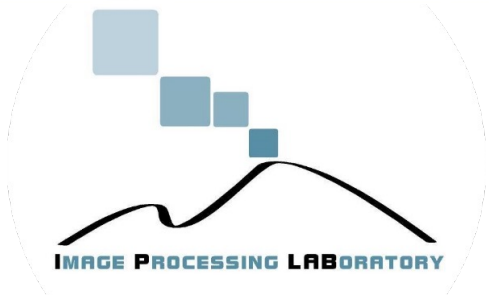
Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Università di Catania

NEXT VISION

Spin-off of the University of Catania



Procedure Understanding from Egocentric Videos

Giovanni Maria Farinella

FPV @ Image Processing Laboratory - http://iplab.dmi.unict.it/fpv

Next Vision - http://www.nextvisionlab.it/

Department of Mathematics and Computer Science - University of Catania

giovanni.farinella@unict.it - www.dmi.unict.it/farinella

Egocentric (First Person) Vision

Egocentric vision or **first-person vision** is a sub-field of computer vision that entails **analyzing images, videos captured by a wearable device**, which is typically worn on the head or on the chest and **naturally approximates the visual field of the camera wearer**.

Consequently, **visual data capture the part of the scene on which the user focuses to carry out the task** at hand and offer a valuable perspective **to understand the user's intents and activities and their context in a naturalistic setting.**

This research area is sometimes referred with the name "Wearable Vision".

What's Relevant in EgoVision? A top-down approach

An Outlook into the Future of Egocentric Vision

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

Abstract *What will the future be? We wonder!* In this survey, we explore the gap between current research in egocentric vision and the ever-anticipated future, where wearable computing, with outward facing cameras and digital overlays, is expected to be integrated in our every day lives. To understand this gap, the article starts by envisaging the future through character-based stories, showcasing through examples the limitations of current technology. We then provide a mapping between this future and previously defined research tasks. For each task, we survey its seminal works, current state-of-the-art methodologies and available datasets, then reflect on shortcomings that limit its applicability to future research. Note that this survey focuses on software models for egocentric vision, independent of any specific hardware. The paper concludes with recommendations for areas of immediate explorations so as to unlock our path to the future always-on, personalised and life-enhancing egocentric vision.

Keywords Egocentric Vision, Future, Survey, Localisation, Scene Understanding, Recognition, Anticipation, Gaze Prediction, Social Understanding, Body Pose Estimation, Hand and Hand-Object Interaction, Person Identification, Summarisation, Dialogue, Privacy

Contents

*: Equal Contribution/First Author
†: Equal Senior Author
C. Plizzari, G. Goletto and T. Tommasi, Politecnico di Torino, Italy · A. Furnari, F. Ragusa and G. M. Farinella, University of Catania, Italy · S. Bansal and D. Damen, University of Bristol, UK. E-mail: Tatiana.Tommasi@polito.it

1 Introduction

Designing and building tools able to support human activities, improve quality of life, and enhance individuals' abilities to achieve their goals is the ever-lasting aspiration of our species. Among all inventions, digital computing has already had a revolutionary effect on human history. Of particular note is mobile technology, currently integrated in our lives through hand-held devices, i.e. *mobile smart phones*. These are nowadays the de facto for outdoor navigation, capturing static and moving footage of our everyday and connecting us to both familiar and novel connections and experiences.

However, humans have been dreaming about the next-version of such mobile technology — wearable computing, for a considerable amount of time. Imaginations

Immagine the Future

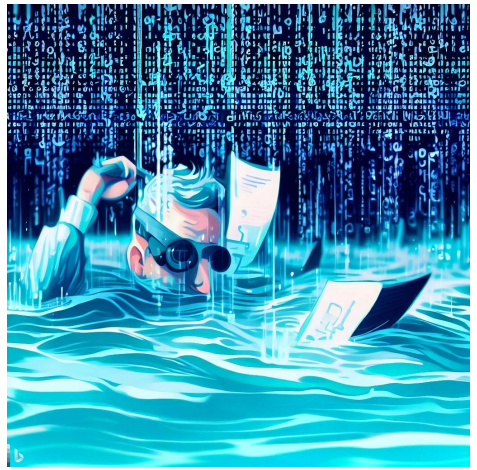
Write Stories in Different Scenarios

Extract Important Tasks from the Stories

Go in-depth with Tasks and Datasets

Rather than being extensive, we considered **seminal** and **state-of-the-art** works

A lot of data!



Plizzari, C., Goletto, G., Furnari, A., Bansal, S., Ragusa, F., Farinella, G. M., Damen., D. & Tommasi, T. (2023). An Outlook into the Future of Egocentric Vision. International Journal of Computer Vision.

What's Relevant in Egovision? Scenarios

An Outlook into the Future of Egocentric Vision

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

Abstract *What will the future be? We wonder!* In this survey, we explore the gap between current research in egocentric vision and the ever-anticipated future, where wearable computing, with outward facing cameras and digital overlays, is expected to be integrated in our every day lives. To understand this gap, the article starts by envisaging the future through character-based stories, showcasing through examples the limitations of current technology. We then provide a mapping between this future and previously defined research tasks. For each task, we survey its seminal works, current state-of-the-art methodologies and available datasets, then reflect on shortcomings that limit its applicability to future research. Note that this survey focuses on software models for egocentric vision, independent of any specific hardware. The paper concludes with recommendations for areas of immediate explorations so as to unlock our path to the future always-on, personalised and life-enhancing egocentric vision.

Keywords Egocentric Vision, Future, Survey, Localisation, Scene Understanding, Recognition, Anticipation, Gaze Prediction, Social Understanding, Body Pose Estimation, Hand and Hand-Object Interaction, Person Identification, Summarisation, Dialogue, Privacy

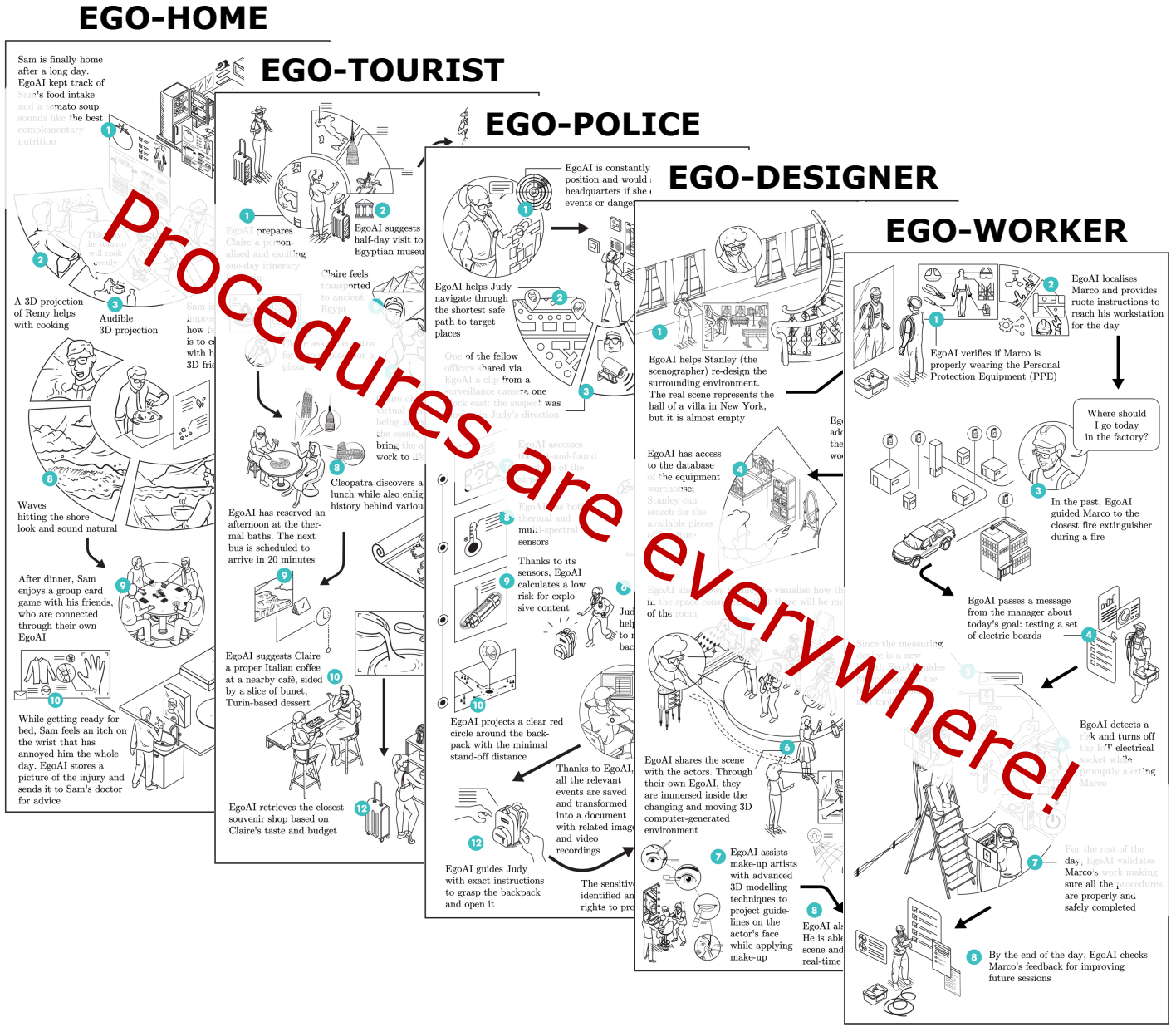
Contents

*: Equal Contribution/First Author
†: Equal Senior Author
C. Plizzari, G. Goletto and T. Tommasi, Politecnico di Torino, Italy · A. Furnari, F. Ragusa and G. M. Farinella, University of Catania, Italy · S. Bansal and D. Damen, University of Bristol, UK. E-mail: Tatiana.Tommasi@polito.it

1 Introduction

Designing and building tools able to support human activities, improve quality of life, and enhance individuals' abilities to achieve their goals is the ever-lasting aspiration of our species. Among all inventions, digital computing has already had a revolutionary effect on human history. Of particular note is mobile technology, currently integrated in our lives through hand-held devices, i.e. *mobile smart phones*. These are nowadays the de facto for outdoor navigation, capturing static and moving footage of our everyday and connecting us to both familiar and novel connections and experiences.

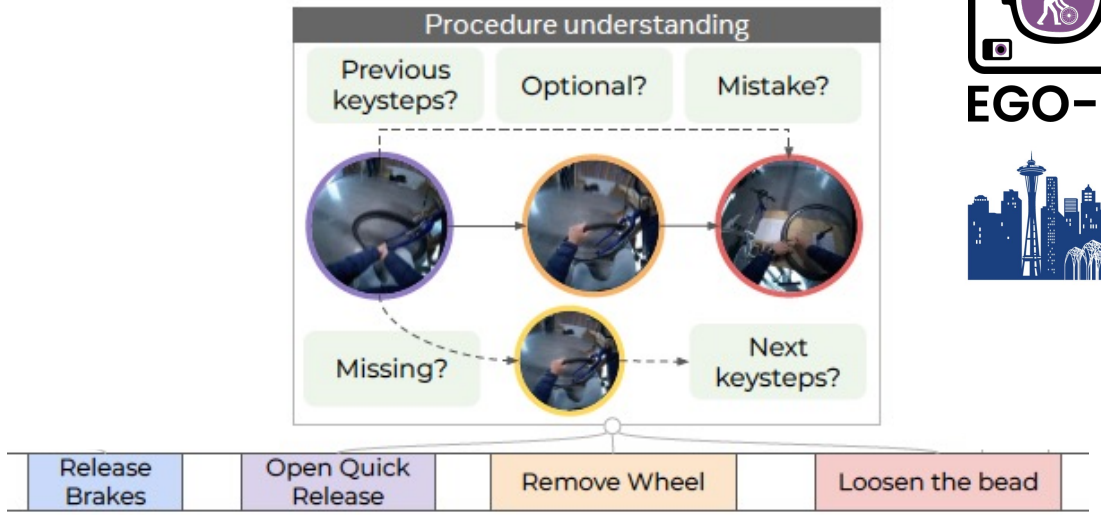
However, humans have been dreaming about the next-version of such mobile technology — wearable computing, for a considerable amount of time. Imaginations



Plizzari, C., Goletto, G., Furnari, A., Bansal, S., Ragusa, F., Farinella, G. M., Damen., D. & Tommasi, T. (2023). An Outlook into the Future of Egocentric Vision. International Journal of Computer Vision.

Procedure Understanding

Procedural activities are sequences of key-steps aimed at achieving specific goals.



EGO-EXO4D

CVPR JUNE 17-21, 2024 SEATTLE, WA

Task: Given a video segment s_t and its previous video segment history, models have to: 1) determine previous keysteps (to be performed before s_t); infer if s_t is 2) optional or 3) a procedural mistake; 4) predict missing keysteps (should have been performed before s_t but were not); and 5) next keysteps (for which dependencies are satisfied).

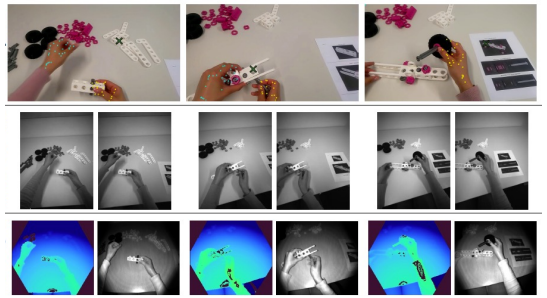
Grauman et al. (2024). Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives. In Conference on Computer Vision and Pattern Recognition (CVPR)

Università di Catania

Lots of Ego-data available for Procedure Understanding

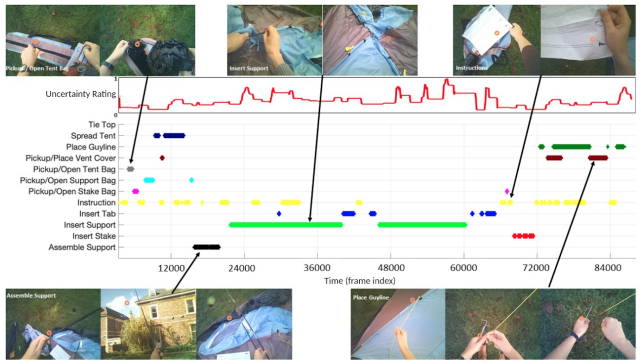
MECCANO

EgoProceL



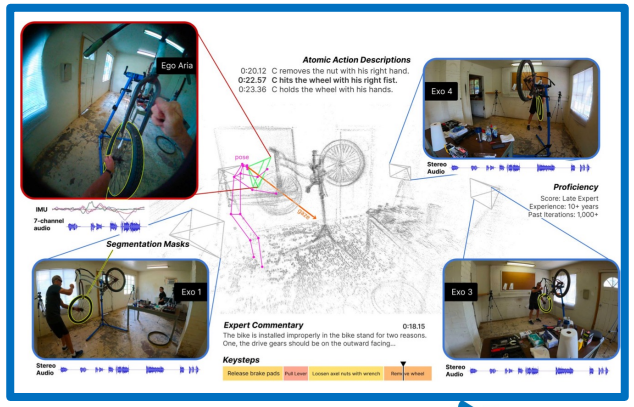
IndustReal

EgoHOS



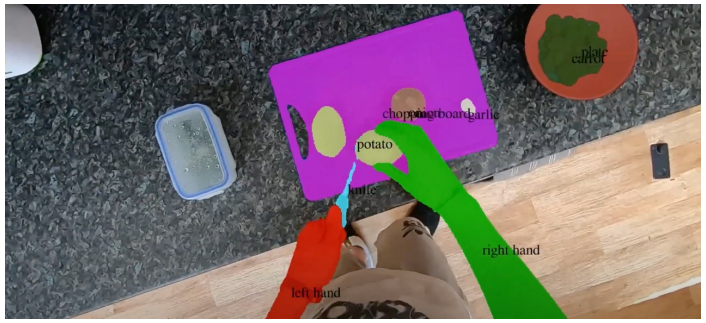
EPIC-Tent

Ego4D

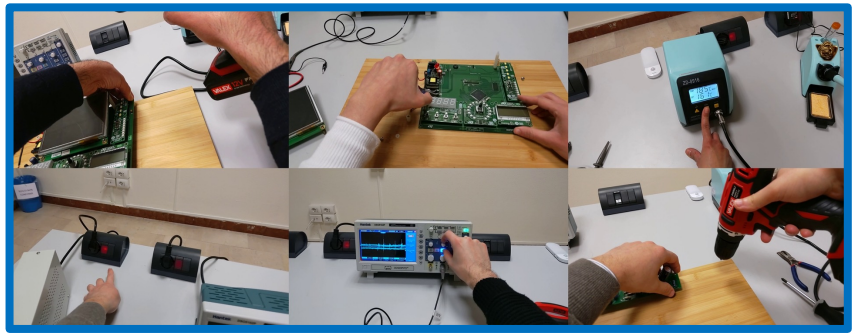


Ego-Exo4D

EPIC-KITCHENS

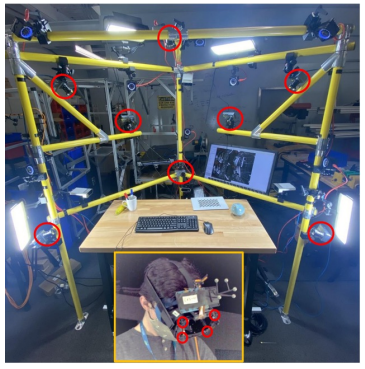


EPIC-KITCHENS VISOR



ENIGMA-51

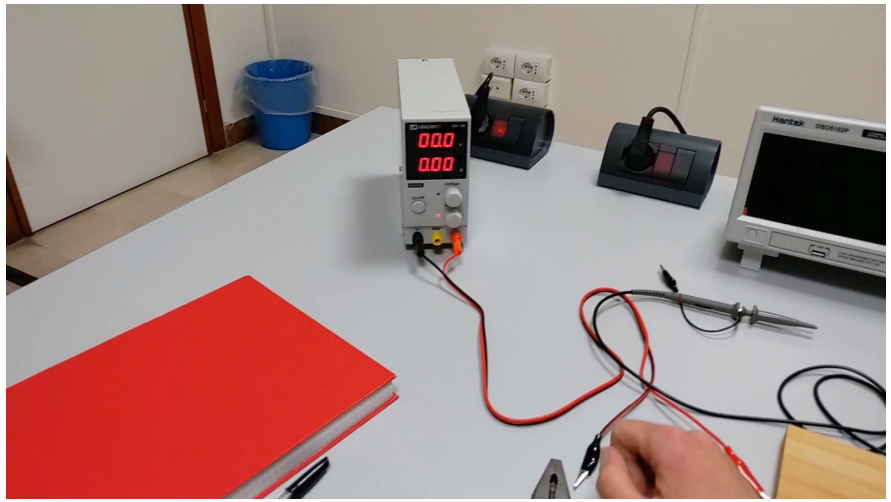
HoloAssist



Assembly101

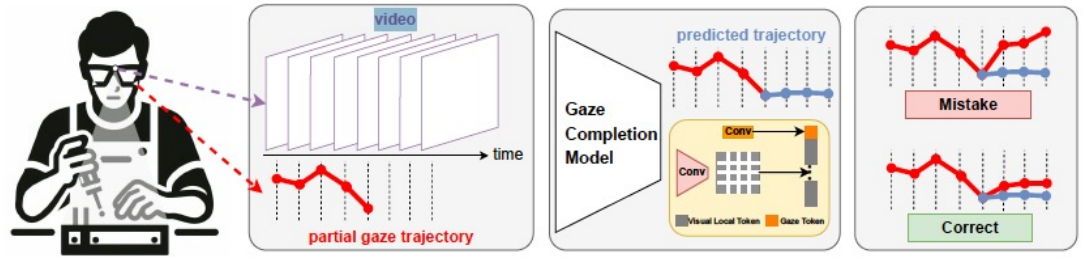
Among Downstream Tasks in Procedure Understanding

1



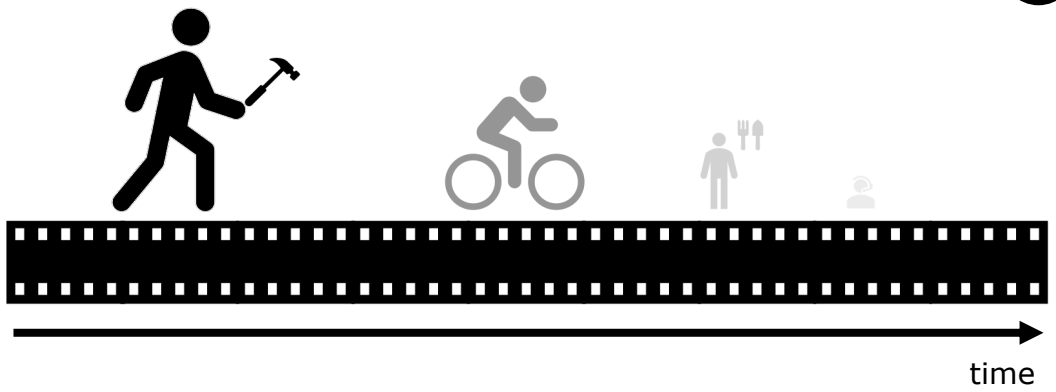
Human-Object Interaction Detection

3



Online Mistake Detection

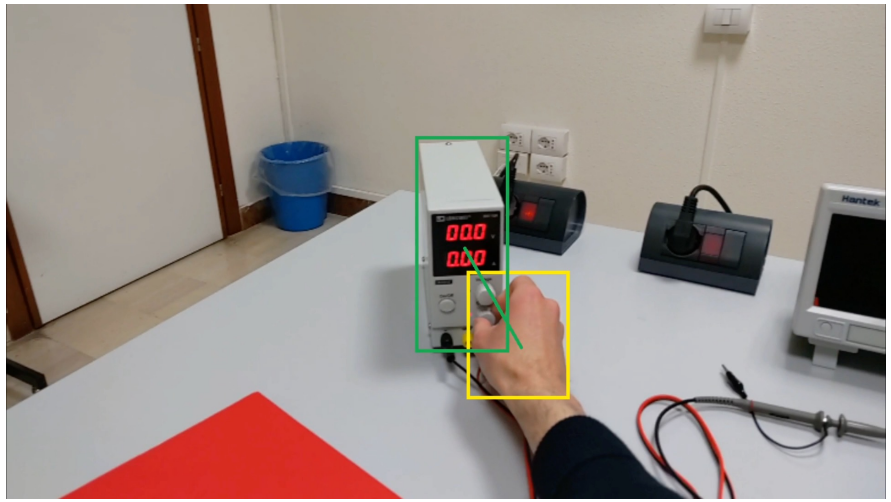
2



LONG-FORM UNDERSTANDING

Among the Tasks in Procedure Understanding

1



Human-Object Interaction Detection

3

Online Mistake Detection

2

LONG-FORM UNDERSTANDING

Università di Catania

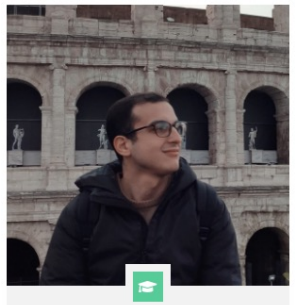
Università di Catania

NEXT VISION

Spin-off of the University of Catania

IMAGE PROCESSING LABORATORY

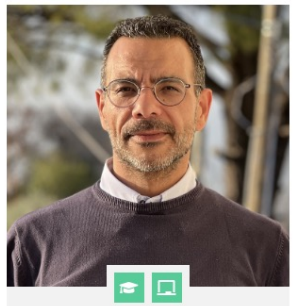
Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?



Rosario Leonardi
FPV@IPLAB
Next Vision s.r.l.

Antonino Furnari
FPV@IPLAB
Next Vision s.r.l.

Francesco Ragusa
FPV@IPLAB
Next Vision s.r.l.



Giovanni Maria Farinella
FPV@IPLAB
Next Vision s.r.l.

ECCV

Poster Session 2
4:30PM TUE 1 OCT

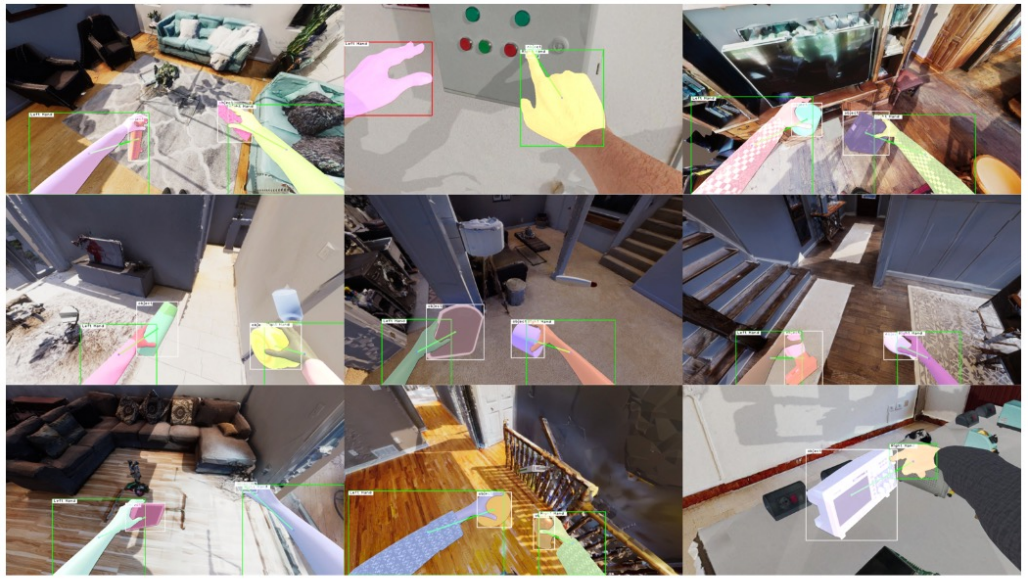
Data and Code: https://fpv-iplab.github.io/HOI-Synth/

Leonardi, R., Furnari, A., Ragusa, F., & Farinella, G. M. (2024). Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection? An Investigation and the HOI-Synth Domain Adaptation Benchmark. European Conference on Computer Vision (ECCV)

Data Generation Pipeline

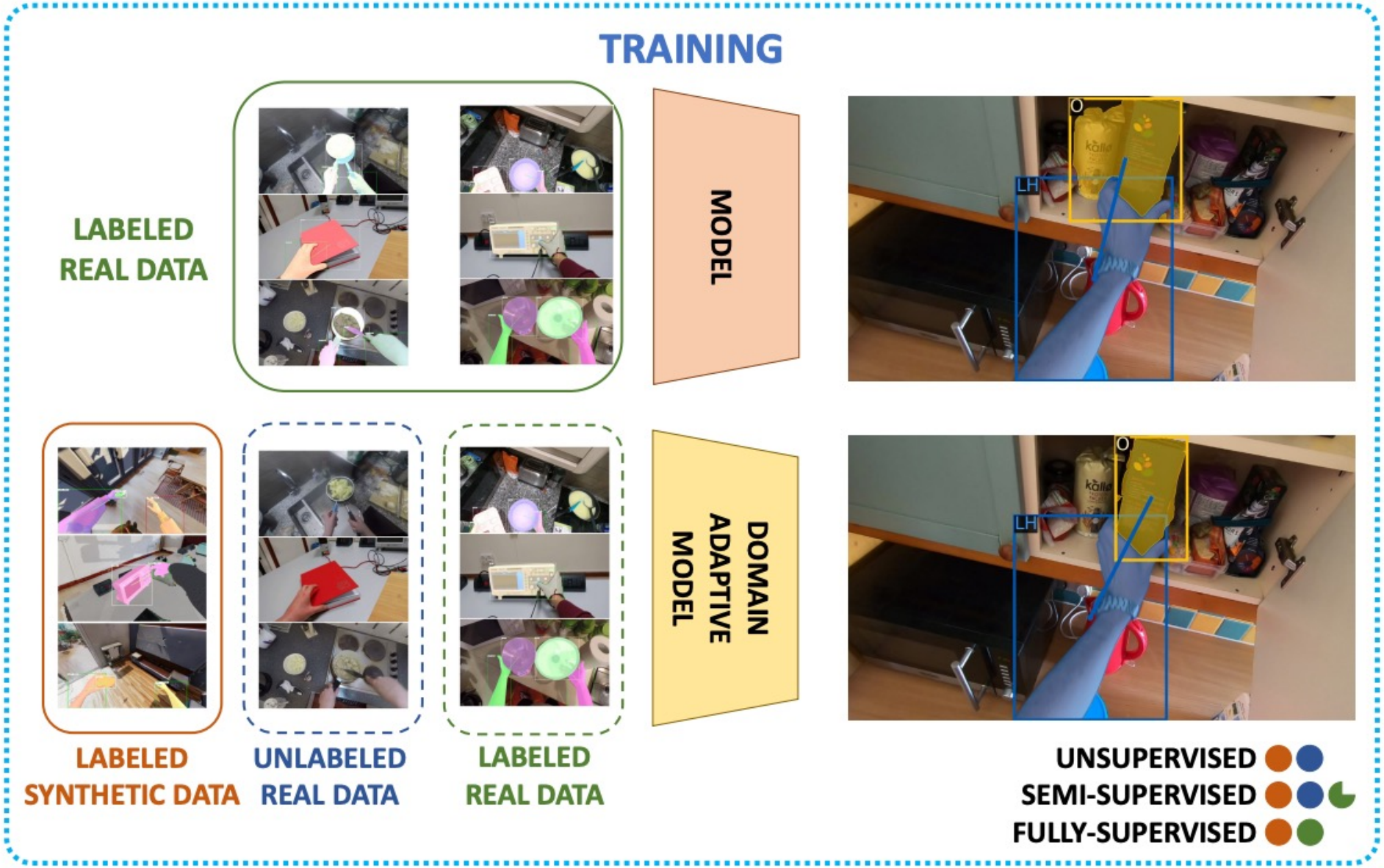
HOI-Synth Dataset – Automatically Labeled

RGB Images	Hand annotations	Object annotations	Interaction frames
75460	141778	101525	101625



Data and Code: https://fpv-iplab.github.io/HOI-Synth/

Can synthetic data help to train models?



Data and Code: https://fpv-iplab.github.io/HOI-Synth/

Results - EPIK-Kitchens VISOR

a) Unsupervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
0%	Synthetic-Only	09.88	28.41	24.89	08.64	01.23
	UDA	**33.33**	**80.16**	**65.98**	**33.47**	**8.35**
Absolute Improvement		**+23.45**	+51.75	+41.09	+24.83	+7.12

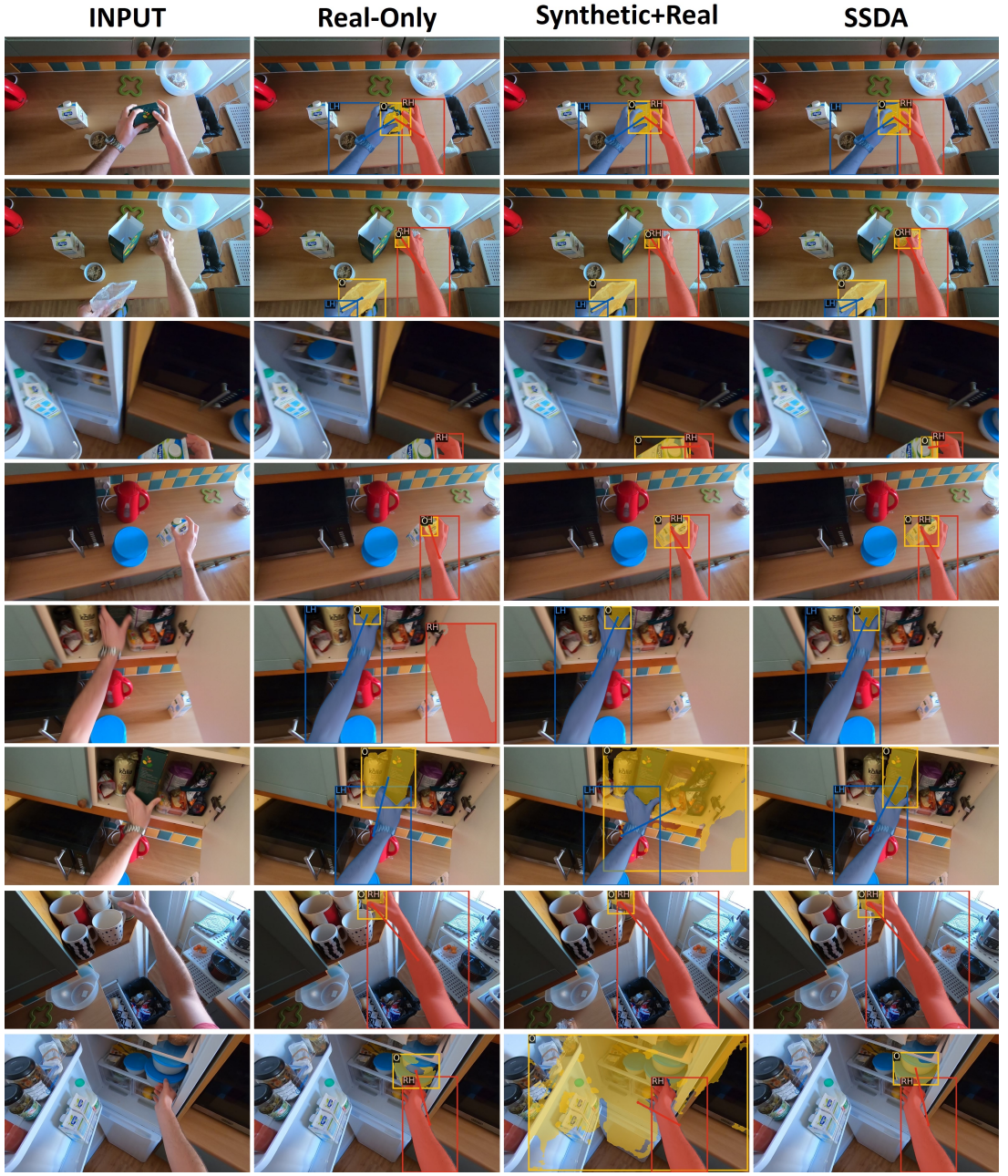
b) Semi-supervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
10% (3,286 images)	Real-Only	38.55	87.45	**83.27**	51.98	19.47
	Synthetic+Real	37.62	86.39	82.85	**52.25**	**23.03**
	SSDA	**44.22**	**89.05**	80.77	46.83	20.41
Absolute Improvement		**+5.67**	+1.60	−0.42	+0.27	+3.56
25% (8,215 images)	Real-Only	37.90	90.14	**85.66**	53.99	17.85
	Synthetic+Real	38.19	89.98	84.67	**55.88**	18.49
	SSDA	**45.55**	**90.37**	84.42	52.59	**22.15**
Absolute Improvement		**+7.65**	+0.23	−0.99	+1.89	+4.30
50% (16,429 images)	Real-Only	38.15	91.16	**86.05**	52.28	17.92
	Synthetic+Real	43.52	**91.34**	85.85	54.09	19.06
	SSDA	**46.47**	90.94	85.73	**58.02**	**23.49**
Absolute Improvement		**+8.32**	+0.18	−0.20	+5.74	+5.57

c) Fully-supervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
100% (32,857 images)	Real-Only	45.33	**92.25**	88.54	**59.24**	24.23
	Synthetic+Real	44.52	91.45	**88.94**	56.55	**27.77**
	FSDA	**46.48**	91.83	87.65	57.63	24.03
Absolute Improvement		**+1.15**	−0.42	+0.40	−1.61	+3.54

Similar behaviour is observed for the EgoHOS and ENIGMA-51 datasets (see the paper)



Among the Tasks in Procedure Understanding

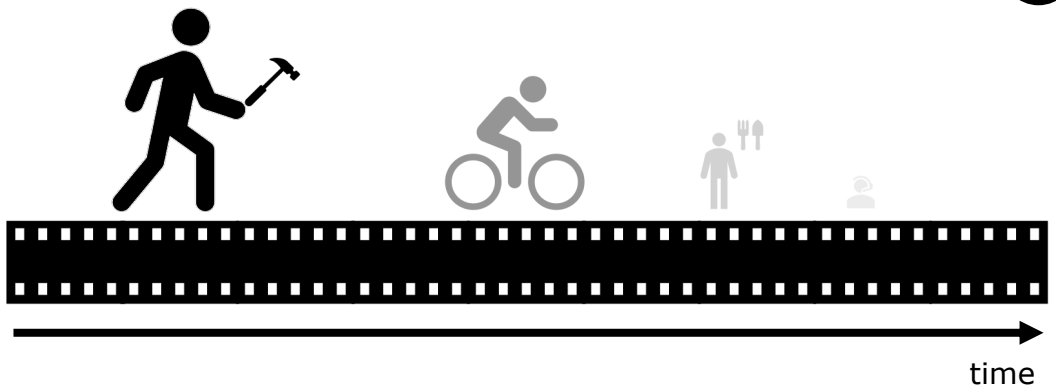
1

Human-Object Interaction Detection

3

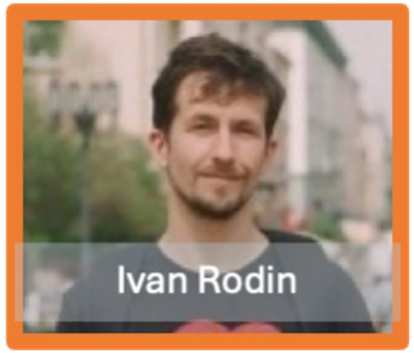
Online Mistake Detection

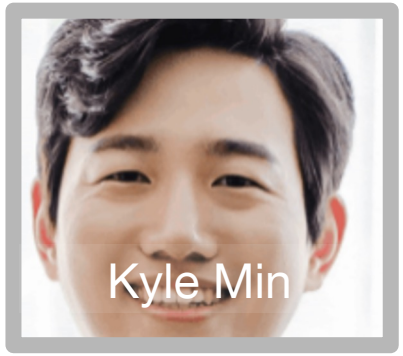
2



LONG-FORM UNDERSTANDING

Egocentric Action Scene Graphs for Long-Form Understanding of Egocentric Videos





CODE AND DATA: https://github.com/fpv-iplab/EASG

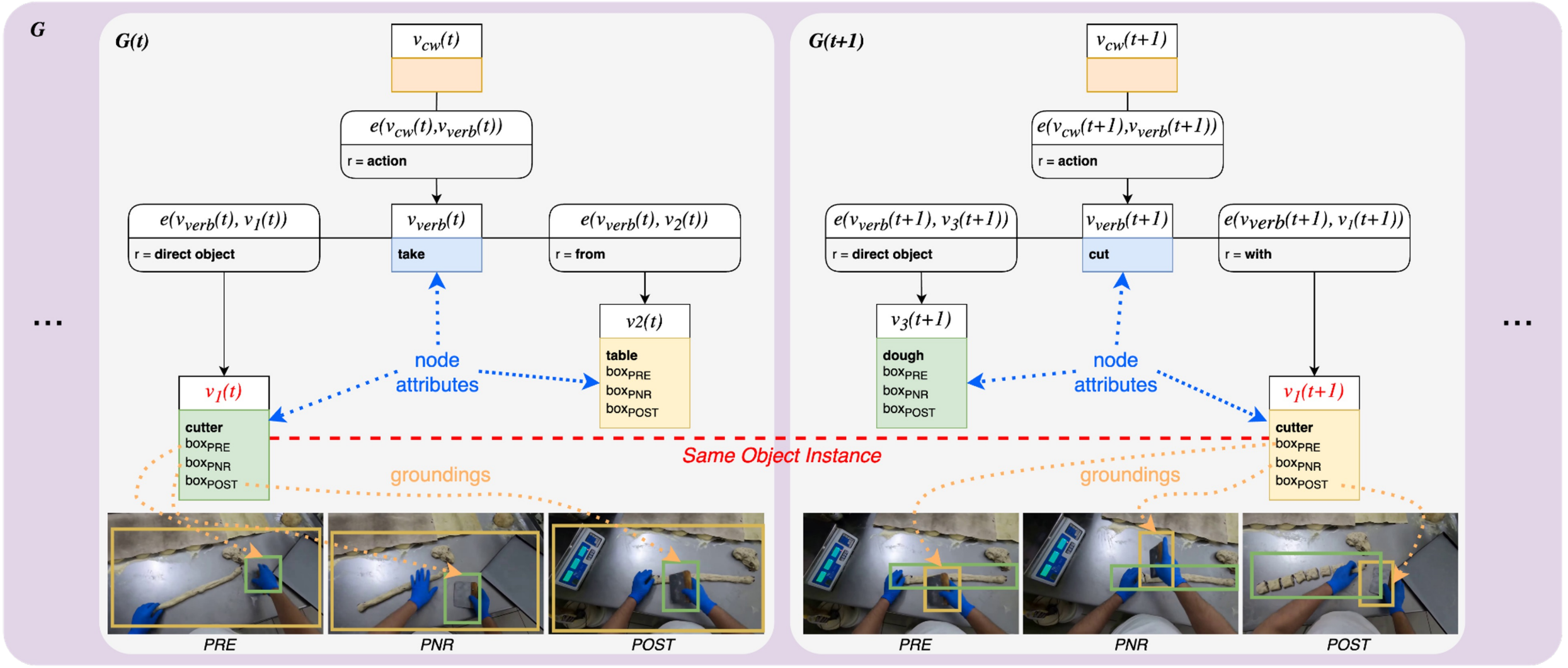
Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, Giovanni Maria Farinella (2024). Action Scene Graphs for Long-Form Understanding of Egocentric Videos. In Conference on Computer Vision and Pattern Recognition (**CVPR 2024**) .

Problem

How to represent egocentric videos for long-term understanding?

- **Understand sequences of activities performed by the camera wearer** in different physical locations
- **Egocentric video is by its own nature long-form**
- Egocentric vision systems require algorithms able to **represent and process video over temporal spans that last in the order of minutes or hours**
- Examples of **applications** are **action anticipation, video summarization, and episodic memory retrieval**
- **Lack of a comprehensive and long-form representation of videos** that algorithms can rely on
- **Popular highlevel human-gathered representations** being in the form of textual narrations, verb-noun action labels, temporal bounds for action segments, object bounding boxes, object state changes, and hand-object interaction states, **are all short-range representations describing temporal spans lasting few seconds.**

Egocentric Action Scene Graphs (EASG)

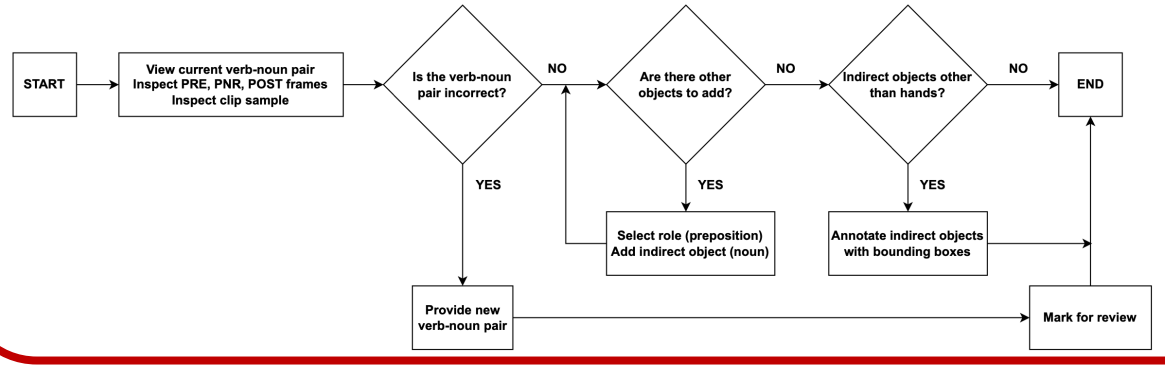


Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, Giovanni Maria Farinella (2024). Action Scene Graphs for Long-Form Understanding of Egocentric Videos. In Conference on Computer Vision and Pattern Recognition (**CVPR 2024**) .

Università di Catania Ego4D-EASG Dataset

Extending Ego4D with Egocentric Action Scene Graph Representations

EASG Annotation Pipeline



Validation procedure	Example Questions (answers in red)
1. Filtering verb-noun pair	Does CW *take bowl* or *press dough*? take bowl
2. Selecting proper preposition in case of multiple edges between two nodes	Select the preposition which is more appropriate: • CW takes bowl **with** left hand ✓ • CW takes bowl **on** left hand
3. Selecting hand(s) if there are different hands with the same preposition	Does CW take bowl with right hand, with left hand or with both hands? left hand
4. Identifying spatial relations	Is the following statement correct: • The bowl is with flour [Y/N] Y • The bowl is from scale [Y/N] N

Example Frame

Dataset	Dynamic	Egocentric	Sequences	Hours	Avg. Len. (seconds)	Avg. Graphs per Vid.	Obj Cls	Verb Cls	Rel Cls
VidVRD [42]	✗	✗	1,000	3	11	3.9*	35	25**	132
VidOR [43]	✗	✗	10,000	99	35	8.8* action + 29.2* spatial	80	42	50
Action Genome [49]	✓	✗	10,000	82	30	5	35	-	25
PVSG [51]	✗	Partly (28%)	400	9	77	382	126	44	57
HOMAGE [38]	✗	paired ego-exo	1,752	25	3	3.8	86	453	29
Ego4D-EASG (Ours)	✓	✓	221	11.4	186	28.3	407	219	16

CODE AND DATA: https://github.com/fpv-iplab/EASG

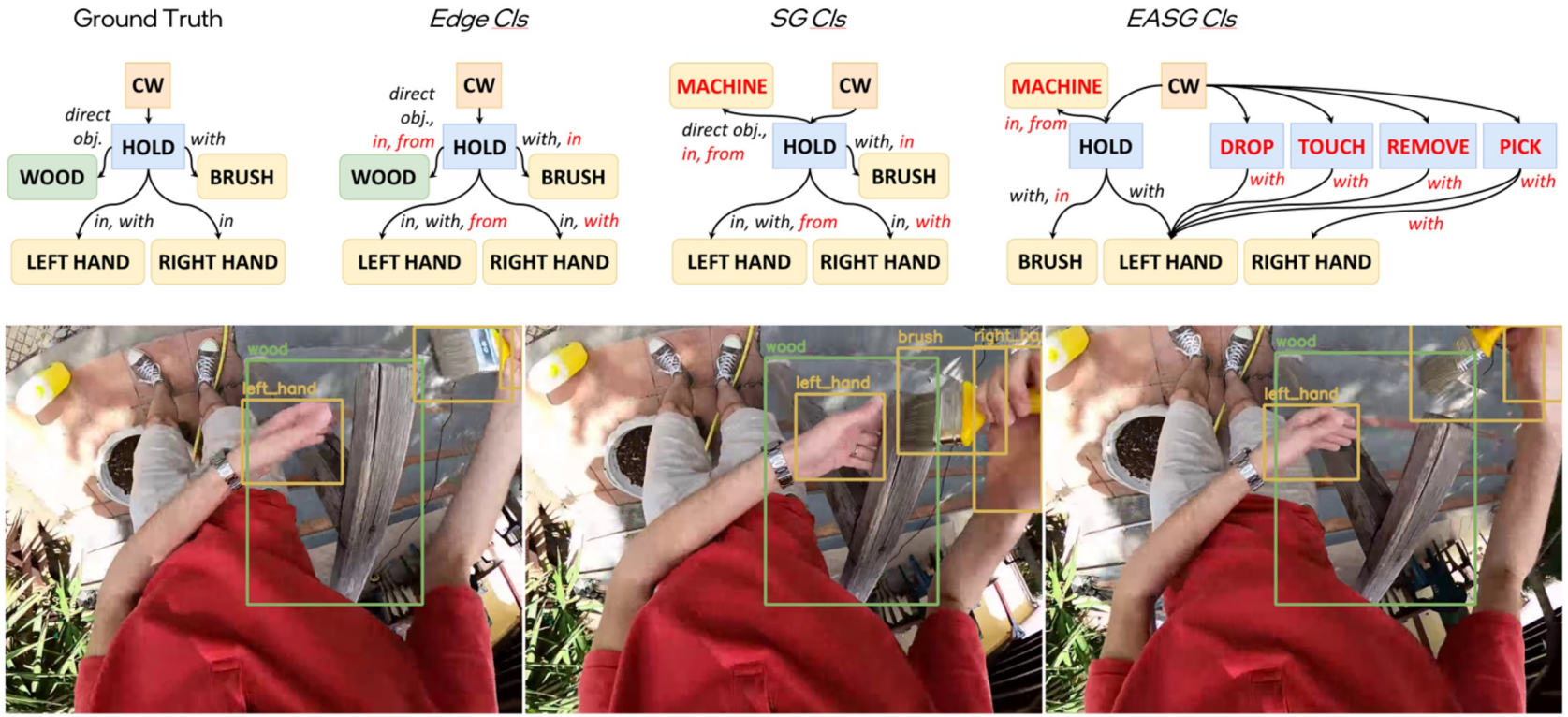
Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, Giovanni Maria Farinella (2024). Action Scene Graphs for Long-Form Understanding of Egocentric Videos. In Conference on Computer Vision and Pattern Recognition (**CVPR 2024**) .

EASG Generation Task

Method	With Constraint									No Constraint								
	Edge Cls			*SG Cls*			*EASG Cls*			*Edge Cls*			*SG Cls*			*EASG Cls*		
	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50	R@10	R@20	R@50
Random Guess	8.0	8.0	8.0	0.2	0.4	1.0	0.0	0.0	0.0	36.5	72.6	99.9	0.3	0.5	1.0	0.0	0.0	0.0
Baseline (Ours)	60.4	60.4	60.4	41.4	44.3	50.6	14.3	16.4	17.9	94.4	99.8	100	51.6	58.2	62.4	14.7	18.3	20.9

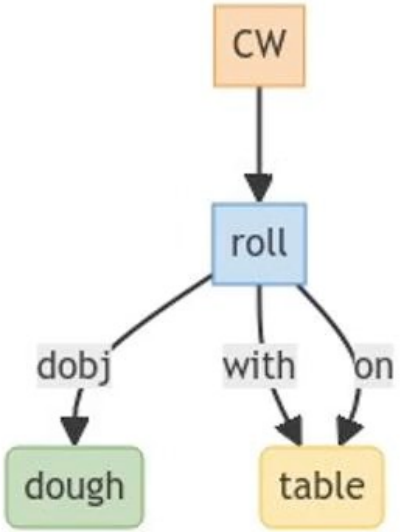
Verb-features: extracted with SlowFast model

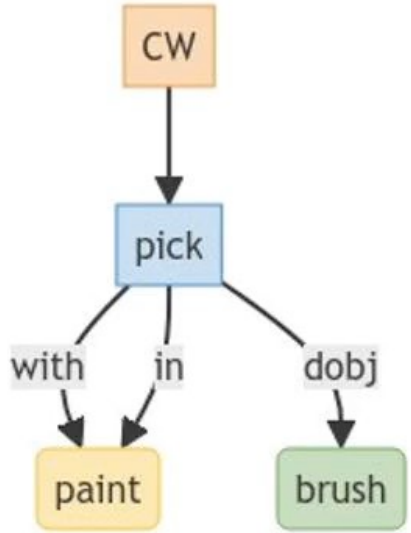
Bbox-features: Faster-RCNN RoIAlign features

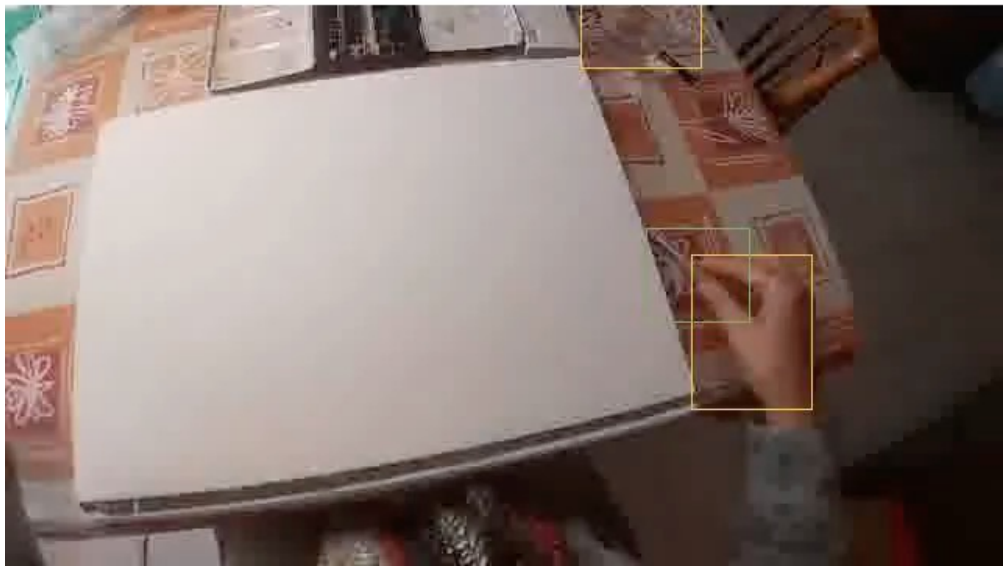


Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, Giovanni Maria Farinella (2024). Action Scene Graphs for Long-Form Understanding of Egocentric Videos. In Conference on Computer Vision and Pattern Recognition (**CVPR 2024**) .

EASG Generation - Qualitative Result Examples







Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, Giovanni Maria Farinella (2024). Action Scene Graphs for Long-Form Understanding of Egocentric Videos. In Conference on Computer Vision and Pattern Recognition (**CVPR 2024**) .

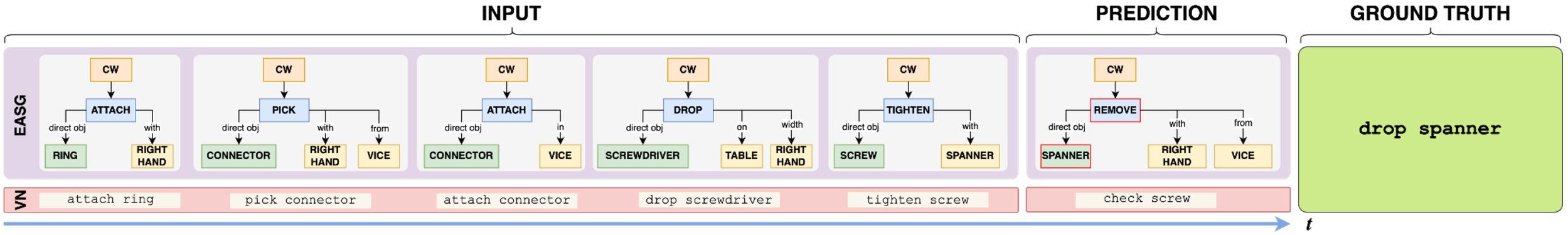
Long-Form Downstream Tasks

Action Anticipation

	Seq. length T	Avg. duration	Verb		Noun		Action	
			Top-1	Top-5	Top-1	Top-5	Top-1	Top-5
V-N	5	19s	2.54	5.01	47.68	62.24	1.28	2.60
EASG	5	19s	3.33	9.53	**48.84**	66.03	1.88	5.24
V-N	20	82s	3.43	8.41	46.69	64.85	2.01	4.98
EASG	20	82s	**5.94**	**15.97**	47.36	**67.26**	**3.40**	**9.24**
Improvement			+2.51	+7.56	+0.67	+2.41	+1.39	+4.26

Summarization

	CIDEr	ROUGE-1	METEOR
V-N	9.42	31.5	26.09
EASG	13.79	33.3	26.30
Narrations	19.99	37.7	29.43



CODE AND DATA: https://github.com/fpv-iplab/EASG

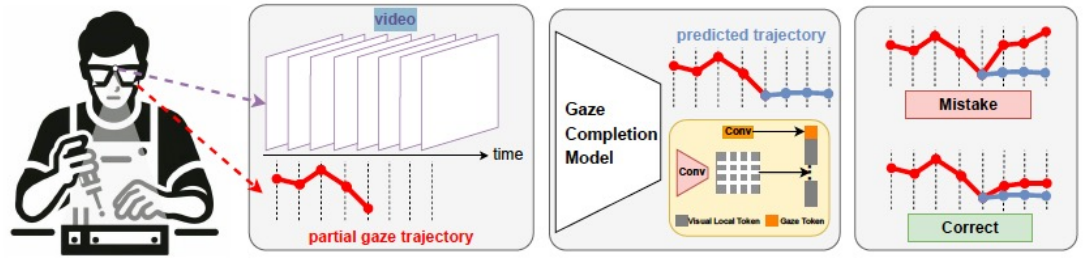
Ivan Rodin, Antonino Furnari, Kyle Min, Subarna Tripathi, Giovanni Maria Farinella (2024). Action Scene Graphs for Long-Form Understanding of Egocentric Videos. In Conference on Computer Vision and Pattern Recognition (**CVPR 2024**) .

Among the Tasks in Procedure Understanding

1

Human Object Interaction Detection

3



Online Mistake Detection

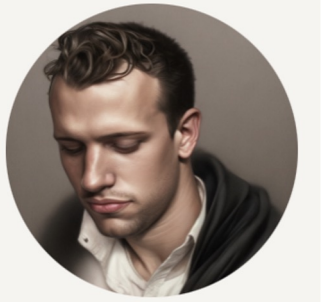
2

LONG-FORM UNDERSTANDING

Università di Catania

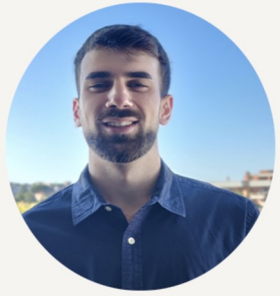
SAPIENZA UNIVERSITÀ DI ROMA

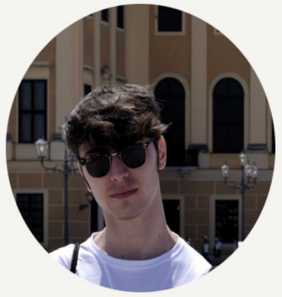
PREGO: online mistake detection in PRocedural EGOcentric videos


Alessandro Flaborea


Guido Maria D'Amely di Melendugno, Ph.D.


Leonardo Plini


Luca Scofano


Edoardo De Matteis


Antonino Furnari


Giovanni M. Farinella


Fabio Galasso

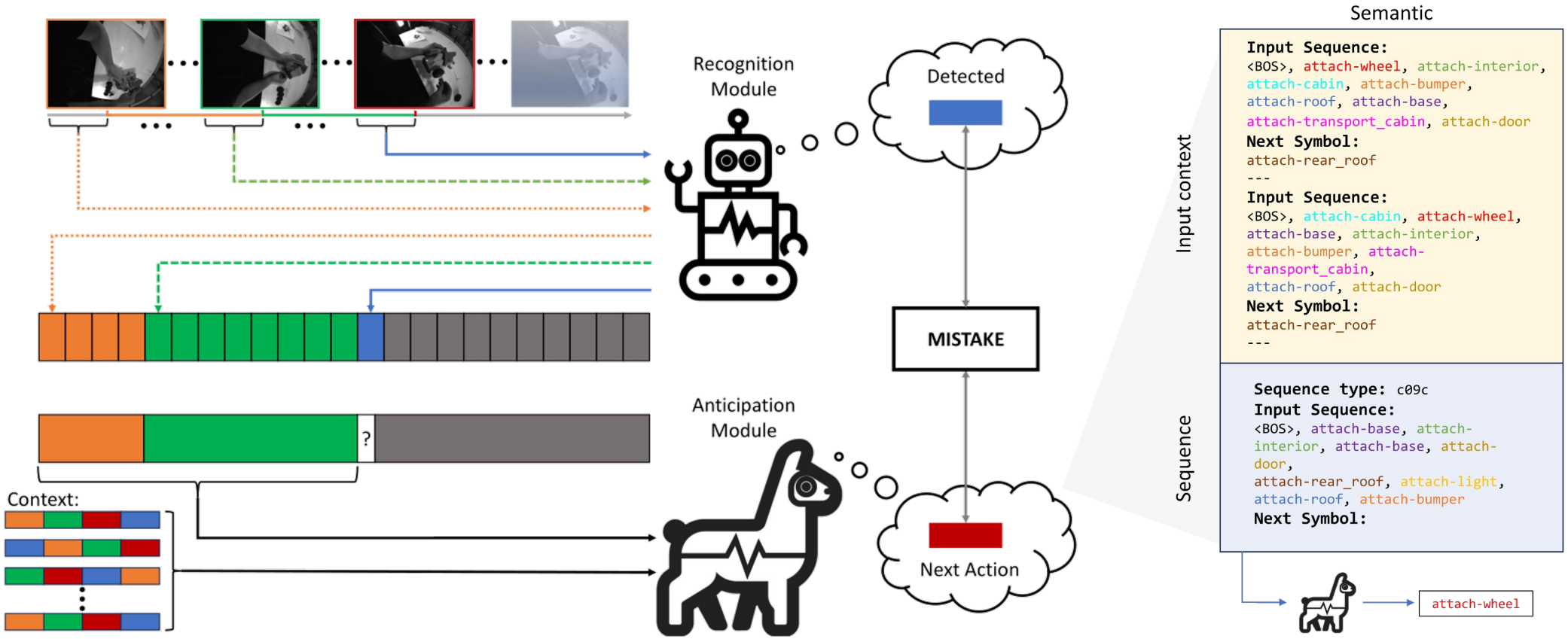
Alessandro Flaborea, Guido D'Amely, Leonardo Plini, Luca Scofano, Edoardo De Matteis, Antonino Furnari, Giovanni Maria Farinella, Fabio Galasso (2024). PREGO: online mistake detection in PRocedural EGOcentric videos. In Conference on Computer Vision and Pattern Recognition (**CVPR 2024**)

Detecting Mistakes in Egocentric Video

- Understand when the user makes a mistake to assist them;
- Current methods rely on **labeled mistakes**, which are hard to obtain;
- We argue that mistake detection should be **one class** and **online**.

Alessandro Flaborea, Guido D'Amely, Leonardo Plini, Luca Scofano, Edoardo De Matteis, Antonino Furnari, Giovanni Maria Farinella, Fabio Galasso (2024). PREGO: online mistake detection in PRocedural EGOcentric videos . In Conference on Computer Vision and Pattern Recognition (**CVPR 2024**)

Detecting Mistakes in Egocentric Video



A model of what "should be"

Alessandro Flaborea, Guido D'Amely, Leonardo Plini, Luca Scofano, Edoardo De Matteis, Antonino Furnari, Giovanni Maria Farinella, Fabio Galasso (2024). PREGO: online mistake detection in PRocedural EGOcentric videos . In Conference on Computer Vision and Pattern Recognition (**CVPR 2024**)

Detecting Mistakes in Egocentric Video

	Step Recog.	Step Antic.	Assembly101-O			Epic-tent-O		
			Precision	Recall	F1 score	Precision	Recall	F1 score
One-step memory	*Oracle*		16.3	30.7	21.3	6.6	26.6	10.6
BERT [7]	*Oracle*		78.2	20.0	31.8	75.0	5.6	10.4
PREGO	*Oracle*	*GPT-3.5*	29.2	75.8	42.1	9.9	73.3	17.4
PREGO	*Oracle*	*LLAMA*	30.7	94.0	46.3	10.7	86.7	19.1
OadTR for MD [35]	*OadTR [35]*	*OadTR [35]*	24.3	18.1	20.7	6.7	21.7	10.2
PREGO	*OadTR [35]*	*LLAMA*	22.1	**94.2**	35.8	9.5	**93.3**	**17.2**
PREGO	*MiniRoad [2]*	*GPT-3.5*	16.2	87.5	27.3	4.3	66.6	8.0
PREGO	*MiniRoad [2]*	*LLAMA*	**27.8**	84.1	**41.8**	**8.6**	20.0	12.0

Alessandro Flaborea, Guido D'Amely, Leonardo Plini, Luca Scofano, Edoardo De Matteis, Antonino Furnari, Giovanni Maria Farinella, Fabio Galasso (2024). PREGO: online mistake detection in PRocedural EGOcentric videos . In Conference on Computer Vision and Pattern Recognition (**CVPR 2024**)

Differentiable Task Graph Learning: Procedural Activity Representation and Online Mistake Detection from Egocentric Videos

Luigi Seminara

Prof. Giovanni Maria Farinella

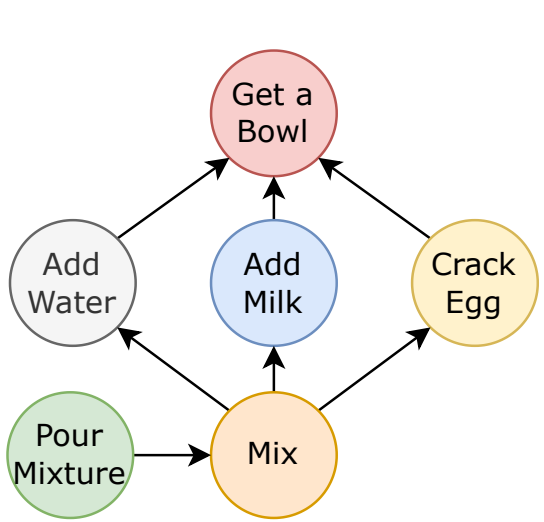
Antonino Furnari

CODE IS AVAILABLE:
https://github.com/fpv-iplab/Differentiable-Task-Graph-Learning

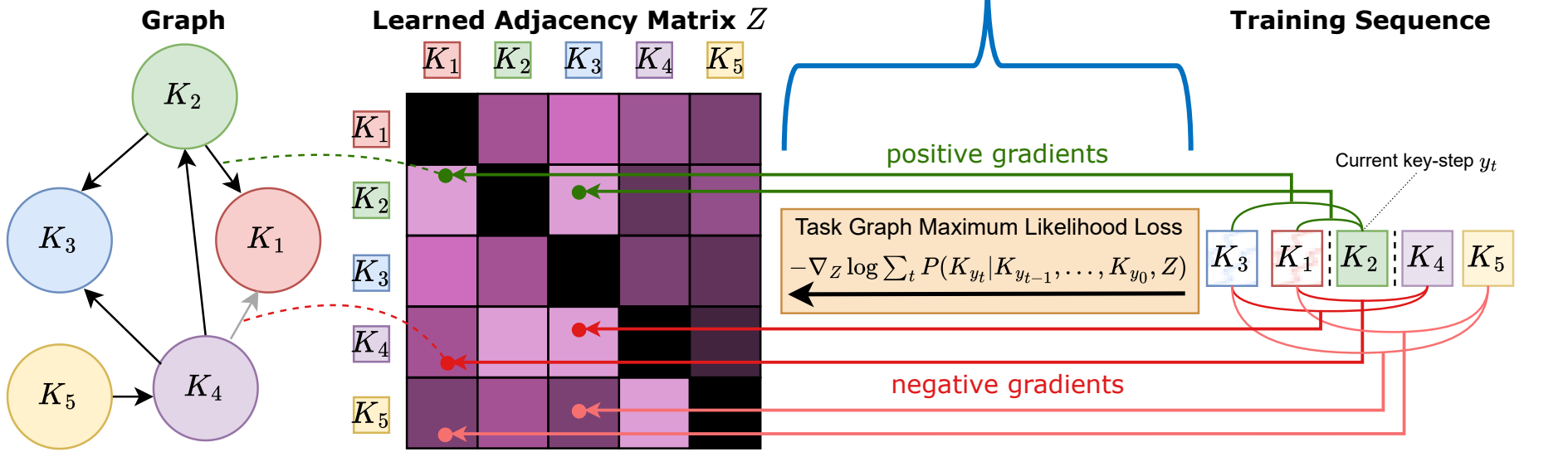
Luigi Seminara, Giovanni Maria Farinella, Antonino Furnari (2024). Differentiable Task Graph Learning: Procedural Activity Representation and Online Mistake Detection from Egocentric Videos. ArXiv: https://arxiv.org/abs/2406.01486

Task Graph Learning

We propose the Task Graph Maximum Likelihood (TGML) loss directly supervises the entries of the adjacency matrix Z generating gradients to maximize the probability of edges from past nodes (K3, K1) to the current node (K2), while minimizing the probability of edges from past nodes to future nodes (K4, K5) in a contrastive manner.



(a) Example Task Graph



(b) Task Graph Learning as Maximum Likelihood Estimation

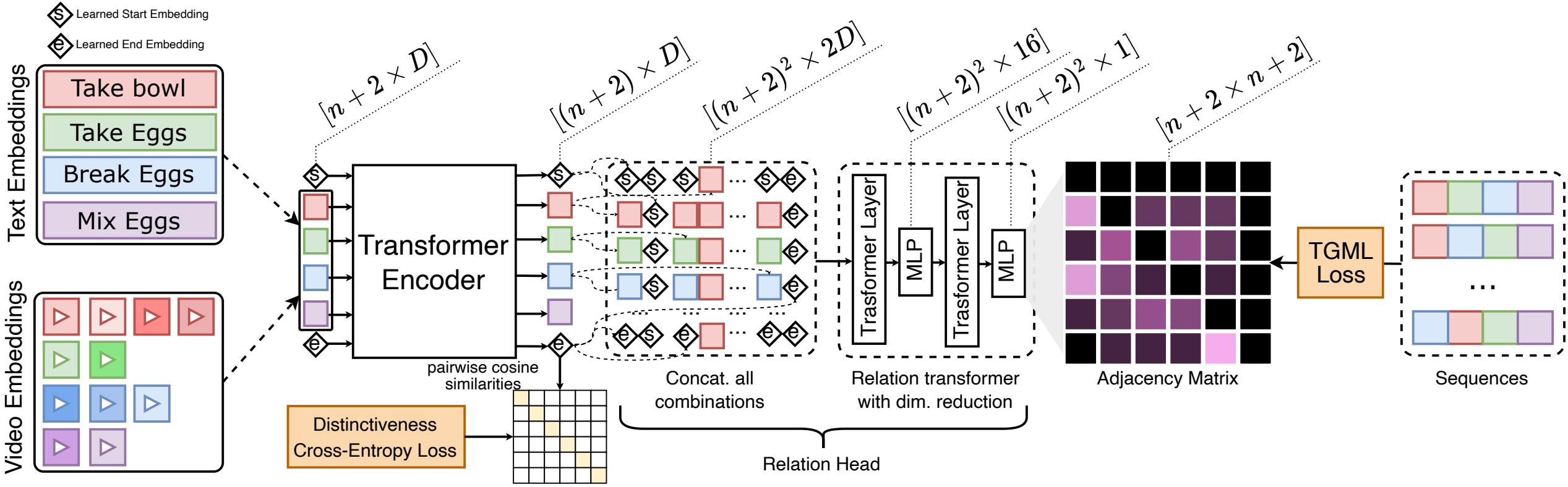
An example task graph encoding dependencies in a "mix eggs" procedure

We learn a task graph which encodes a partial ordering between actions represented as an adjacency matrix Z

Task Graphs are learned from input action sequences

Task Graph Transformer

CODE IS AVAILABLE:
https://github.com/fpv-iplab/Differentiable-Task-Graph-Learning



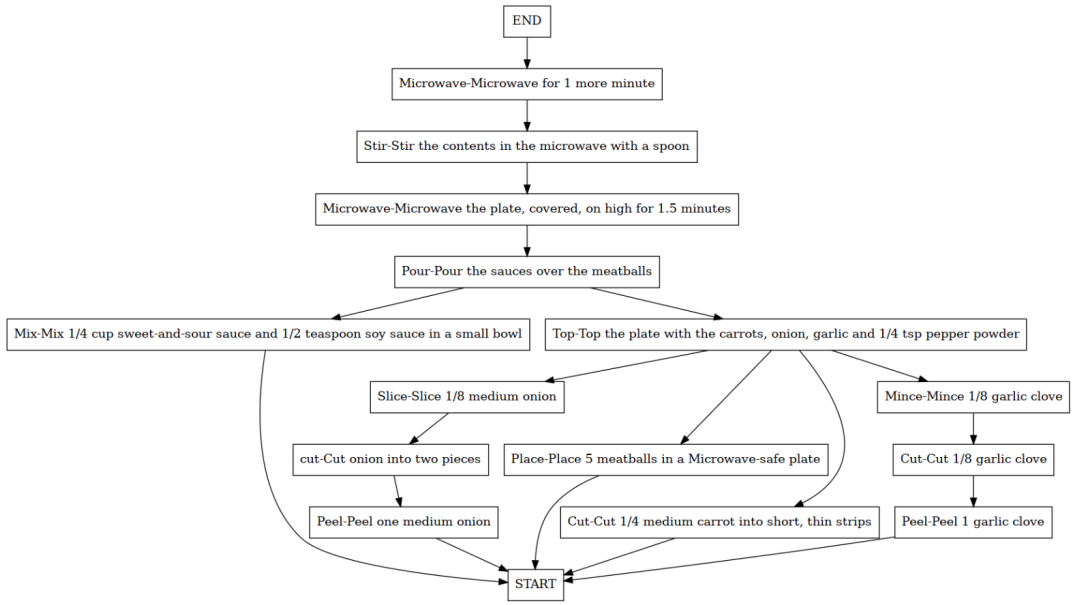
See paper for the details

Luigi Seminara, Giovanni Maria Farinella, Antonino Furnari (2024). Differentiable Task Graph Learning: Procedural Activity Representation and Online Mistake Detection from Egocentric Videos. ArXiv: https://arxiv.org/abs/2406.01486

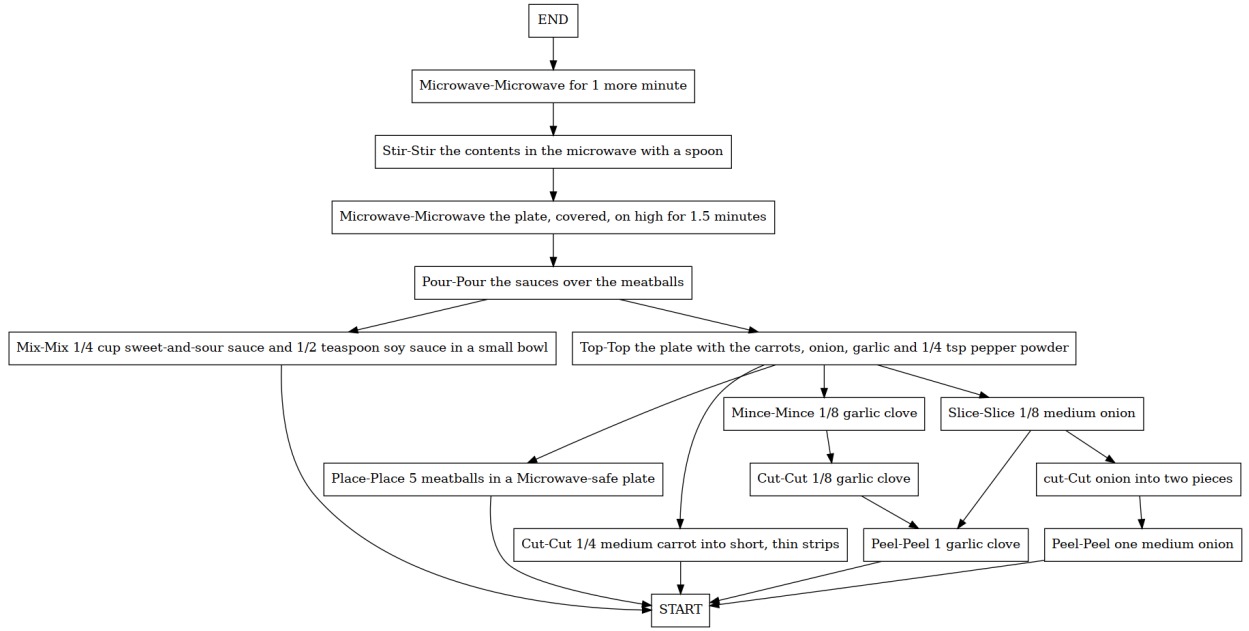
Qualitative Results

CODE IS AVAILABLE:
https://github.com/fpv-iplab/Differentiable-Task-Graph-Learning

Ground Truth



Generated



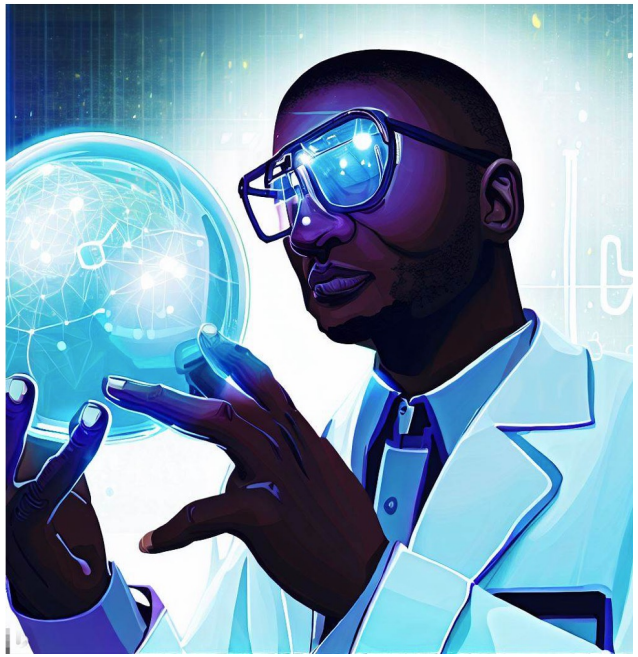
Luigi Seminara, Giovanni Maria Farinella, Antonino Furnari (2024). Differentiable Task Graph Learning: Procedural Activity Representation and Online Mistake Detection from Egocentric Videos. ArXiv: https://arxiv.org/abs/2406.01486

Online Mistake Detection with Generated Task Graphs

Table 3: Online mistake detection results. Results obtained with ground truth action sequences are denoted with *, while results obtained on predicted action sequences are denoted with $^+$.

	Assembly101							EPIC-Tent						
	Avg	Correct			Mistake			Avg	Correct			Mistake		
Method	F_1	F_1	Prec	Rec	F_1	Prec	Rec	F_1	F_1	Prec	Rec	F_1	Prec	Rec
Count-Based* [3]	26.2	9.5	5.1	86.2	42.9	97.8	27.5	56.6	92.5	92.8	92.2	20.7	20.0	21.4
LLM*	29.3	15.1	8.3	87.2	43.4	96.7	27.9	47.7	86.3	82.4	90.6	9.1	13.3	6.9
MSGI* [35]	33.1	22.7	13.1	84.4	43.5	93.4	28.3	44.5	66.9	51.6	95.2	22.0	73.3	12.9
PREGO* [12]	39.4	32.6	89.7	19.9	46.3	30.7	94.0	32.1	45.0	95.7	29.4	19.1	10.7	86.7
MSG^{2}* [18]	56.1	63.9	51.5	84.2	48.2	73.6	35.8	54.1	92.9	94.1	91.7	15.4	13.3	18.2
TGT-text (Ours)*	62.8	69.8	56.8	90.6	55.7	84.1	41.7	**64.1**	**93.8**	94.1	93.5	**34.5**	33.3	35.7
DO (Ours)*	**75.9**	**90.2**	98.2	83.4	**61.6**	46.7	90.4	58.3	93.5	94.8	92.4	23.1	20.0	27.3
Improvement*	+19.8	+26.3			+13.4			+7.5	+0.9			+12.5		
Count-Based$^+$ [3]	23.1	2.5	1.3	60.0	**43.7**	97.8	28.1	38.1	68.3	54.9	90.1	7.9	26.7	4.7
LLM$^+$	28.1	15.1	7.8	65.5	42.3	89.5	27.7	35.9	61.6	46.7	90.4	10.2	40.0	5.8
MSGI$^+$ [35]	28.4	14.0	7.8	67.9	42.7	90.7	28.0	40.4	59.2	42.9	95.5	21.6	80.0	12.5
PREGO$^+$ [12]	32.5	23.1	68.8	13.9	41.8	27.8	84.1	29.4	41.6	97.9	26.4	17.2	9.5	93.3
MSG^{2+} [18]	46.2	59.1	51.2	70.0	33.2	44.5	26.5	45.2	67.5	52.4	95.1	22.9	73.3	13.6
TGT-text (Ours)$^+$	53.0	67.8	62.3	74.5	38.2	46.2	32.6	43.8	**69.5**	55.8	92.1	18.2	53.3	11.0
DO (Ours)$^+$	**53.5**	**78.9**	85.0	73.5	28.1	22.5	37.3	**46.5**	69.3	54.4	95.2	**23.7**	73.3	14.1
Improvement$^+$	+7.3	+19.8			-5.5			+1.3	+1.2			+1.2		

Luigi Seminara, Giovanni Maria Farinella, Antonino Furnari (2024). Differentiable Task Graph Learning: Procedural Activity Representation and Online Mistake Detection from Egocentric Videos. ArXiv: https://arxiv.org/abs/2406.01486

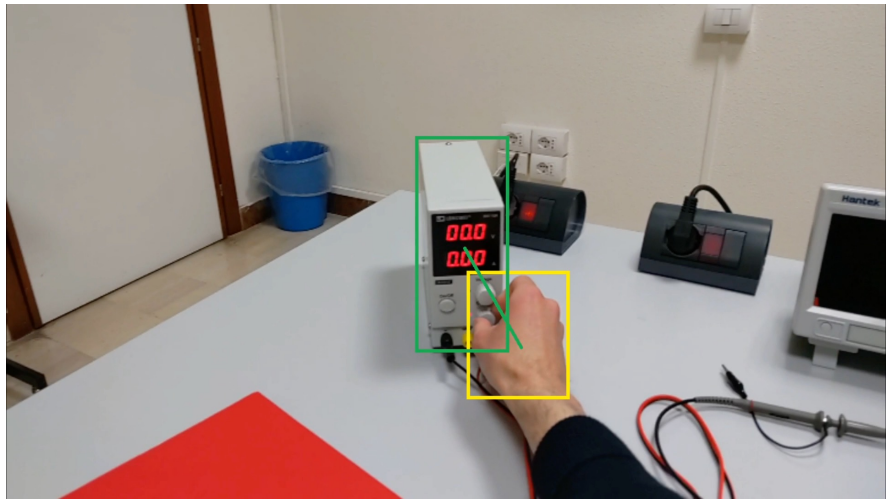


Take home messages

Procedure Understanding opens **many research challenges** as well as the **new and useful applications in different domains to support humans where they live and work** (e.g. Ego-Worker, Ego-Home, ...);

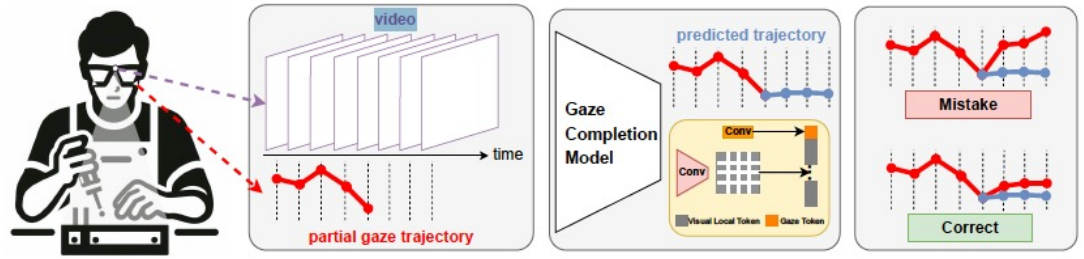
Procedure Understanding from Egocentric Videos

1



Human Object Interaction Detection

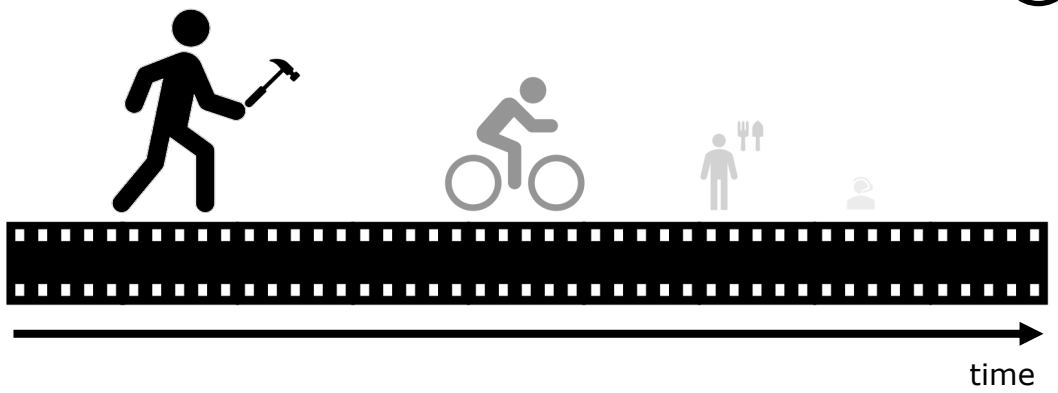
3



Online Mistake Detection

The research community has collected **huge amount of data** and has defined **fundamental research tasks and benchmarks**;

2



LONG-FORM UNDERSTANDING

HOI, Long-form Understanding (Anticipation, Summarization), and Mistake Detection are **key tasks in this context**;

Procedure Understanding from Egocentric Videos

Università di Catania

NEXT VISION
Spin-off of the University of Catania

Image Processing Laboratory

ECCV
Poster Session 1
10:30AM TUE 1 OCT

Synchronization is All You Need: Exocentric-to-Egocentric Transfer for Temporal Action Segmentation with Unlabeled Synchronized Video Pairs


Camillo Quattrocchi


Antonino Furnari


Daniele Di Mauro


Valerio Giuffrida


Giovanni Maria Farinella

C. Quattrocchi, A. Furnari, D. Di Mauro, M. V. Giuffrida, G. M. Farinella (2024). Synchronization is All You Need: Exocentric-to-Egocentric Transfer for Temporal Action Segmentation with Unlabeled Synchronized Video Pairs. European Conference on Computer Vision (ECCV)

Temporal Action Segmentation

Università di Catania

Image Processing Laboratory

Unsupervised Mistake Detection in Egocentric Procedural Video by Detecting Unpredictable Gaze


Michele Mazzamuto


Antonino Furnari


Giovanni Maria Farinella

Michele Mazzamuto, Antonino Furnari, and Giovanni Maria Farinella (2024). Eyes Wide Unshut: Unsupervised Mistake Detection in Egocentric Procedural Video by Detecting Unpredictable Gaze. ArXiv: https://arxiv.org/abs/2406.08379

Gaze-Based Mistake Detection

arXiv

Università di Catania

Universidad Zaragoza

ECCV
Poster Session 4
4:30PM WED 2 OCT

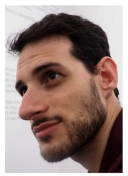
AFF-ttention! Affordances and Attention models for Short-Term Object Interaction Anticipation


Lorenzo Mur-Labadia


Ruben Martinez-Cantin


José Guerrero-Campo


Giovanni Maria Farinella


Antonino Furnari

Lorenzo Mur-Labadia, Ruben Martinez-Cantin, Josechu Guerrero, Giovanni Maria Farinella, Antonino Furnari (2024). AFF-ttention! Affordances and Attention models for Short-Term Object Interaction Anticipation. In European Conference on Computer Vision (ECCV)

Object-Interaction Anticipation

FAIR

Procedure Understanding from Egocentric Videos

Despite promising attempts there are **still a lot of progress to be done** to solve challenges in this research context. **Join the effort!**

Finanziato dall'Unione europea NextGenerationEU
Ministero dell'Università e della Ricerca
Italiadomani PIANO NAZIONALE DI RIPRESA E RESILIENZA
FAIR Future Artificial Intelligence Research

Università di Catania

NEXT VISION

Spin-off of the University of Catania

IMAGE PROCESSING LABORATORY

Thank you for your attention

Procedure Understanding from Egocentric Videos

Giovanni Maria Farinella

FPV@Image Processing Laboratory - http://iplab.dmi.unict.it/fpv

Next Vision - http://www.nextvisionlab.it/

Department of Mathematics and Computer Science - University of Catania

giovanni.farinella@unict.it - www.dmi.unict.it/farinella