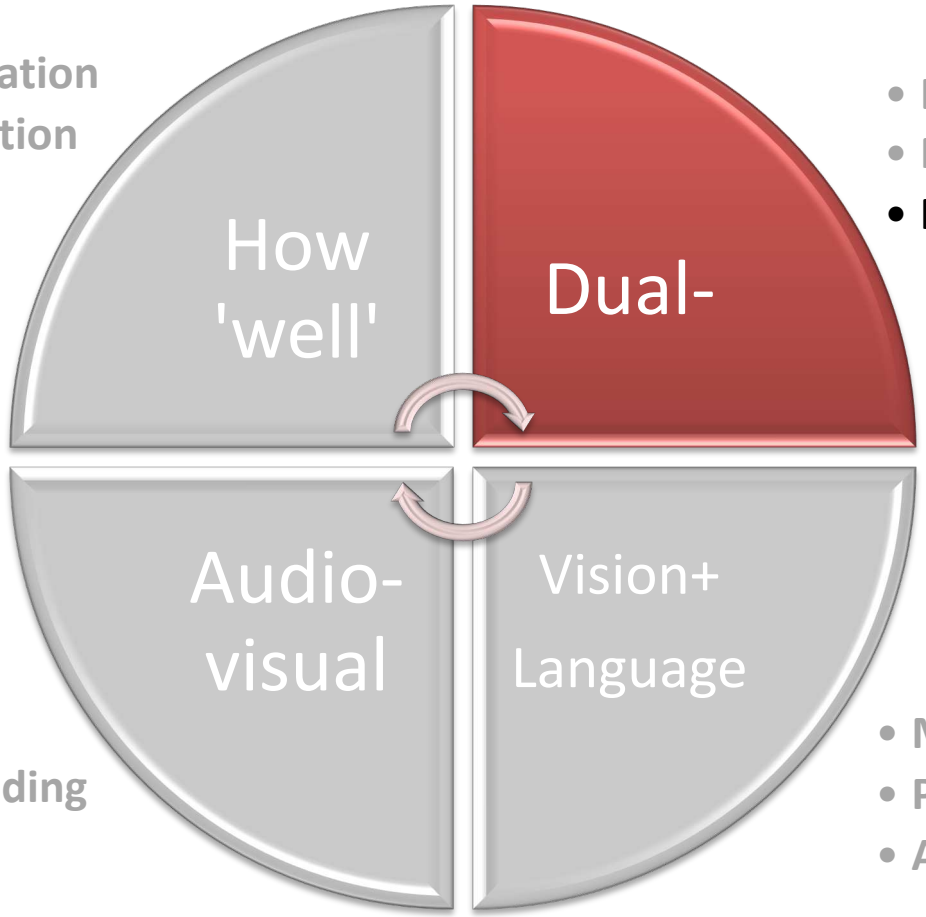
Multi-Modal Domain Adaptation for Fine-Grained Action Recognition

with: Jonathan Munro

	D2→ D1	D3→ D1	D1→ D2	D3→ D2	D1→ D3	D2→ D3	Mean
MM Source-only	42.5	44.3	42.0	**56.3**	41.2	46.5	45.5
AdaBN [29]	44.6	47.8	47.0	54.7	40.3	48.8	47.2
MMD [32]	43.1	48.3	46.6	55.2	39.2	48.5	46.8
MCD [45]	42.1	47.9	46.5	52.7	43.5	51.0	47.3
MM-SADA	**48.2** ▲+5.7	**50.9**▲+6.6	**49.5**▲+7.5	56.1▼-0.2	**44.1**▲+2.9	**52.7** ▲+6.3	**50.3** ▲+4.8
Supervised target	62.8	62.8	71.7	71.7	74.0	74.0	69.5

Fine-Grained Object Interactions

- Skill Determination
- Action Completion

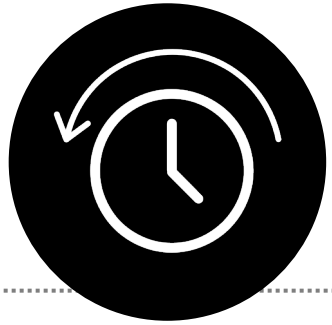


- DDLSTM
- Multi-modal UDA
- **Retro-Actions**

- Temporal Binding

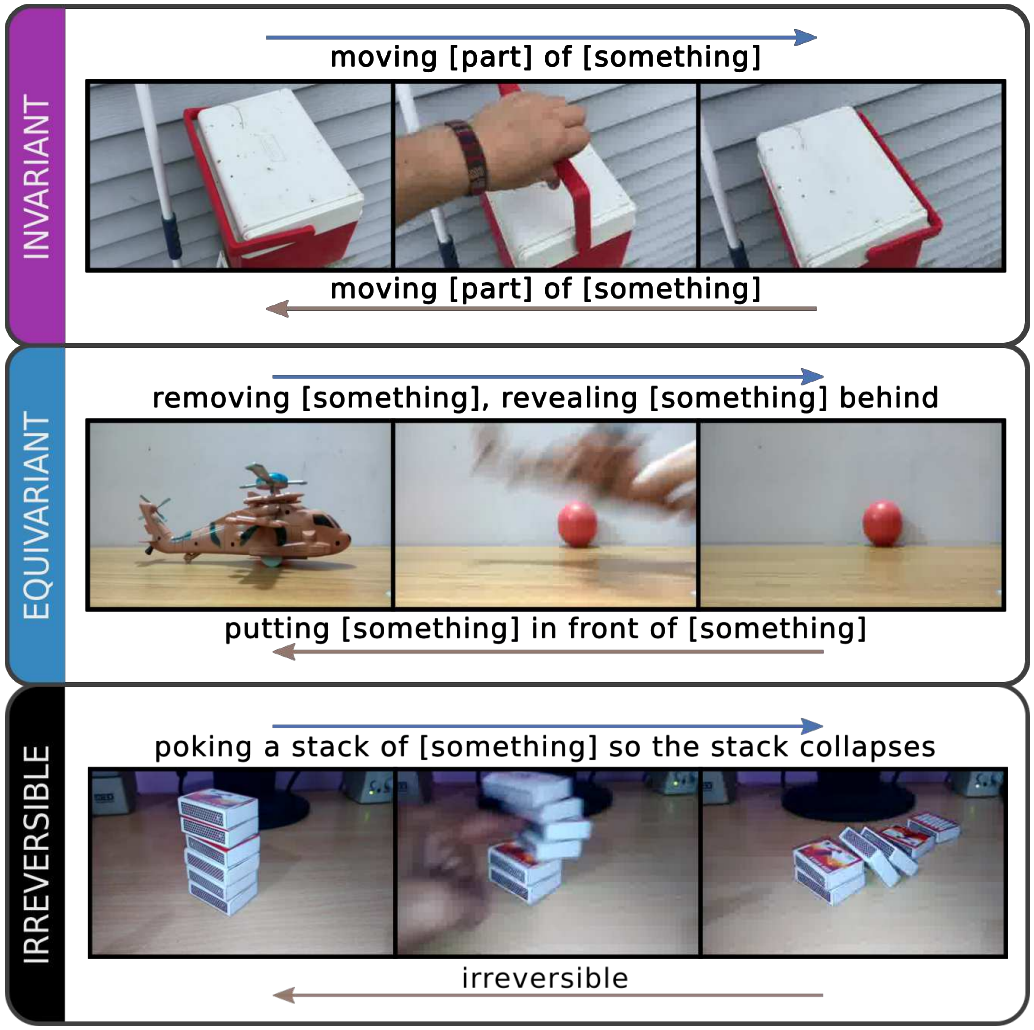
- Multi-Verb Labels
- Part-of-Speech
- Adverbs

Retro-actions



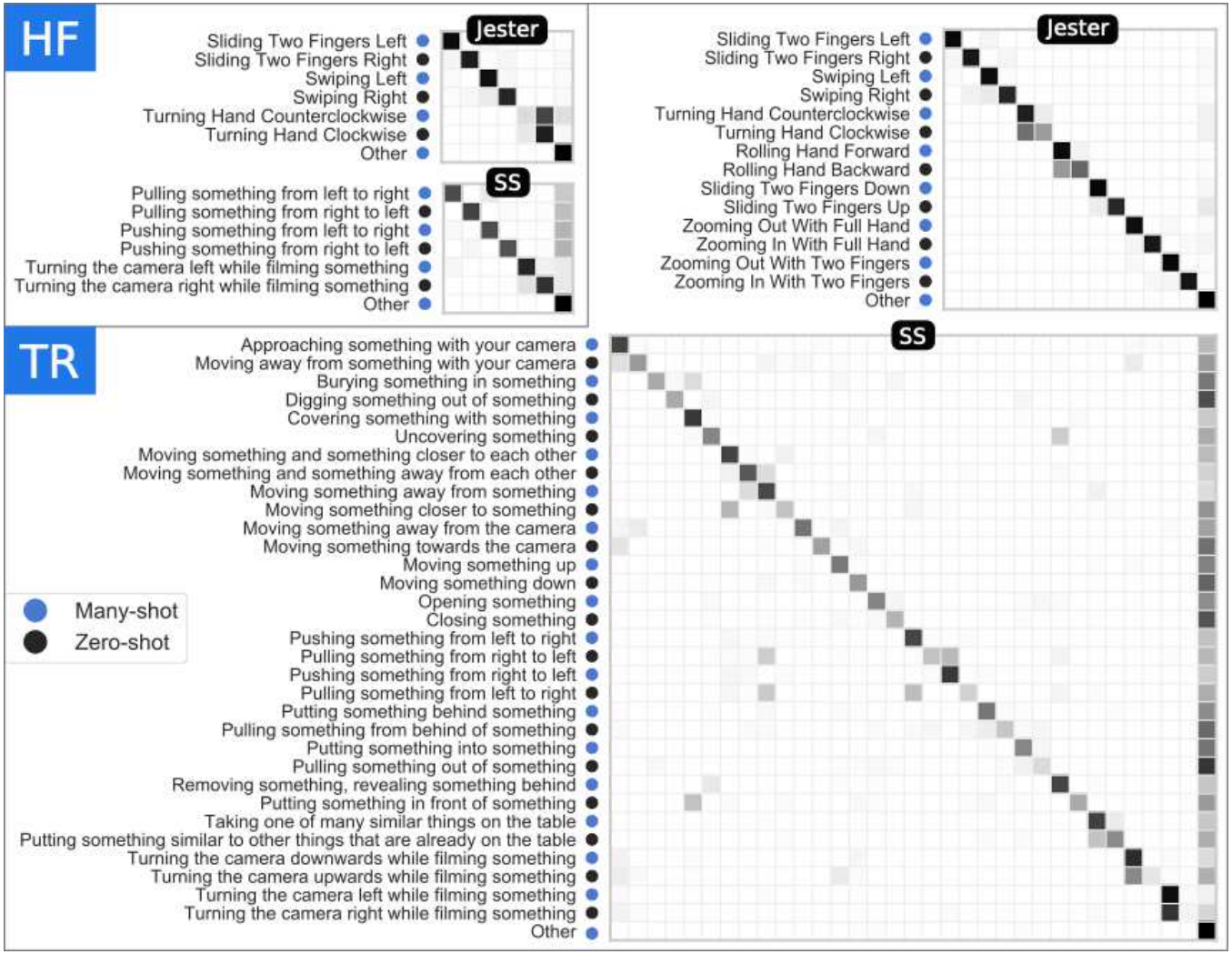
W Price, D Damen (2019). Retro-Actions: Learning 'Close' by Time-Reversing 'Open' Videos. ICCV MDALC Workshop

Retro-actions



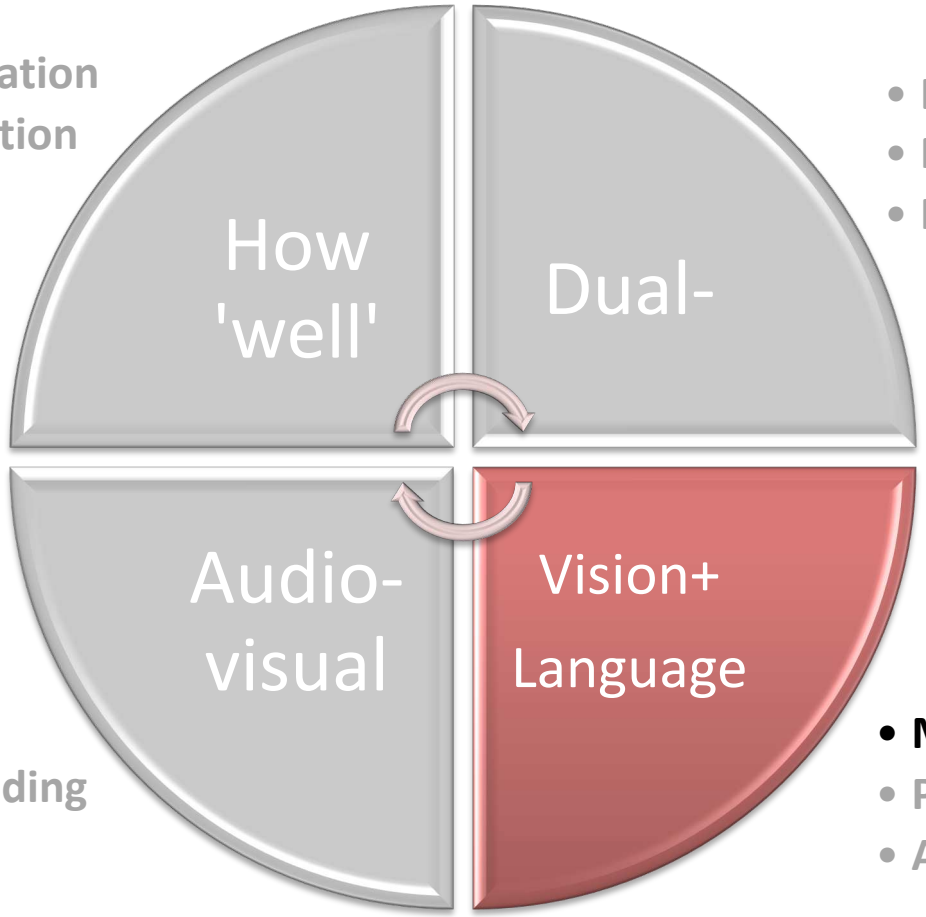
W Price, D Damen (2019). Retro-Actions: Learning 'Close' by Time-Reversing 'Open' Videos. ICCV MDALC Workshop

Retro-actions – Zero-Shot Learning



Fine-Grained Object Interactions

- Skill Determination
- Action Completion

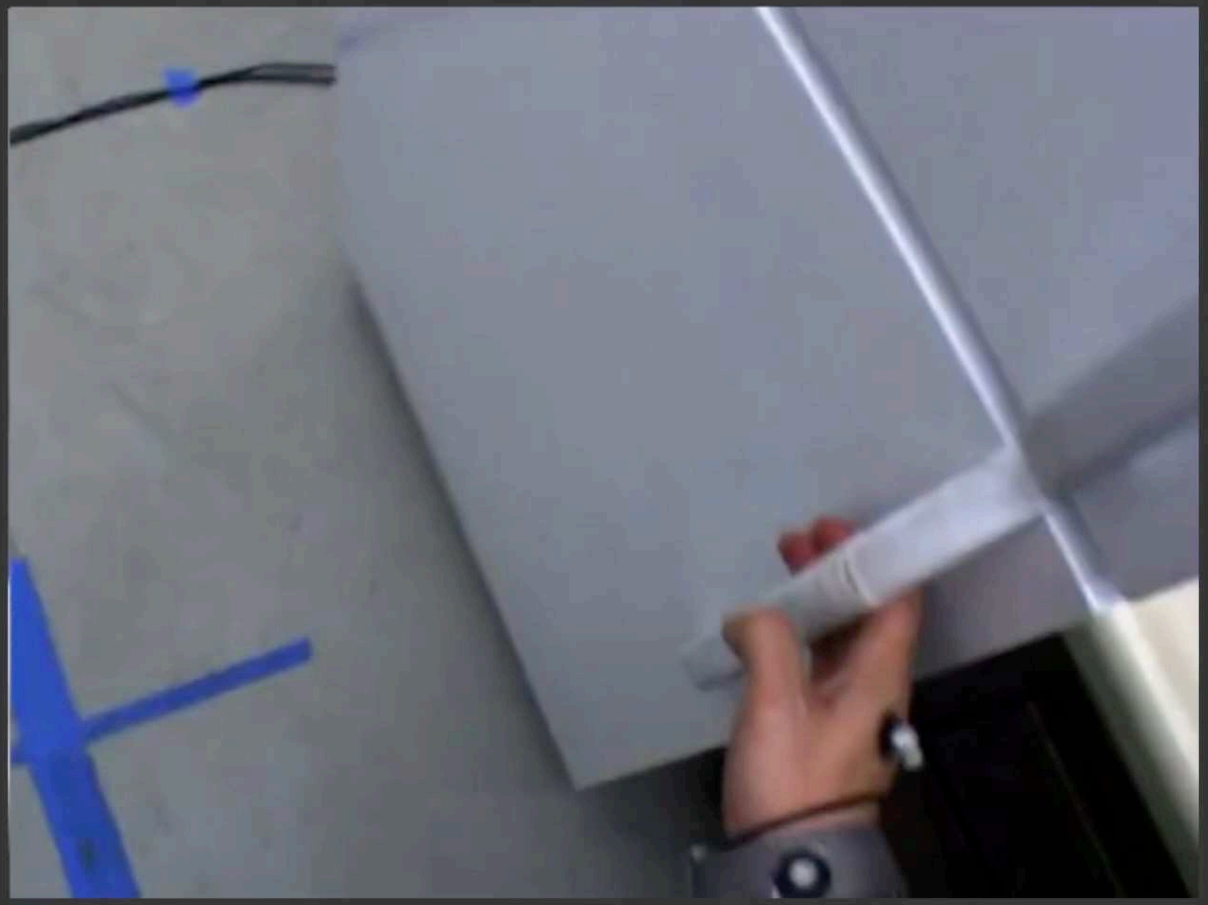


- DDLSTM
- Multi-modal UDA
- Retro-Actions

- Temporal Binding

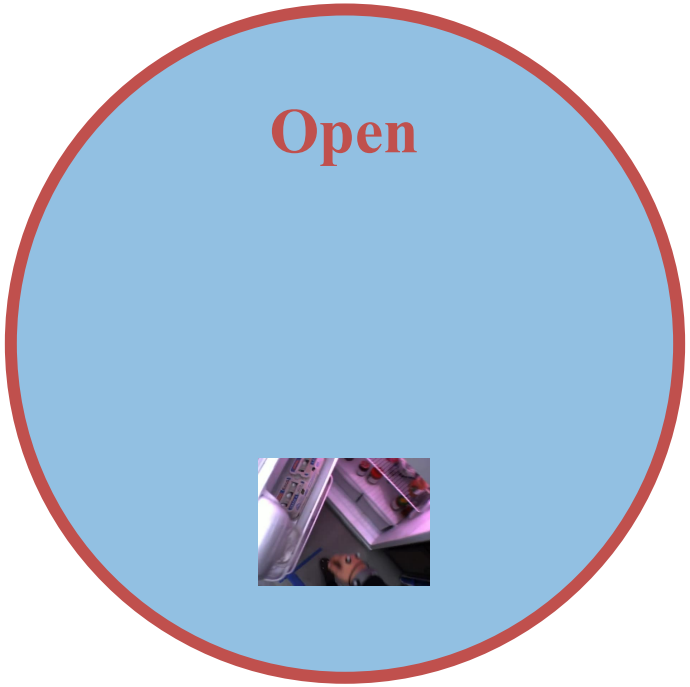
- **Multi-Verb Labels**
- Part-of-Speech
- Adverbs

The *Verbs* Dilemma

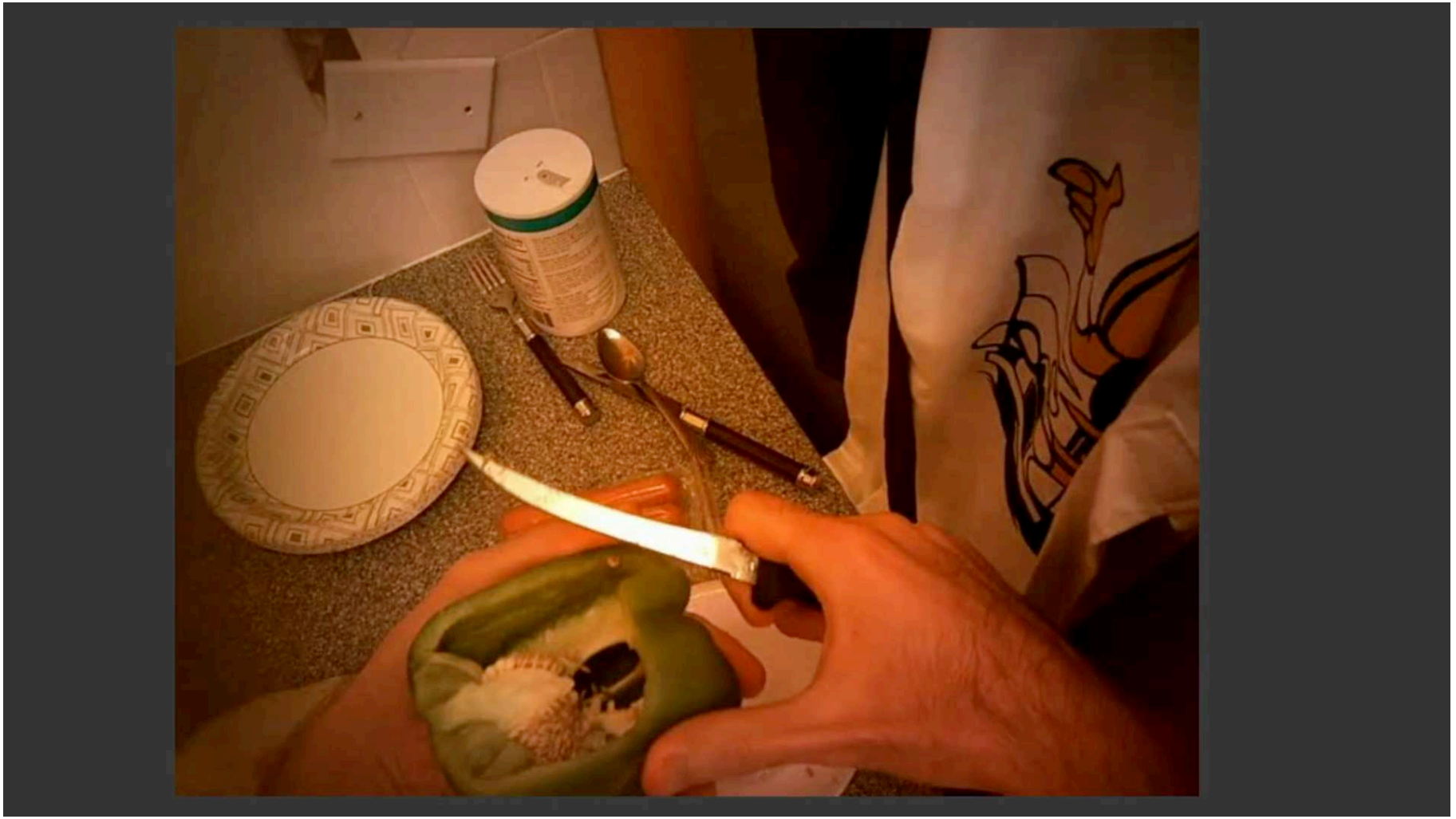


M Wray and D Damen (2019). Learning Visual Actions Using Multiple Verb-Only Labels. BMVC

The *Verbs* Dilemma

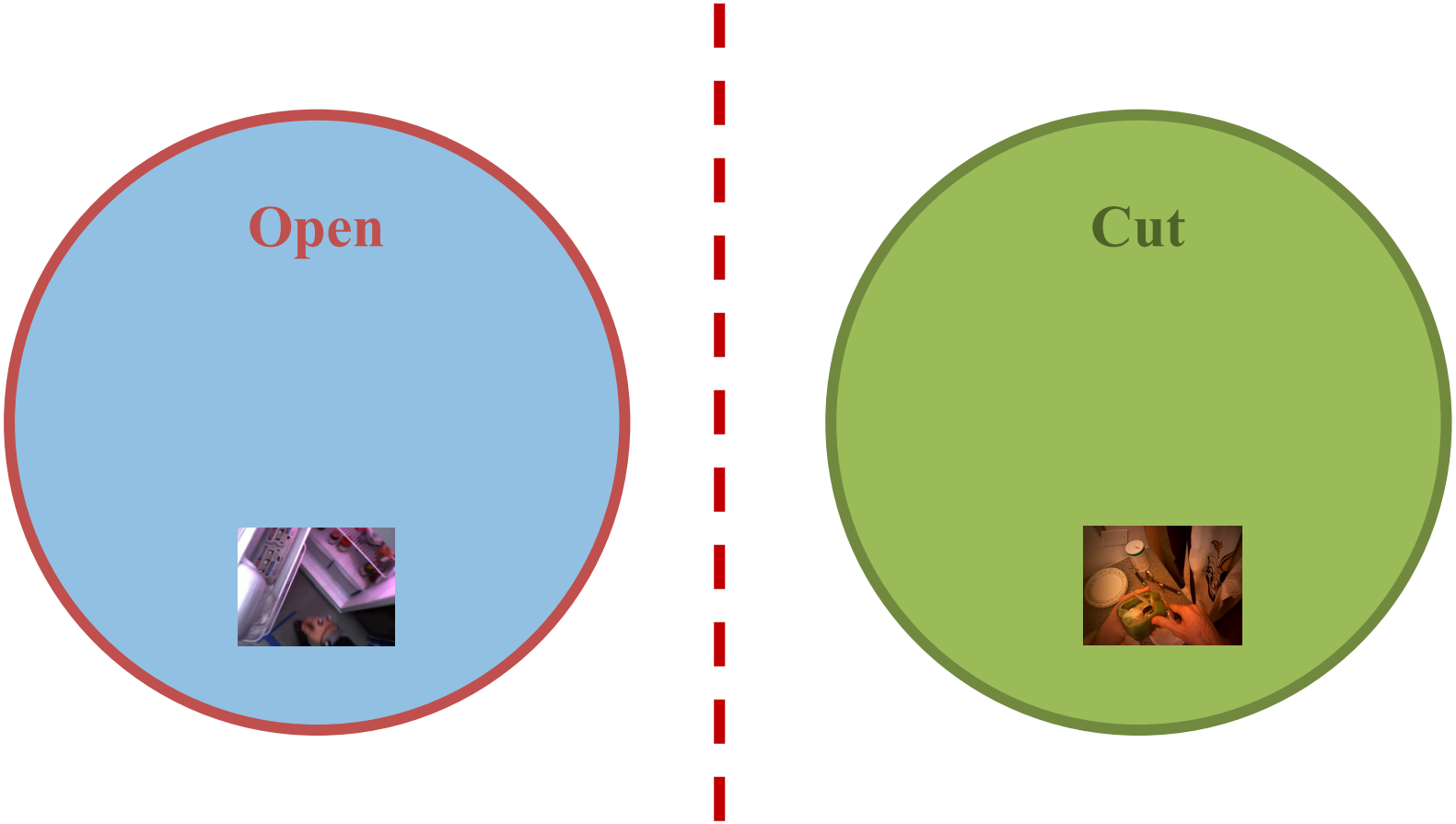


The *Verbs* Dilemma

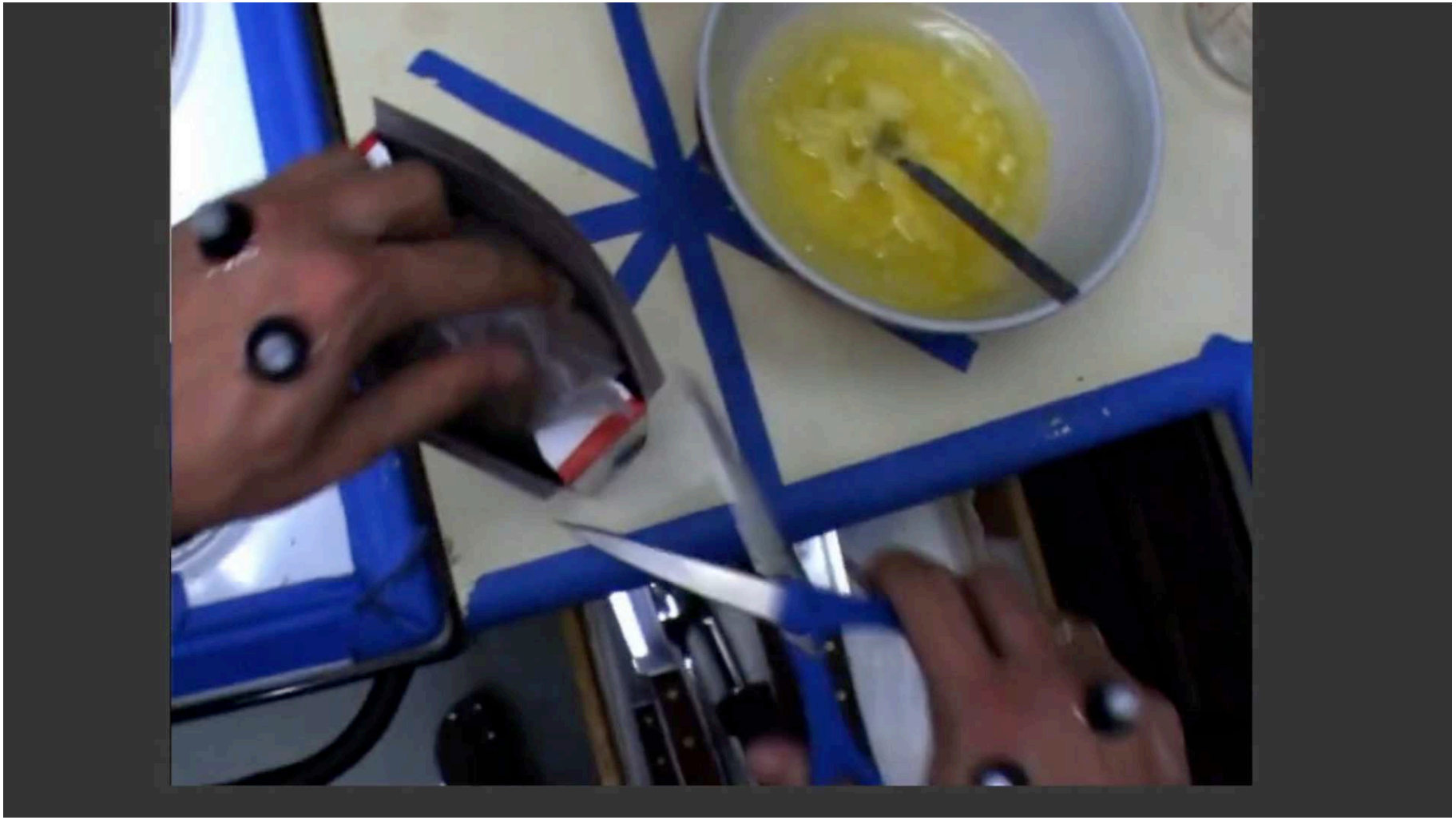


M Wray and D Damen (2019). Learning Visual Actions Using Multiple Verb-Only Labels. BMVC

The *Verbs* Dilemma

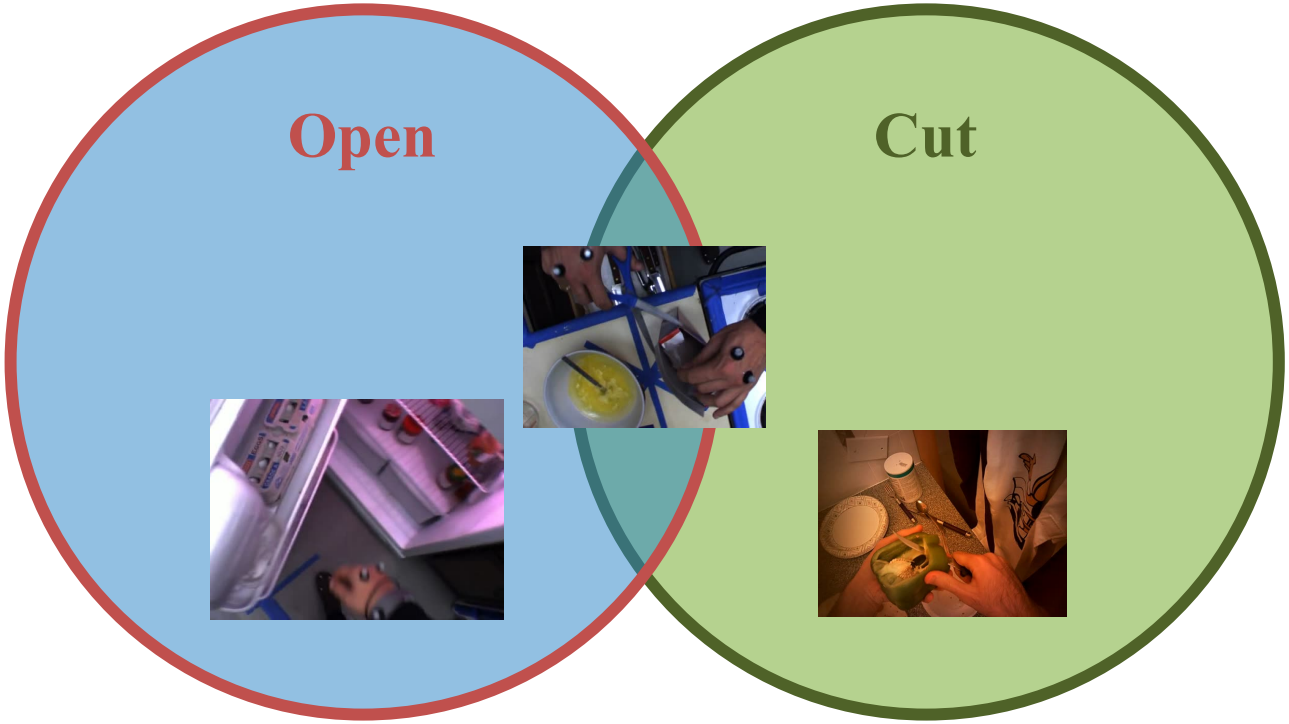


The *Verbs* Dilemma



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

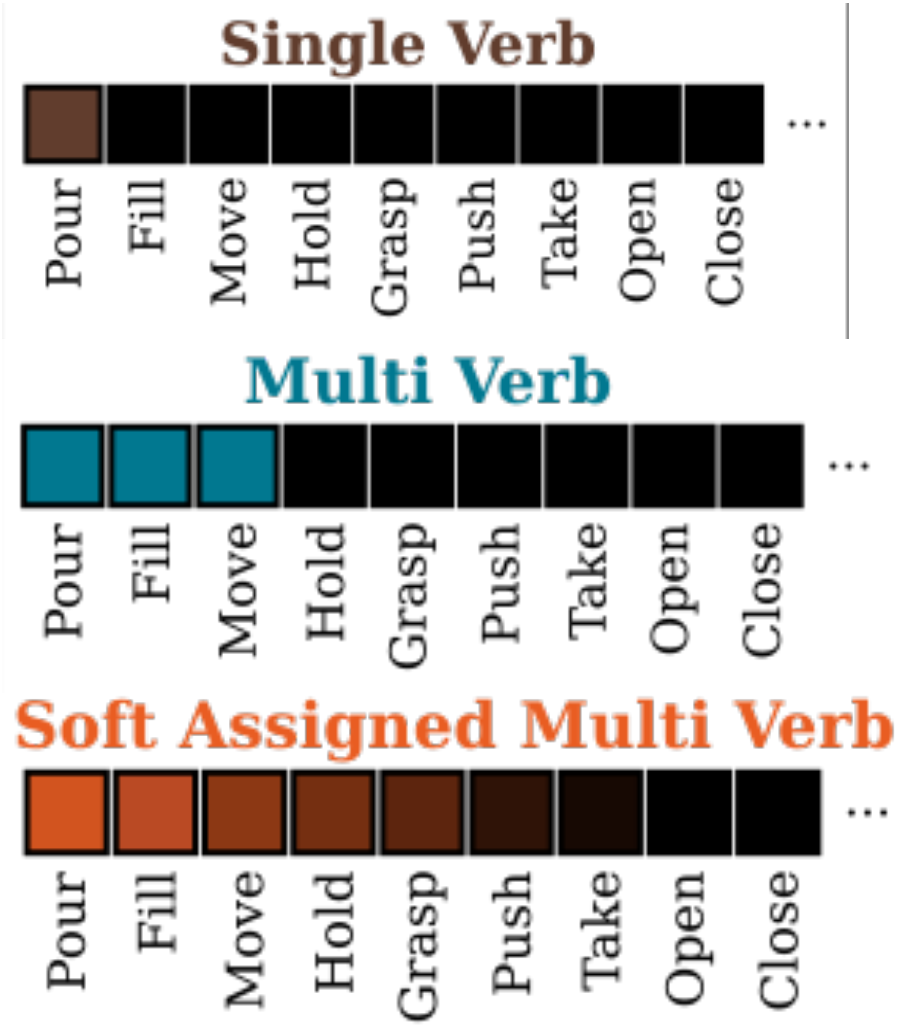
The *Verbs* Dilemma



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

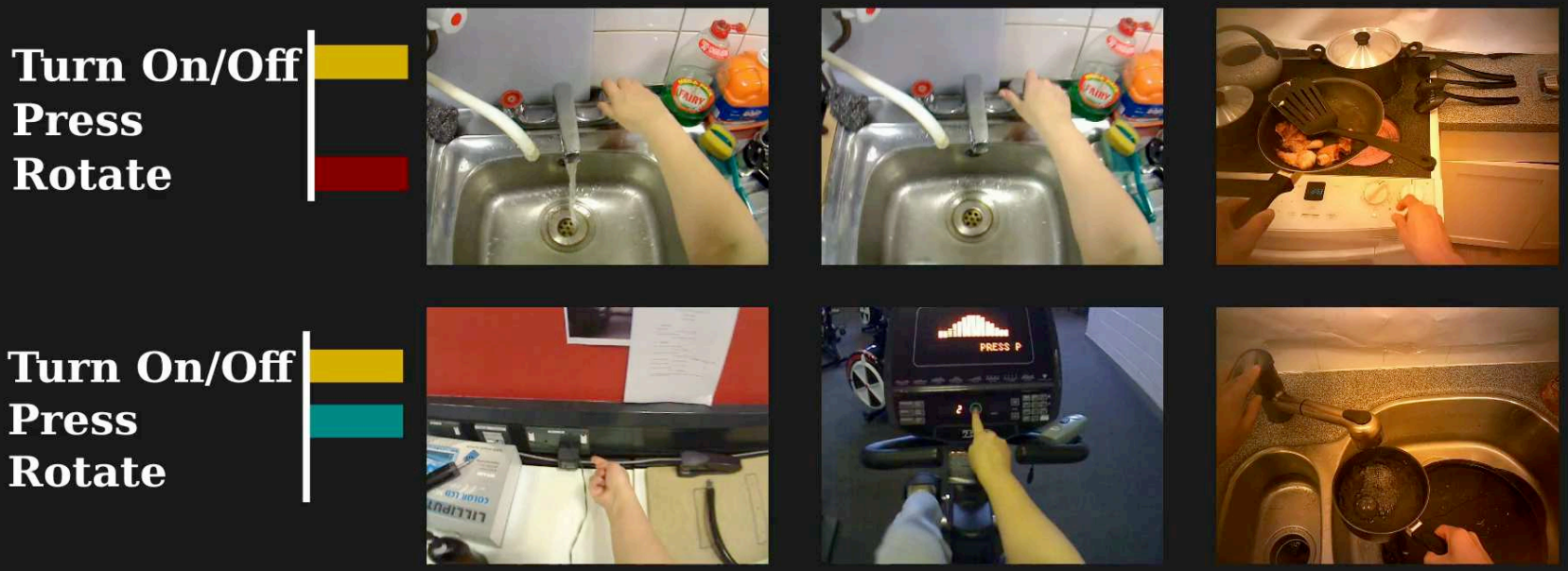
The *Verbs* Dilemma



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

The *Verbs* Dilemma

Top 3 retrieved classes across all datasets.

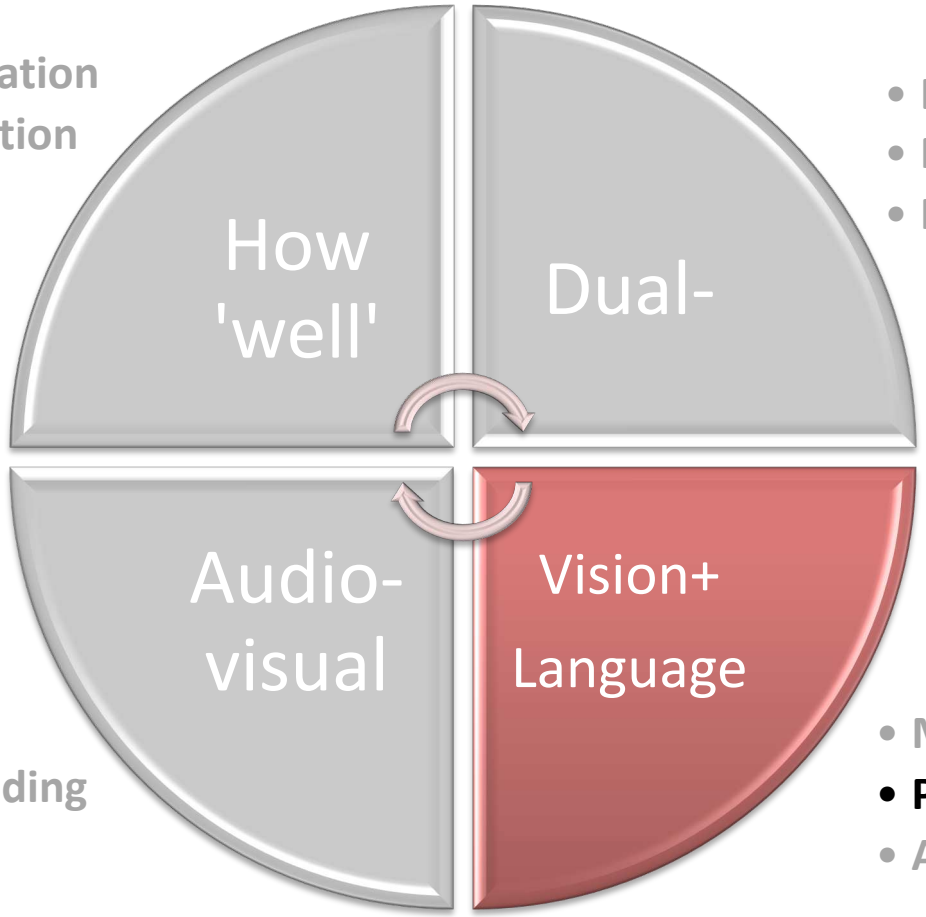


Labelling Method can differentiate turn On/Off tap by pressing and by rotating.

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

Fine-Grained Object Interactions

- Skill Determination
- Action Completion



- DDLSTM
- Multi-modal UDA
- Retro-Actions

- Temporal Binding

- Multi-Verb Labels
- **Part-of-Speech**
- Adverbs

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

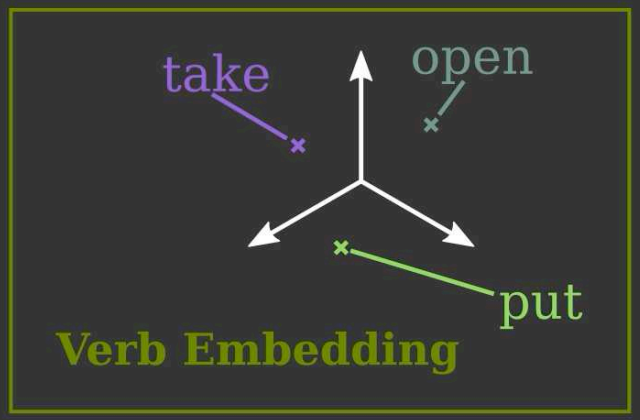
Fine-Grained Action Retrieval

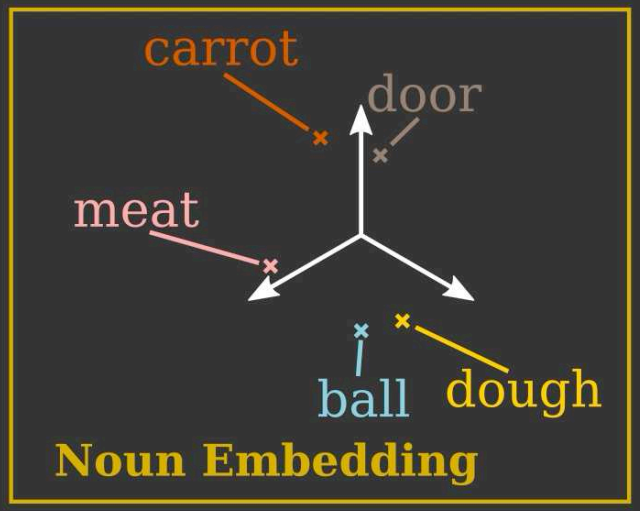
with: Michael Wray
Gabriela Csurka
Diane Larlus

We embed the video and representations

[put]

[meat, ball, dough]

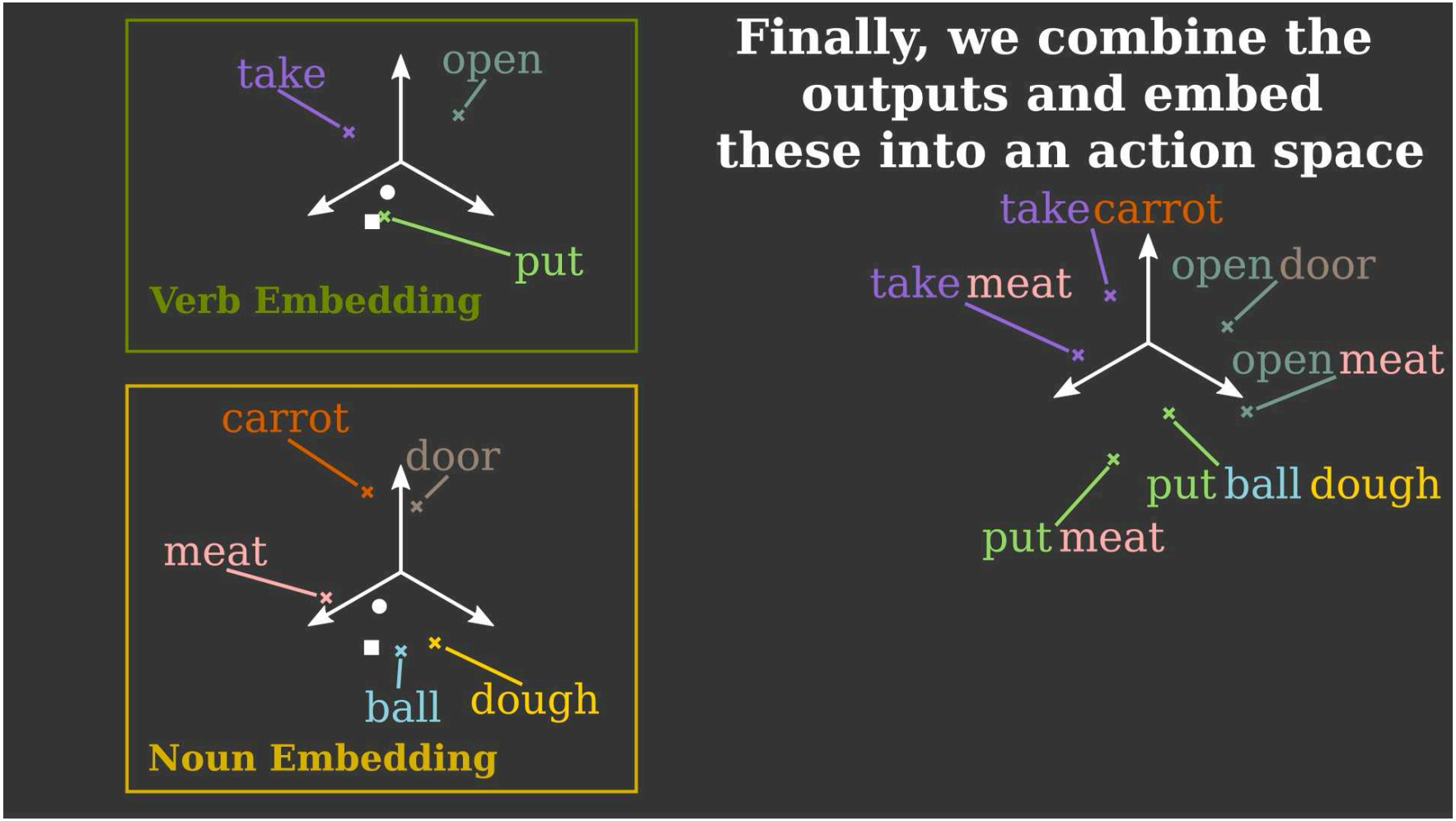




M Wray, D Larlus, G Csurka, D Damen (2019). Fine-Grained Action Retrieval through Multiple Parts-of-Speech Embeddings. ICCV

Fine-Grained Action Retrieval

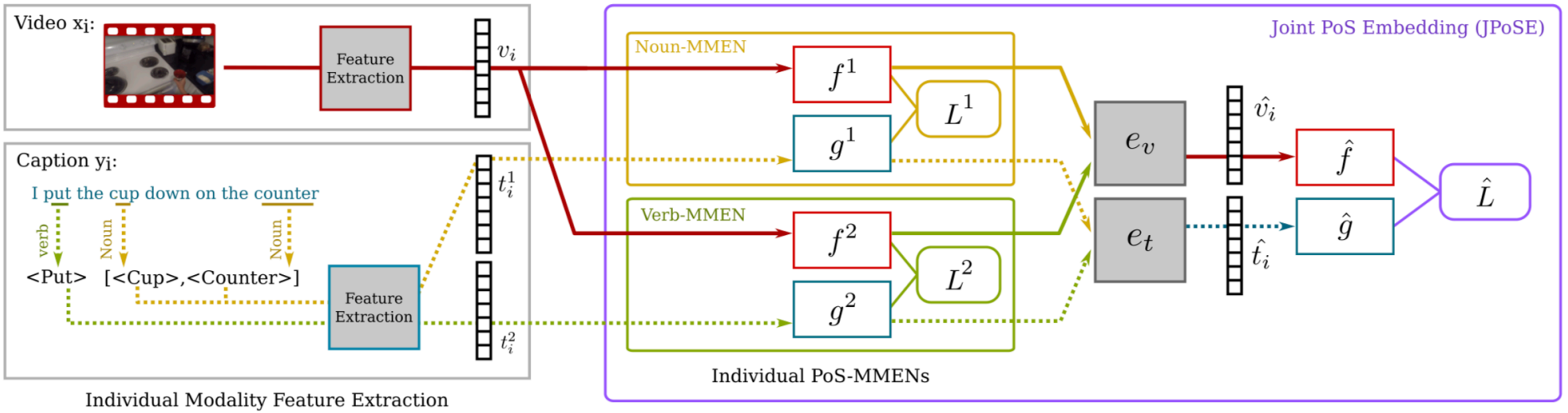
with: Michael Wray
Gabriela Csurka
Diane Larlus



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

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.

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

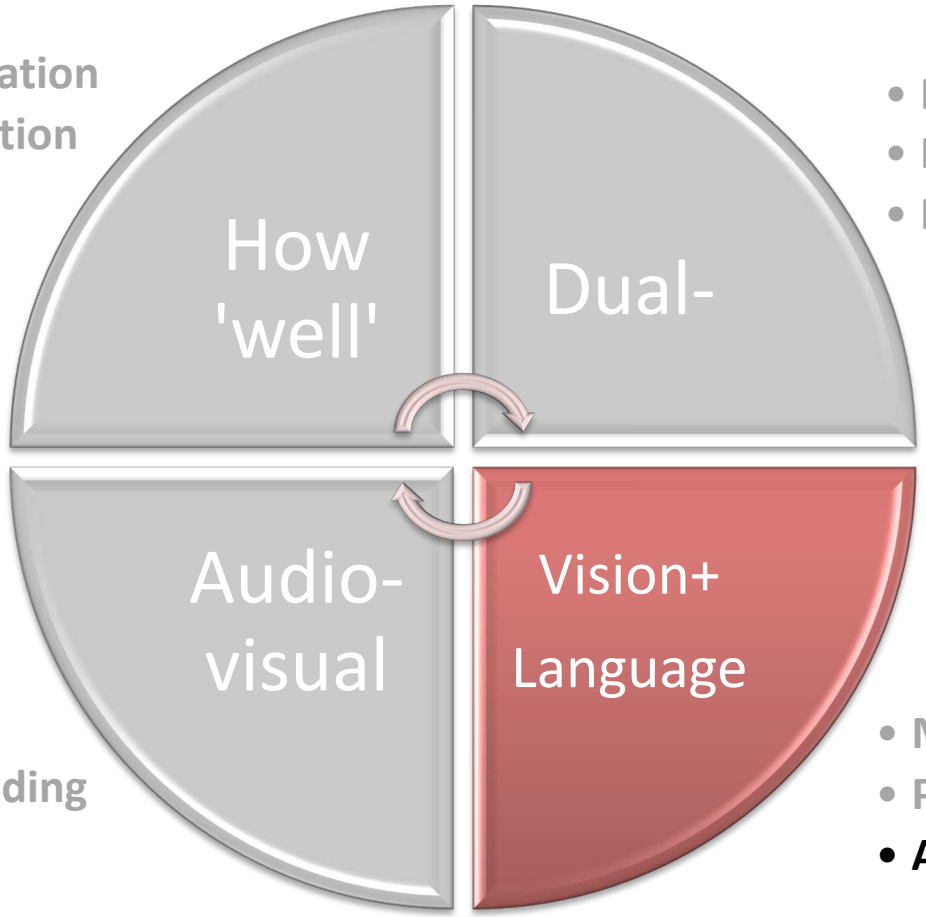
Fine-Grained Action Retrieval

with: Michael Wray
Gabriela Csurka
Diane Larlus

Maximum activation examples for a neuron in a noun PoS Embedding (Cutting Board) - Figure 4

Fine-Grained Object Interactions

- Skill Determination
- Action Completion



- DDLSTM
- Multi-modal UDA
- Retro-Actions

- Temporal Binding

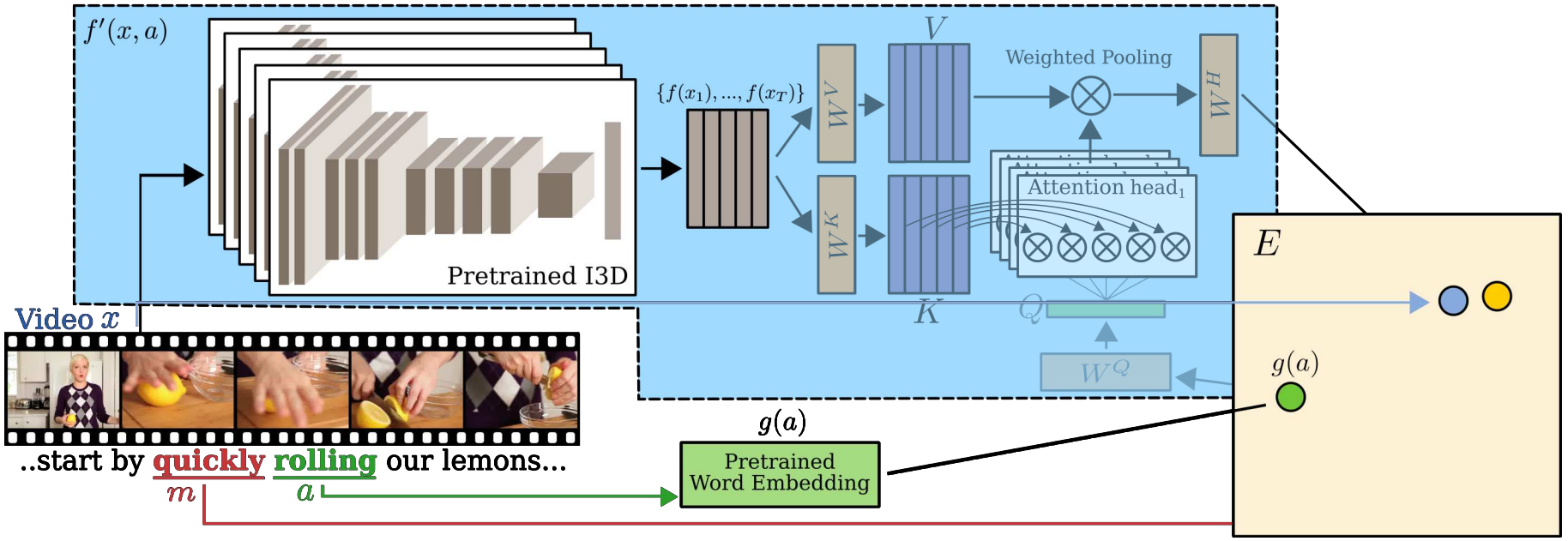
- Multi-Verb Labels
- Part-of-Speech
- **Adverbs**

Action Modifiers: Learning from Adverbs in Instructional Videos

with: Hazel Doughty
Ivan Laptev
Walterio Mayol-Cuevas

Action Modifiers: Learning from Adverbs in Instructional Videos

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)

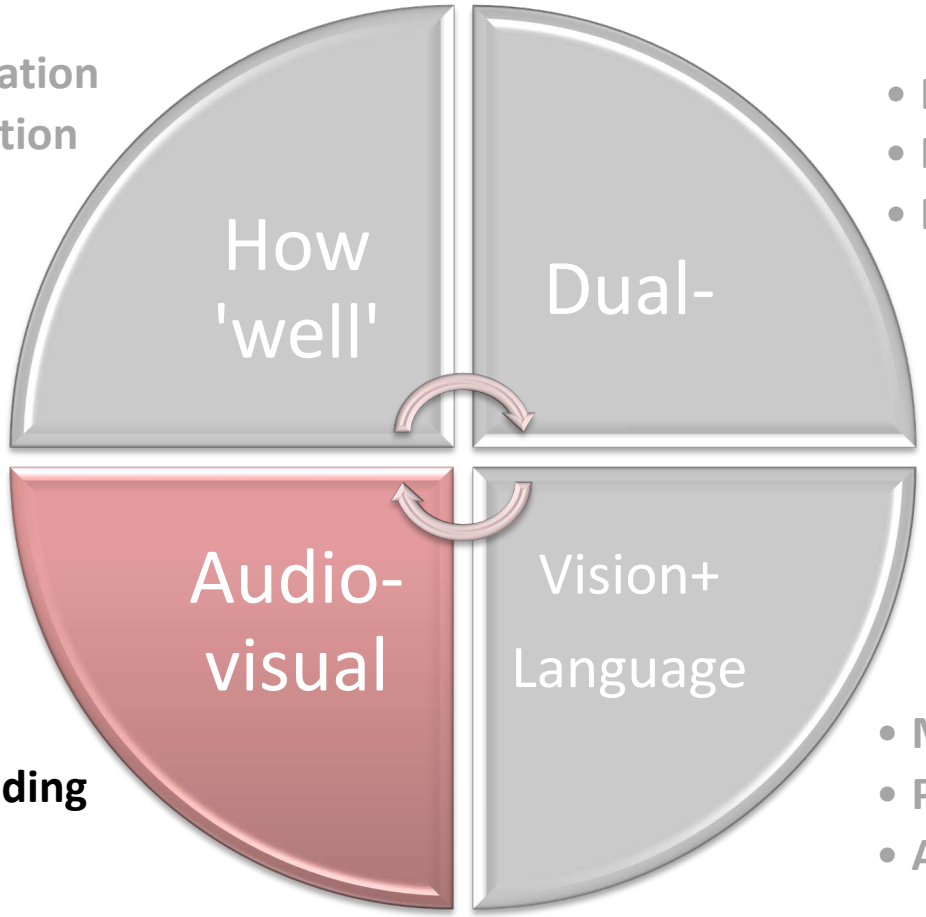
Action Modifiers: Learning from Adverbs in Instructional Videos

with: Hazel Doughty
Ivan Laptev
Walterio Mayol-Cuevas

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

Fine-Grained Object Interactions

- Skill Determination
- Action Completion



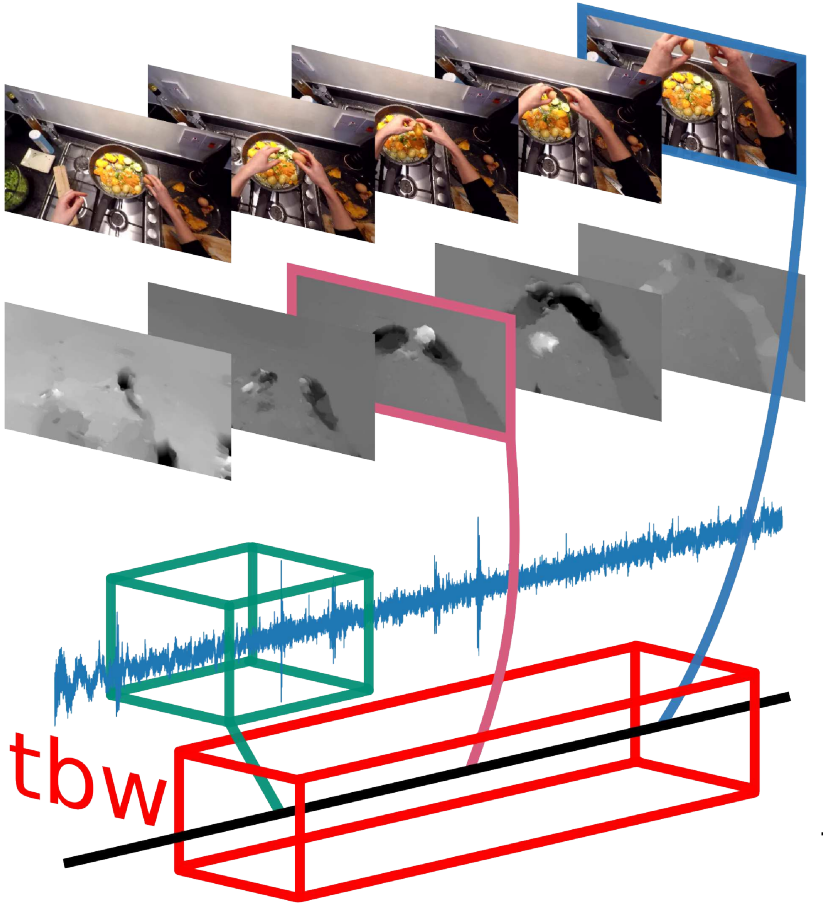
- DDLSTM
- Multi-modal UDA
- Retro-Actions

- **Temporal Binding**

- Multi-Verb Labels
- Part-of-Speech
- Adverbs

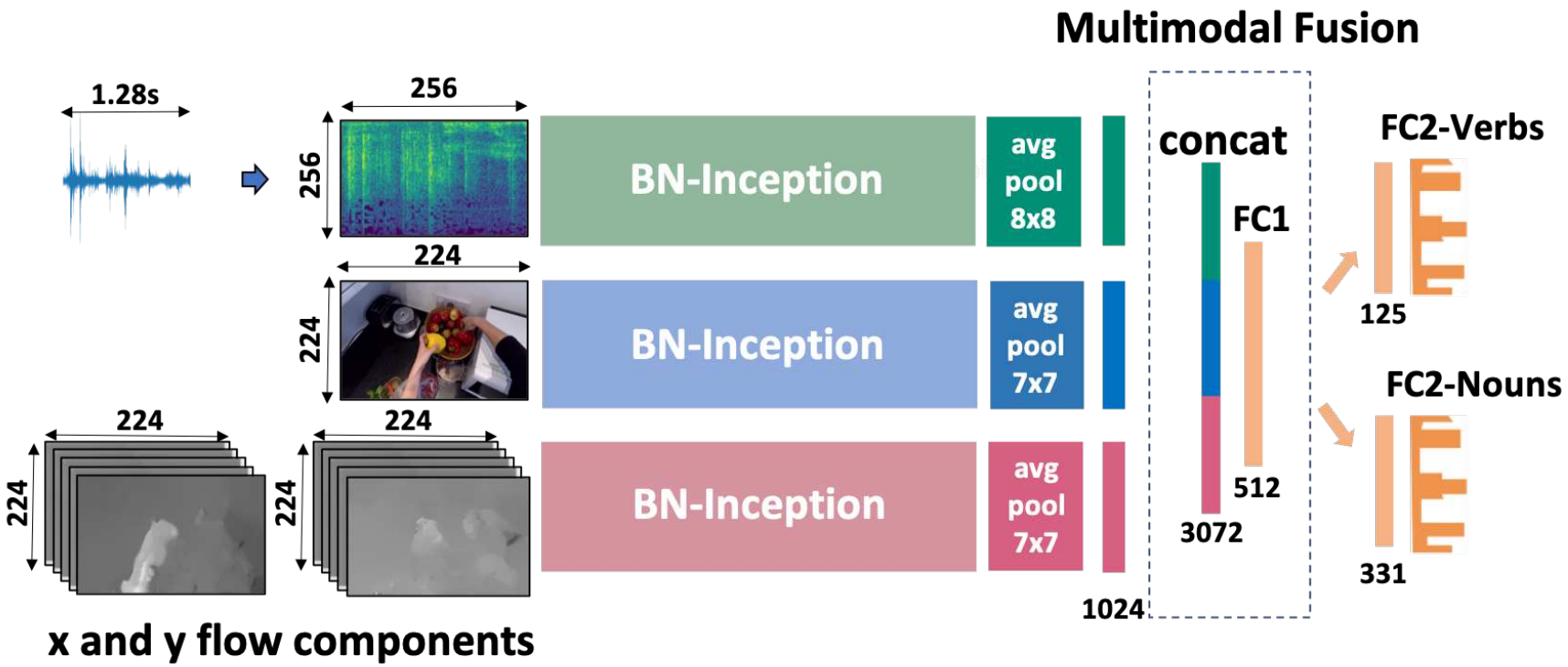
Audio-Visual Temporal Binding for Egocentric Action Recognition

with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



Audio-Visual Temporal Binding for Egocentric Action Recognition

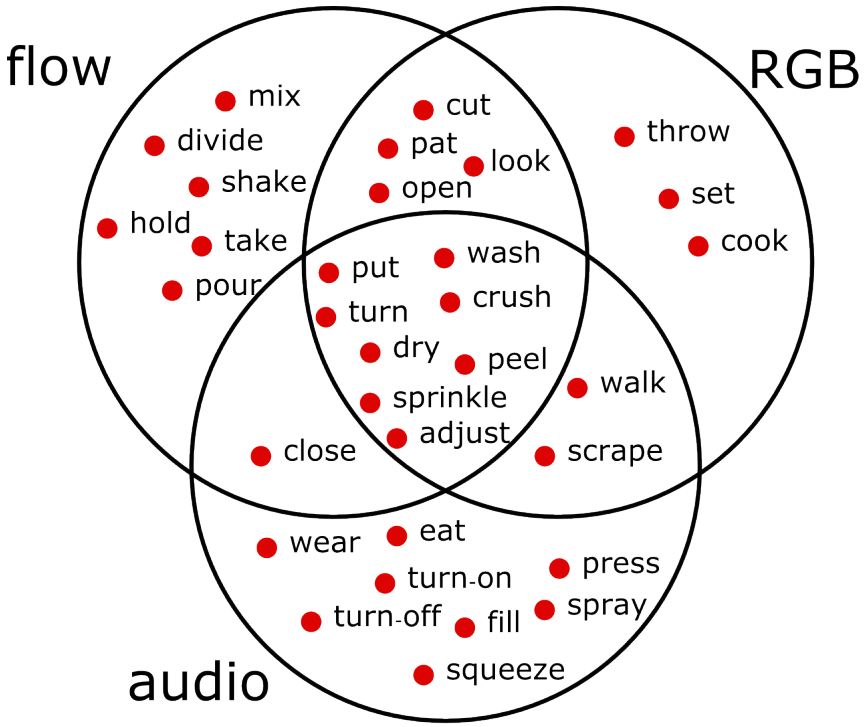
with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



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

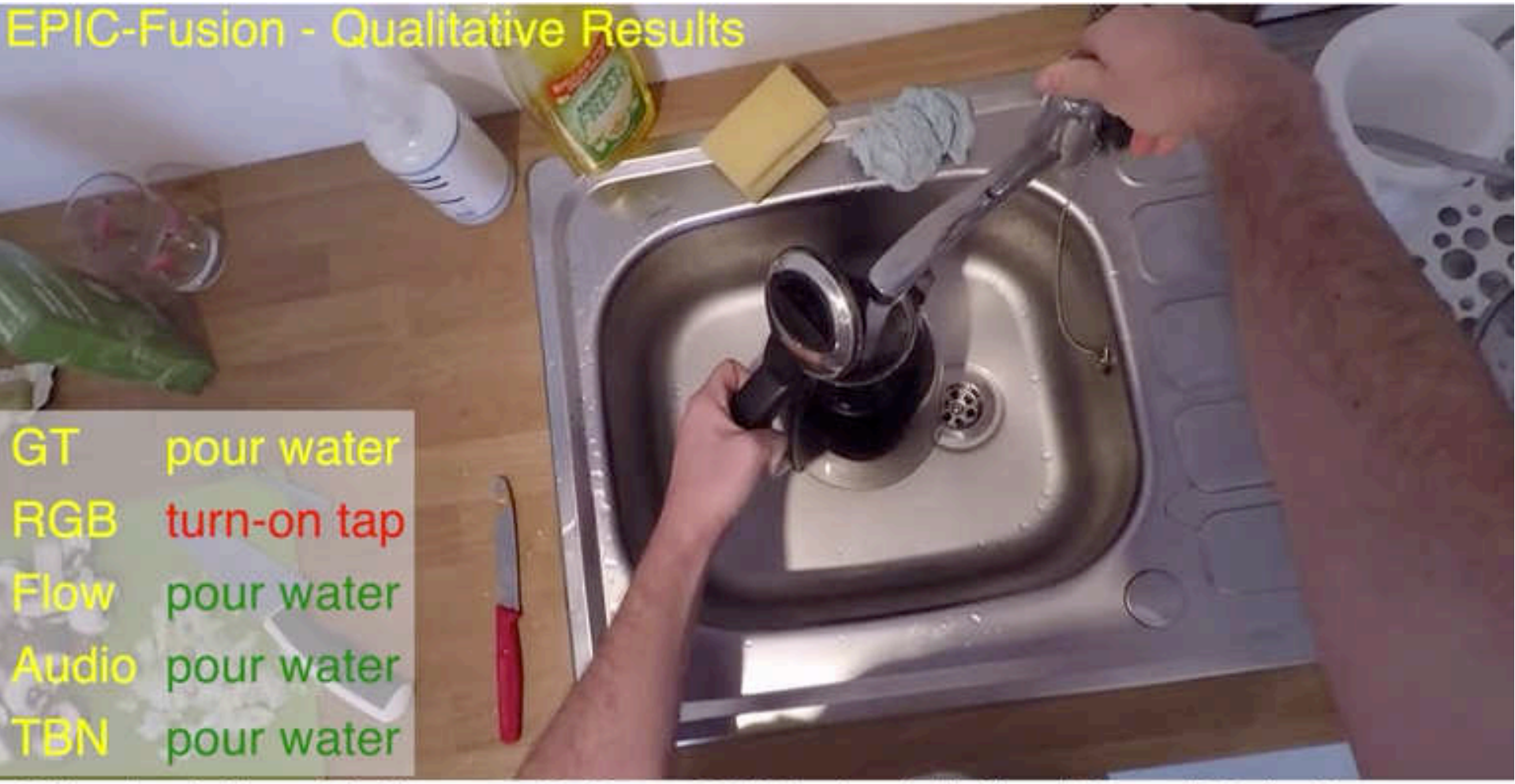
with: Vangelis Kazakos
Arsha Nagrani
Andrew Zisserman



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



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

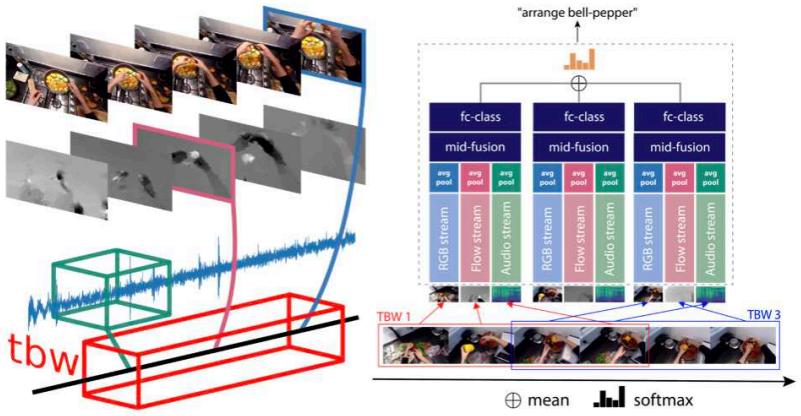
Audio-Visual Temporal Binding for Egocentric Action Recognition

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]

June 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

EPIC KITCHENS

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

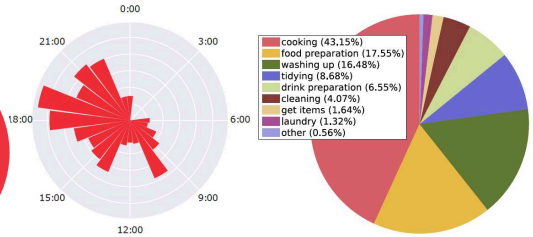

Will Price


Michael Wray

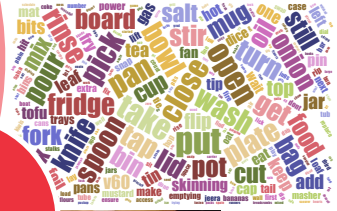
Scaling Egocentric Vision

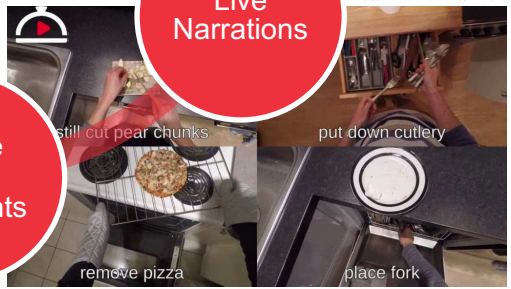
Data Collection

Native Environment, Natural Interactions



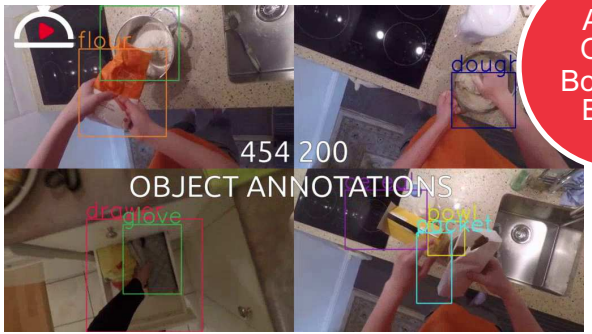
Live Narrations





Dense Action Segments

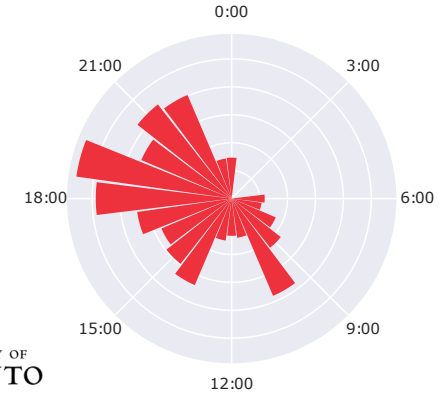
Active Object Bounding Boxes



Benchmark and Challenges

Data Collection

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

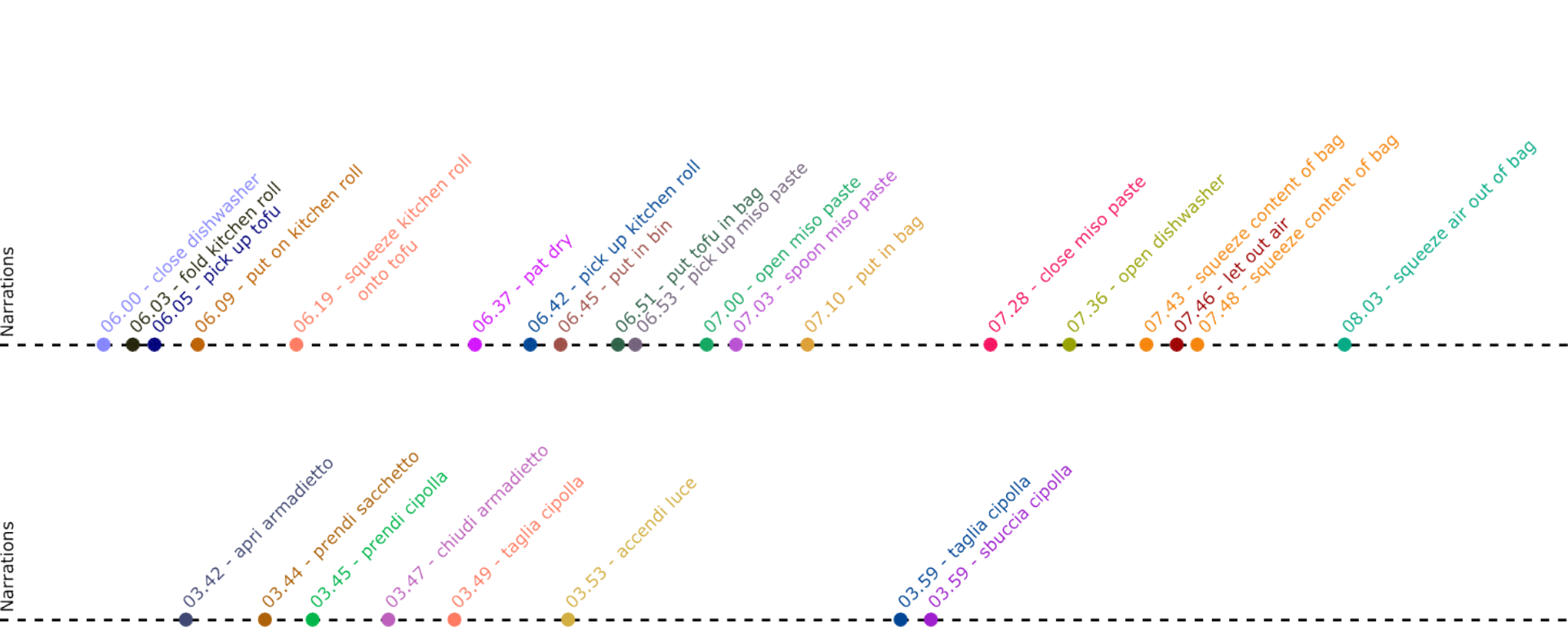


UNIVERSITY OF TORONTO

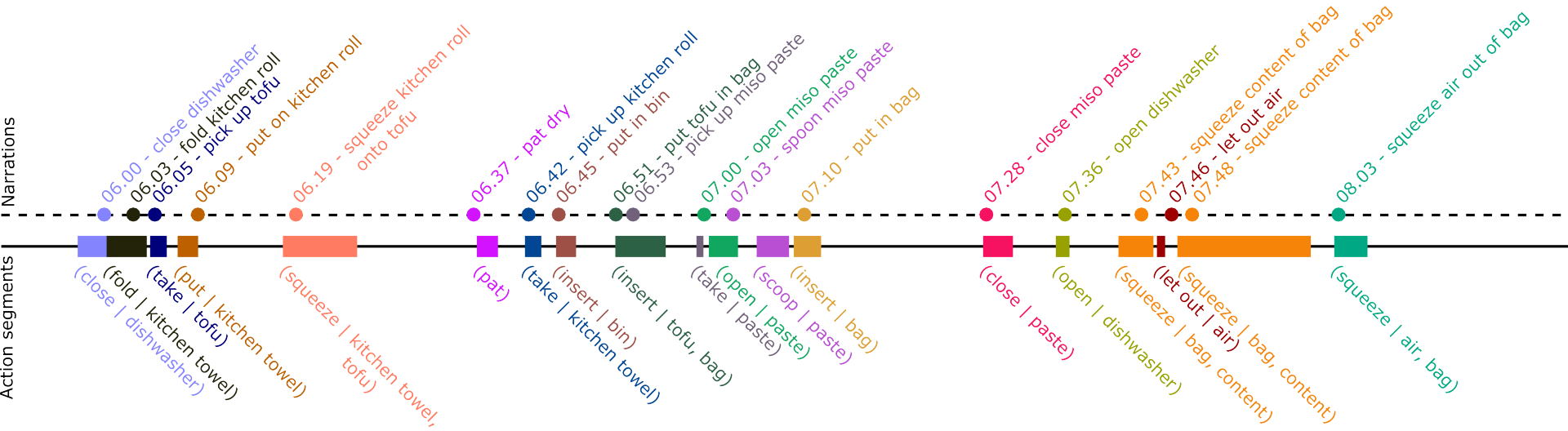
University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

Annotations (1) - Narrations



Narrations to Action Segments



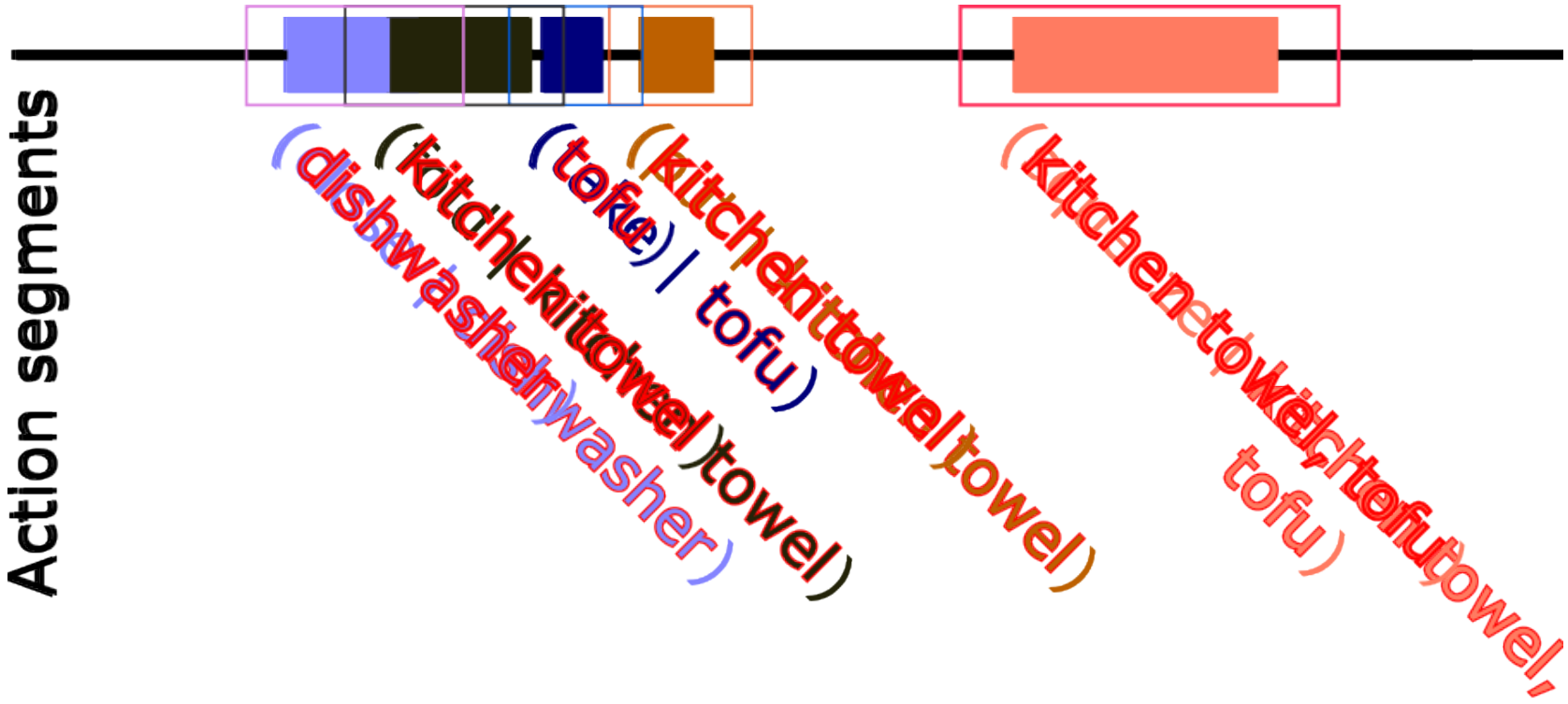
UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

39 000
ACTION SEGMENTS

Annotations (3) – Object Bounding Boxes



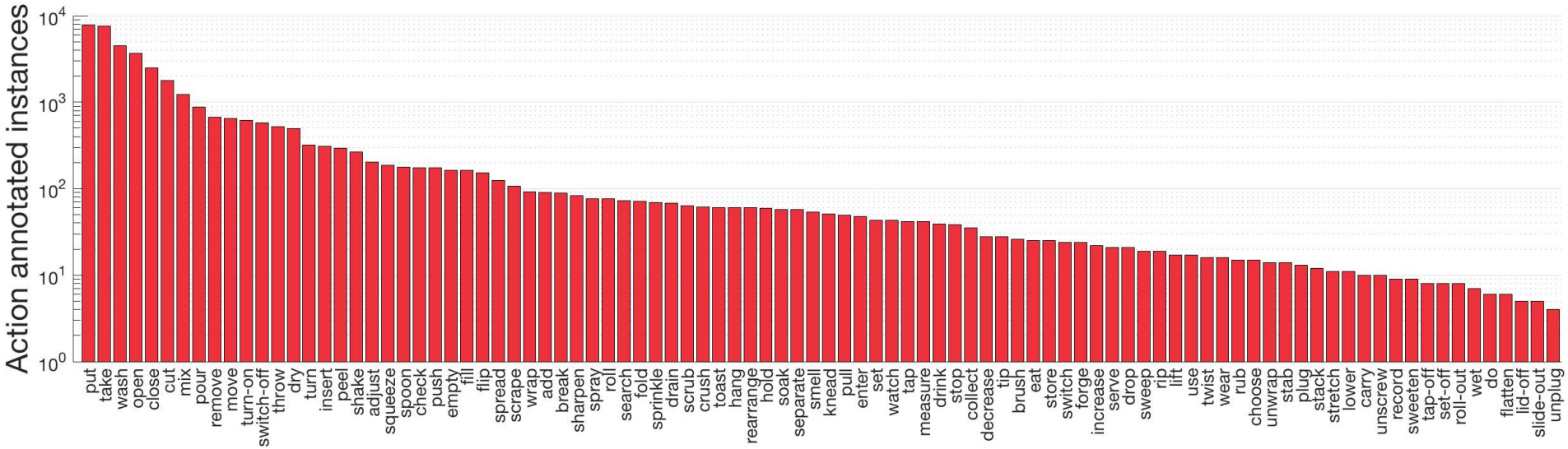
454 200
OBJECT ANNOTATIONS

Annotations (4) – Verb and Noun Classes

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

- 125 verb classes
- 331 noun classes

Annotations Statistics



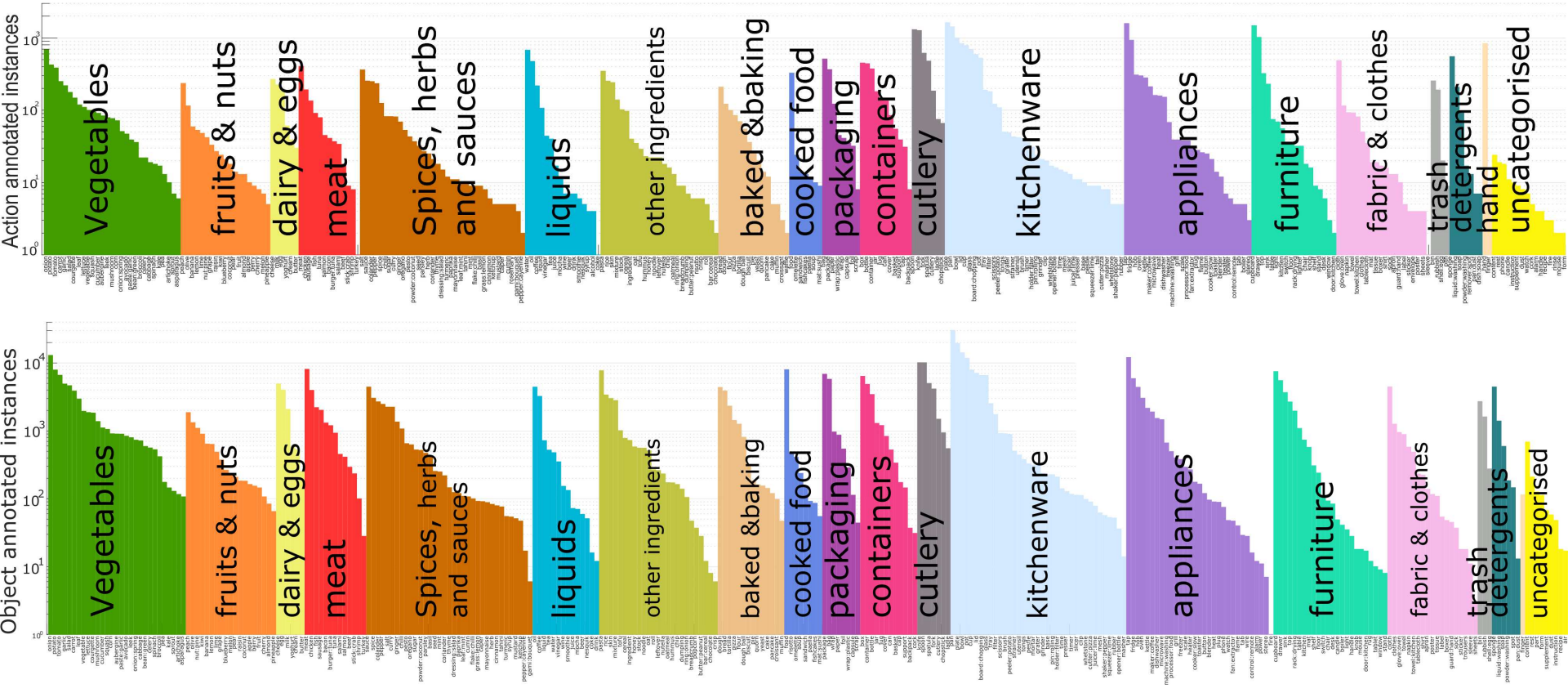
UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

EPIC KITCHENS

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

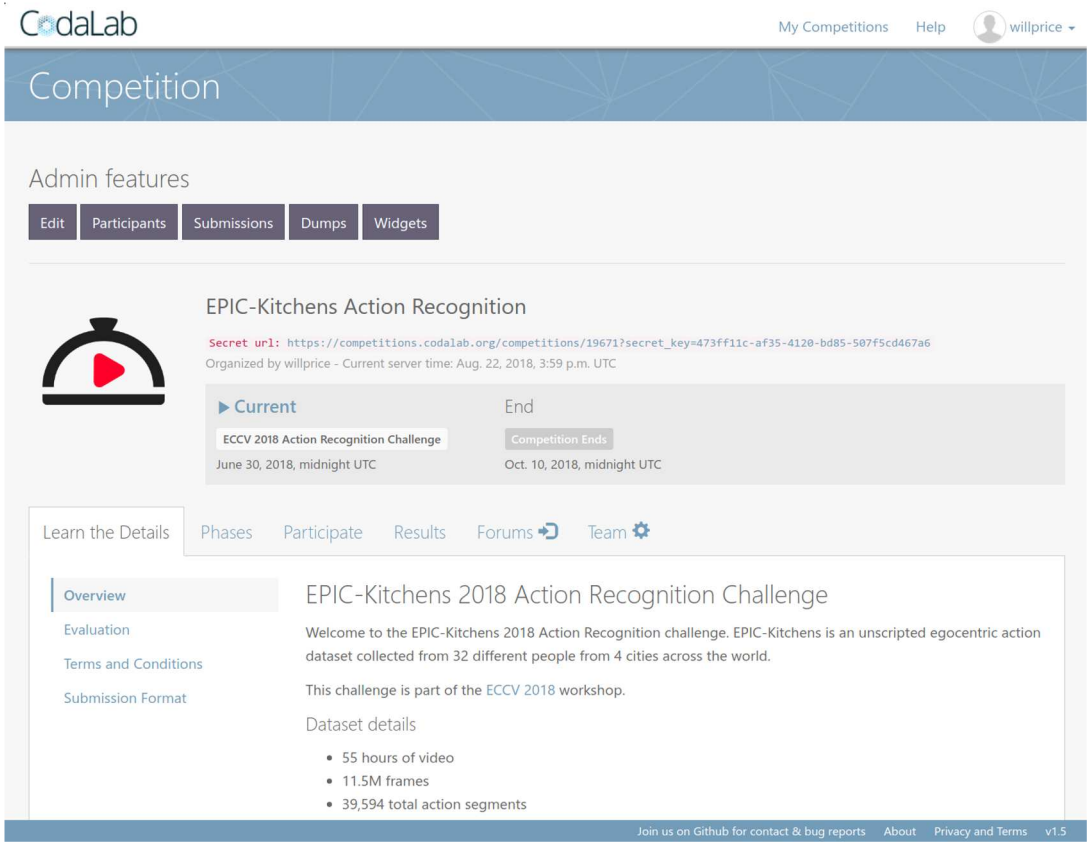
Table 4: Statistics of test splits: seen (S1) and unseen (S2) kitchens

	#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

Open Challenges

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.

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

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

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

EPIC KITCHENS

Action Recognition Challenge

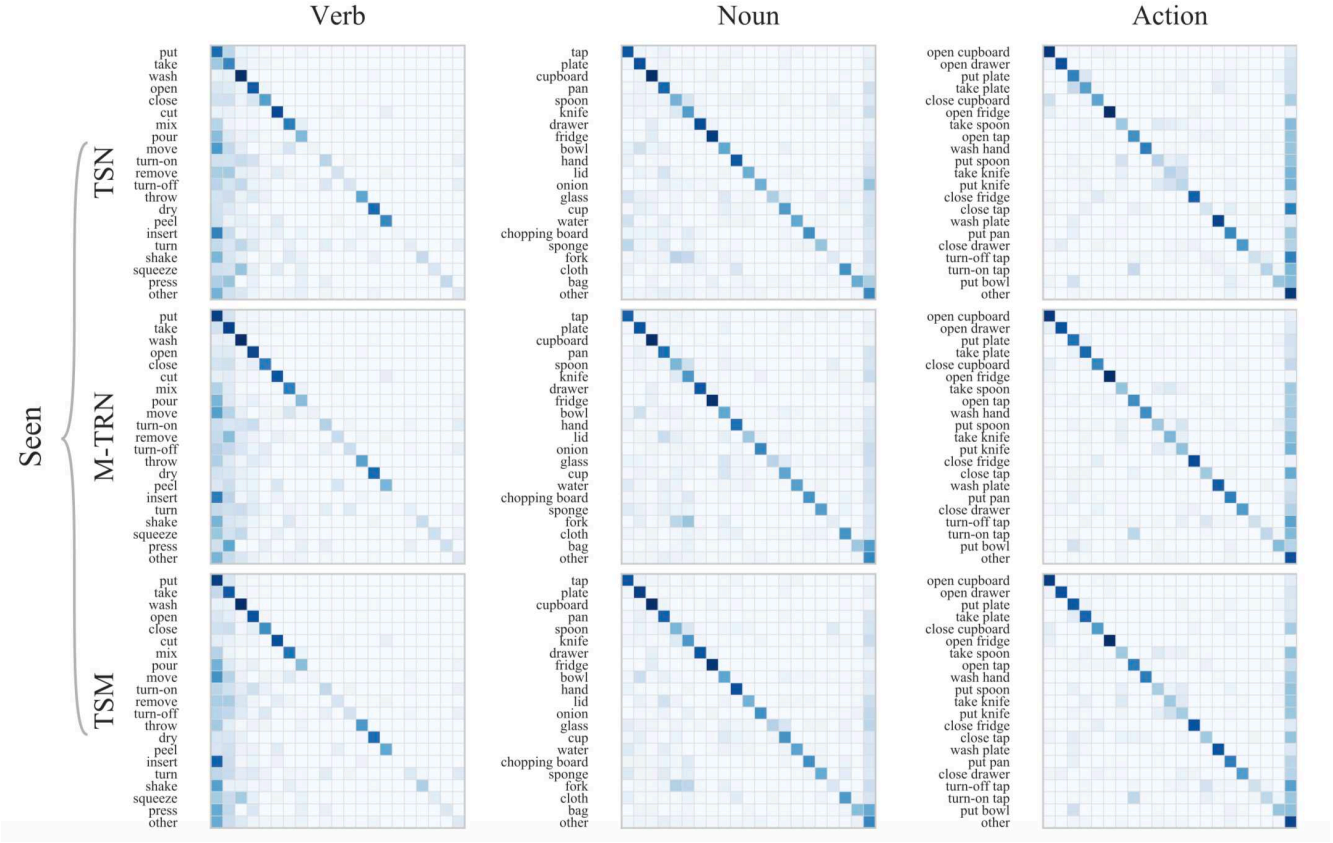
Seen Kitchens (S1)															
#	User	Entries	Date of Last Entry	Team Name	Top-1 Accuracy (%)			Top-5 Accuracy (%)			Precision (%)			Recall (%)		
					Verb ▲	Noun ▲	Action ▲	Verb ▲	Noun ▲	Action ▲	Verb ▲	Noun ▲	Action ▲	Verb ▲	Noun ▲	Action ▲
1	action_banks	11	03/02/20		63.22 (7)	46.49 (5)	41.34 (1)	87.34 (9)	69.98 (5)	63.50 (1)	53.75 (5)	43.18 (4)	24.28 (1)	40.27 (8)	42.82 (5)	25.67 (1)
2	wasun	11	02/22/20	UTS_BAIDU	69.85 (1)	51.14 (1)	41.18 (2)	90.67 (2)	74.44 (1)	62.13 (2)	58.85 (4)	45.06 (1)	23.45 (4)	45.23 (4)	48.38 (1)	25.35 (2)
3	aptx4869lm	12	01/30/20	GT-WISC-MPI	68.51 (2)	49.96 (2)	38.75 (3)	89.33 (4)	72.30 (3)	58.99 (3)	51.04 (10)	44.00 (3)	23.70 (3)	43.70 (5)	47.32 (2)	23.92 (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 (8)	41.26 (7)	20.84 (5)	46.73 (2)	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 (4)	60.73 (3)	44.89 (2)	24.01 (2)	46.81 (1)	43.88 (4)	22.92 (4)
6	Sudhakaran	30	08/10/19	FBK_HuPBA	63.34 (6)	44.75 (6)	35.54 (6)	89.01 (5)	69.88 (6)	57.18 (5)	63.21 (1)	42.26 (5)	19.76 (7)	37.77 (12)	41.28 (7)	21.19 (6)
7	tnet	16	03/06/20		64.74 (5)	44.75 (6)	34.81 (7)	88.82 (6)	68.98 (7)	55.49 (7)	52.68 (7)	42.18 (6)	20.76 (6)	45.29 (3)	41.97 (6)	20.06 (7)
8	antoninofurnari	1	07/19/19		56.93 (11)	43.05 (7)	33.06 (8)	85.68 (15)	67.12 (8)	55.32 (8)	50.42 (12)	39.84 (8)	18.91 (8)	37.82 (10)	38.11 (8)	19.12 (8)
9	cvg_uni_bonn	7	01/07/20		58.63 (9)	41.44 (8)	29.81 (9)	88.73 (7)	66.57 (9)	48.64 (11)	50.32 (13)	37.67 (9)	18.30 (9)	41.37 (6)	37.61 (9)	18.07 (10)

UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

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

	GFLOP/s		Params (M)	
Model	RGB	Flow	RGB	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

Models Released

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

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

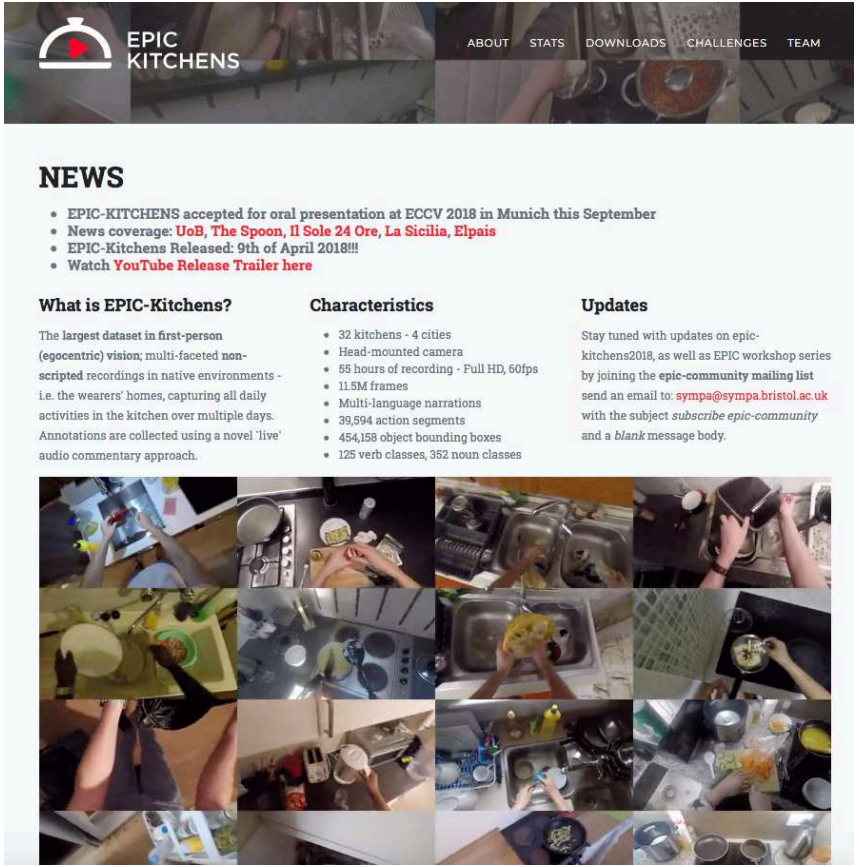
UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA

More?

http://epic-kitchens.github.io



UNIVERSITY OF TORONTO

University of BRISTOL

UNIVERSITÀ degli STUDI di CATANIA