First Person (Egocentric) Vision for Human-Centric Assistance: History, Building Blocks, and Applications

Antonino Furnari, Francesco Ragusa

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

Department of Mathematics and Computer Science - University of Catania

Next Vision s.r.l., Italy

furnari@dmi.unict.it - http://www.antoninofurnari.it/

francesco.ragusa@unict.it - https://iplab.dmi.unict.it/ragusa/

http://iplab.dmi.unict.it/fpv - https://www.nextvisionlab.it/

Before we begin…

The slides of this tutorial are available online at:

http://www.antoninofurnari.it/talks/iciap2022

Agenda

1) Part I: Definitions, motivations, history and research trends [14.00 - 15.30] – Antonino Furnari
   a) What is first person vision? What is it for?
   b) What makes it different from third person vision?
   c) History of First Person Vision: visions, ideas, research, devices;
   d) Where do we go from here? Research trends, datasets and challenges.

Coffe Break [15.30 – 16.00]

1) **Part II: Building Blocks for First Person Vision Systems [16.00 – 18.00] – Francesco Ragusa**
   a) **Data Acquisition & Datasets;**
   b) **Fundamental Task in First Person Vision:**
      i) **Localization;**
      ii) **Object Detection and Recognition;**
      iii) **Egocentric Human-Object Interaction;**
      iv) **Action/Activities;**
      v) **Anticipation.**
   c) **Example Applications;**
   d) **Conclusion.**

Part 2

Building Blocks for First Person Vision Systems

Data Acquisition – Video Quality

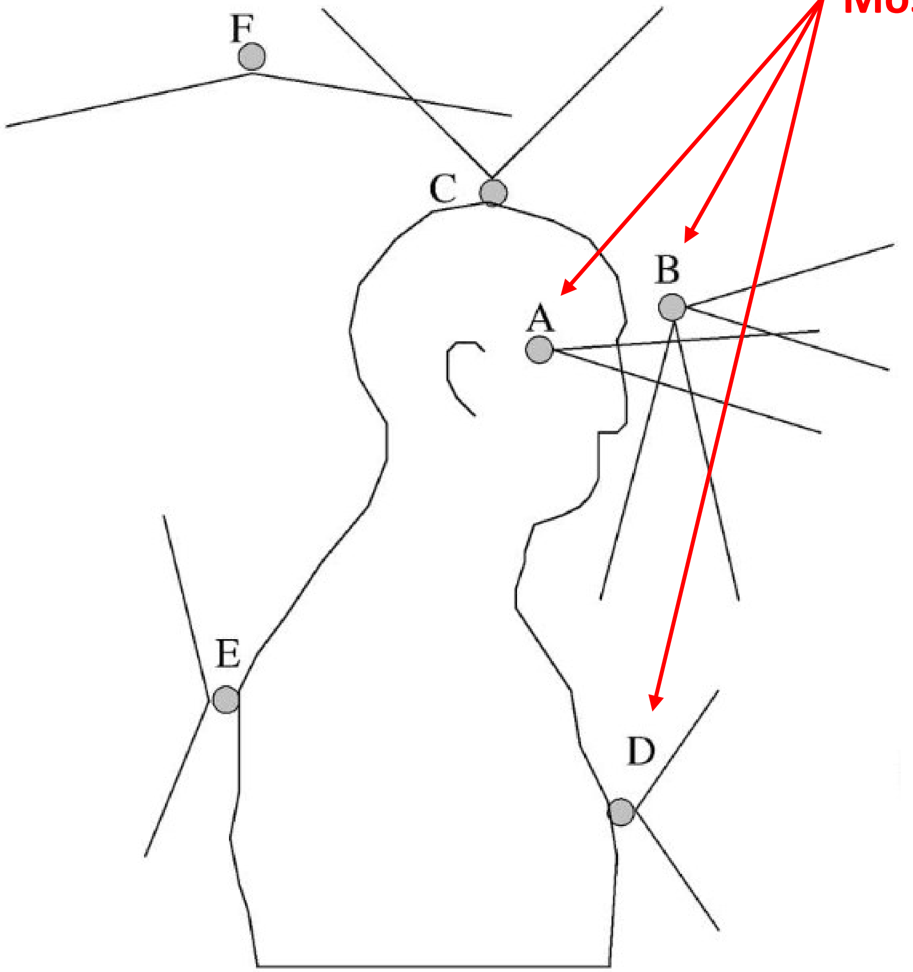
- Try to get a high quality camera to get high quality images!
- Egocentric video is subject to motion blur and exposure issues.

High Quality Video Obtained with a GoPro

Average Quality Video

Data Acquisition – Camera Wearing Modalities

Most Common Wearing Modalities **A,B: head mounted, D: chest mounted**



A

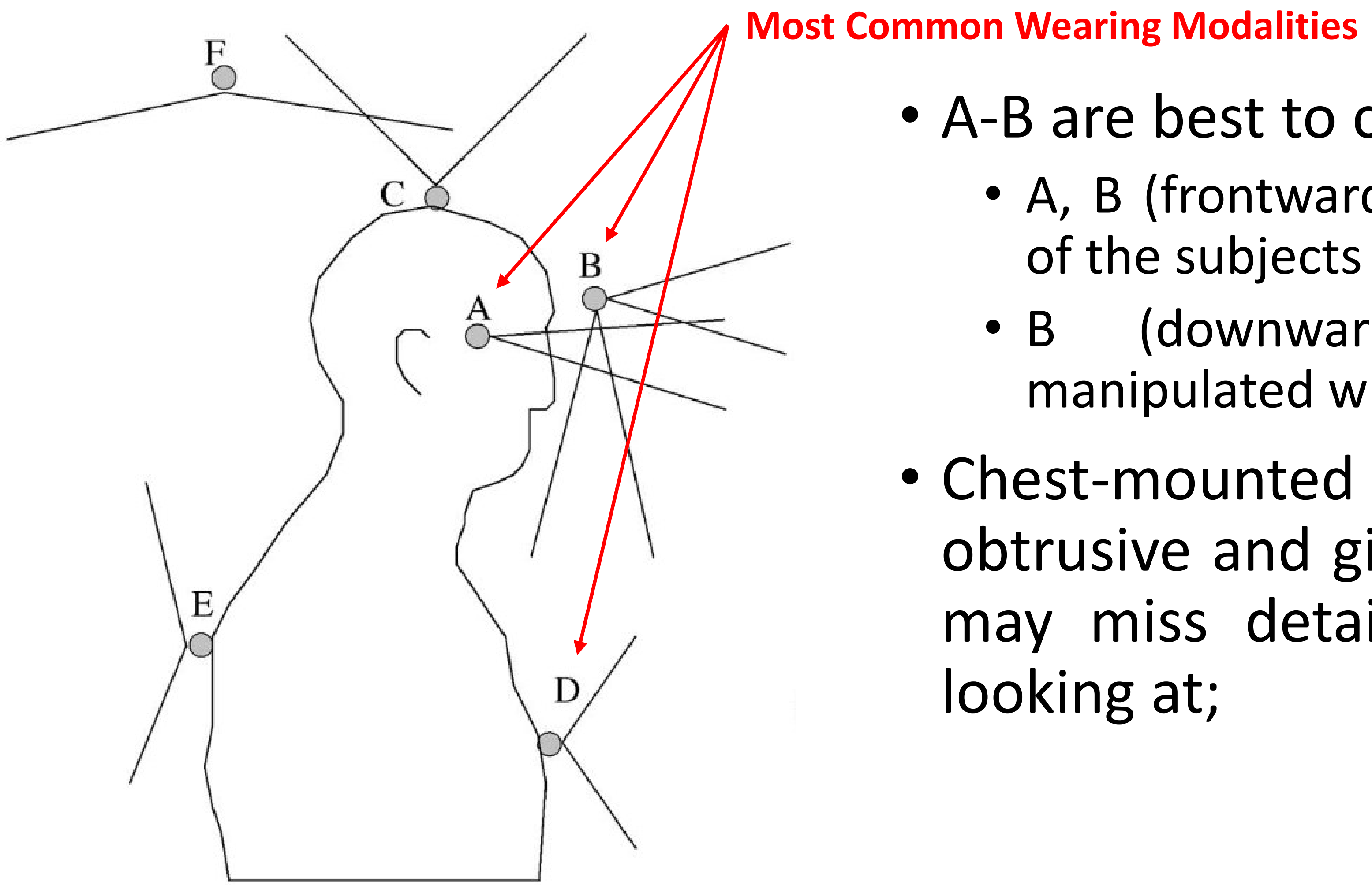
B (frontward)

B (downward)

D

Mayol-Cuevas, W. W., Tordoff, B. J., & Murray, D. W. (2009). On the choice and placement of wearable vision sensors. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *39*(2), 414-425.

Data Acquisition – Camera Wearing Modalities (2)



- A-B are best to capture objects:
  - A, B (frontward) to capture objects in front of the subjects (e.g., paintings in a museum);
  - B (downward) to capture objects manipulated with hands (e.g., kitchen);
- Chest-mounted cameras (D) are less obtrusive and give stable video, but they may miss details on what the user is looking at;

Mayol-Cuevas, W. W., Tordoff, B. J., & Murray, D. W. (2009). On the choice and placement of wearable vision sensors. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *39*(2), 414-425.

Data Acquisition – Field of View (FOV)

A wide FOV allows to capture more scene but introduces distortion.

Narrow Angle

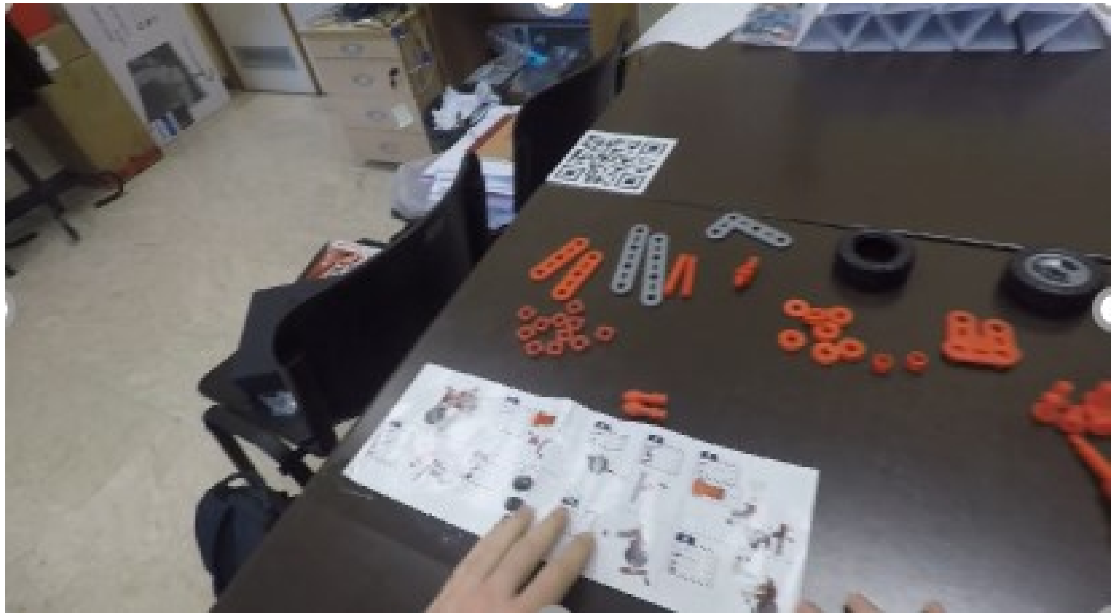
Wide Angle

Data Acquisition – Field of View (FOV)

A wide FOV allows to capture more scene but introduces distortion.

Narrow Angle

Wide Angle



Data Acquisition – Other Modalities – Depth

- If you can acquire depth, do it!
- Depth can improve scene understanding by highlighting the position of objects and hands;

Wan, S., & Aggarwal, J. K. (2015). Mining discriminative states of hands and objects to recognize egocentric actions with a wearable RGBD camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 36-43).

Data Acquisition – Other Modalities – Gaze

Gaze can give information on what the user is paying attention to.

However, gaze trackers generally require a calibration process (and some expertise).

F. Ragusa, A. Furnari, S. Livatino, G. M. Farinella. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. WACV 2021 (ORAL) (https://arxiv.org/abs/2010.05654).

Datasets

- If you are trying to solve a specific FPV problem, chances are that someone already collected/labeled data that is suitable for you.
- Search on the internet first!
- In particular, there are quite a few dataset focusing on action/activity recognition;
- In the following, a (non-exhaustive) list of datasets.

Datasets (non-exhaustive)

Dataset	URL	Settings	Annotations	Goal
EGO4D	https://ego4d-data.org/	931 participants performing different activities in different domains.	Different temporal and spatial annotations related to 5 benchmarks	Episodic Memory, Hand-Object Interaction, Audio-Visual Diarization, Social Interactions, Forecasting
EPIC-KITCHENS-100	https://epic-kitchens.github.io/2020-100	Subjects performing unscripted actions in their native kitchens.	Temporal segments	Action recognition, detection, anticipation, retrieval.
MECCANO	https://iplab.dmi.unict.it/MECCANO/	20 subjects assembling a toy motorbike.	Temporal segments, active objects, human-object interactions	Action recognition, Active object detection, Egocentric Human-Object Interaction Detection
ASSEMBLY101	https://assembly-101.github.io/	53 subjects assembling in a cage settings 101 children's toys.	Temporal segments, 3D hand poses	Action recognition, Action Anticipation, Temporal Segmentation

Datasets (non-exhaustive)

Dataset	URL	Settings	Annotations	Goal
EPIC-KITCHENS 2018	https://epic-kitchens.github.io/2018	32 subjects performing unscripted actions in their native environments	action segments, object annotations	Action recognition, Action Anticipation, Object Detection
Charade-Ego	https://allenai.org/plato/charades/	paired first-third person videos	action classes	Action recognition
EGTEA Gaze+	http://ai.stanford.edu/~alireza/GTEA/	32 subjects, 86 sessions, 28 hours	action segments, gaze, hand masks	Understading daily activities, action recognition
ADL	https://www.csee.umbc.edu/~hpirsiav/papers/ADLdataset/	20 subjects performing daily activities in their native environments	activity segments, objects	Detecting activities of daily living
CMU kitchen	http://www.cs.cmu.edu/~espriggs/cmu-mmac/annotations/	multimodal, 18 subjects cooking 5 different recipes: brownies, eggs, pizza, salad, sandwiche	action segments	Understading daily activities
EgoSeg	http://www.vision.huji.ac.il/egoseg/	Long term actions (walking, running, driving, etc.)	long term activity	Temporal Segmentation, Indexing

Datasets (non-exhaustive)

Dataset	URL	Settings	Annotations	Goal
First-Person Social Interactions	http://ai.stanford.edu/~alireza/Disney/	8 subjects at disneyworld	Activities: walking, waiting, gathering, sitting, buying something, eating, etc.	Recognizing social interactions
UEC Dataset	http://www.cs.cmu.edu/~kkitani/datasets/	two choreographed datasets with different egoactions (walk, jump, climb, etc.) + 6 youtube sports videos	activities	Unsupervised activity recognition
JPL	http://michaelryoo.com/jpl-interaction.html	interaction with a robot	activities performed on the robot + pose	Interaction recognition/prediction
Multimodal Egocentric Activity Dataset	http://people.sutd.edu.sg/~1000892/dataset	15 seconds clips of 20 activities	activity (walking, elevator, etc.)	Life-logging
LENA: An egocentric video database of visual lifelog	http://people.sutd.edu.sg/~1000892/dataset	13 activities performed by 10 subjects (Google Glass)	activity (walking, elevator, etc.)	Life-logging

Datasets (non-exhaustive)

Dataset	URL	Settings	Annotations	Goal
FPPA	http://tamaraberg.com/prediction/Prediction.html	Five subjects performing 5 daily actions	activity (drinking water, putting on clothes, etc.)	Temporal prediction
UT Egocentric	http://vision.cs.utexas.edu/projects/egocentric/index.html	3-5 hours long videos capturing a person's day	important regions	Summarization
VINST/ Visual Diaries	http://www.csc.kth.se/cvap/vinst/NovEgoMotion.html	31 videos capturing the visual experience of a subject walkin from metro station to work	location id, novel egomotion	Novelty detection
Bristol Egocentric Object Interaction (BEOID)	https://www.cs.bris.ac.uk/~damen/BEOID/	8 subjects, six locations. Interaction with objects and environment	gaze, objects, mode of interaction (pick, plug, etc.)	Provide assistance on object usage
Object Search Dataset	https://github.com/Mengmi/deepfuturegaze_gan	57 sequences of 55 subjects on search and retrieval tasks	gaze	gaze prediction

Datasets (non-exhaustive)

Dataset	URL	Settings	Annotations	Goal
UNICT-VEDI	http://iplab.dmi.unict.it/VEDI/	different subjects visiting a museum	location, observed objects	localizing visitors of a museum and estimating their attention
UNICT-VEDI-POI	http://iplab.dmi.unict.it/VEDI_POIs/	different subjects visiting a museum	object bounding boxes annotations, observed objects	recognizing points of interest observed by the visitors
Simulated Egocentric Navigations	http://iplab.dmi.unict.it/SimulatedEgocentricNavigations/	simulated navigations of a virtual agent within a large building	3-DOF pose of the agent in each image	egocentric localization
EgoCart	http://iplab.dmi.unict.it/EgocentricShoppingCartLocalization/	egocentric images collected by a shopping cart in a retail store	3-DOF pose of the shopping cart in each image	egocentric localization
Unsupervised Segmentation of Daily Livign Activities	http://iplab.dmi.unict.it/dailylivingactivities	egocentric videos of daily activities	activities	unsupervised segmentation with respect to the activities

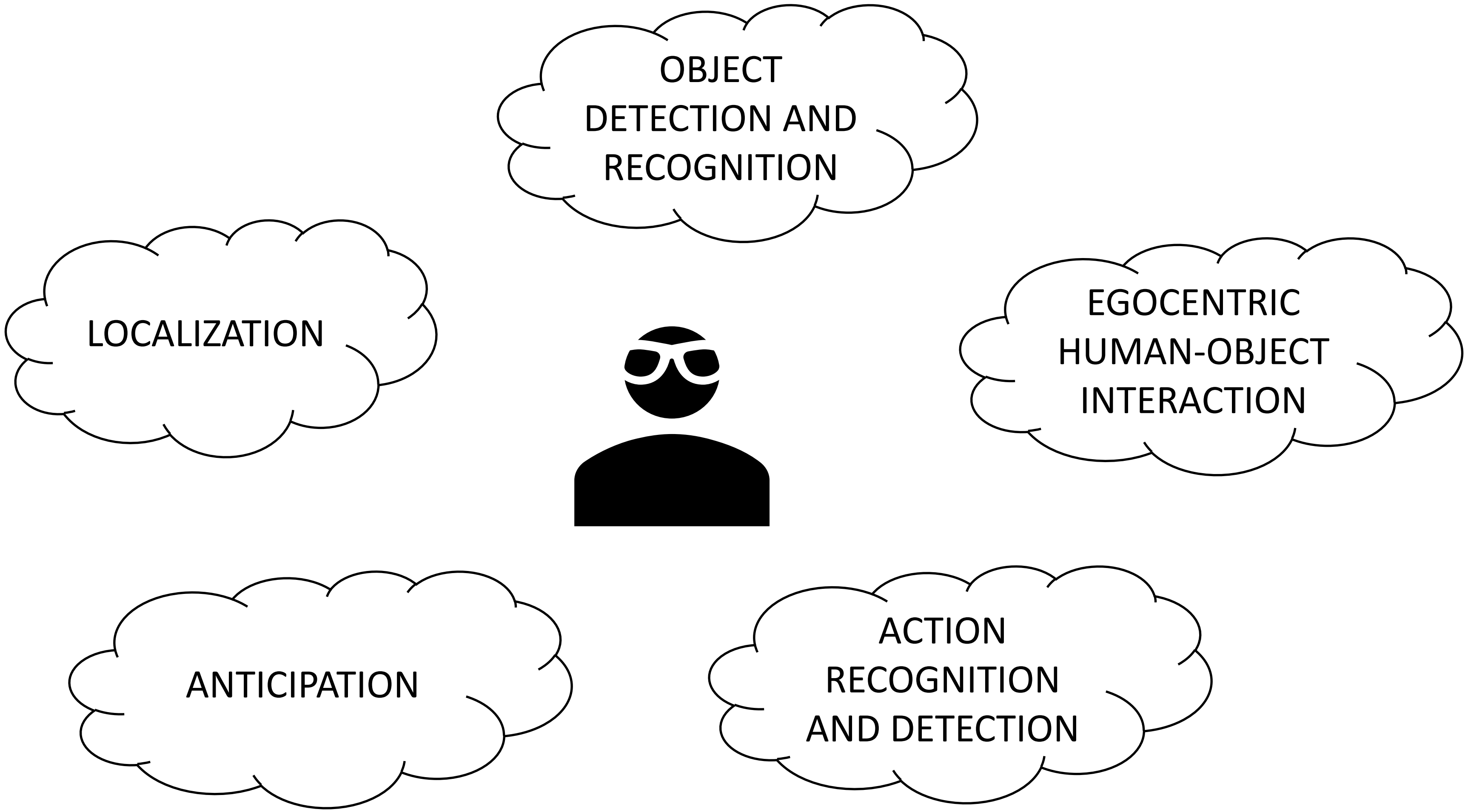
Datasets (non-exhaustive)

Dataset	URL	Settings	Annotations	Goal
Visual Market Basket Analysis	http://iplab.dmi.unict.it/vmba/	egocentric images colelcted by a shopping cart in a retail store	class-location of each image	egocentric localization
Location Based Segmentation of Egocentric Videos	http://iplab.dmi.unict.it/PersonalLocationSegmentation/	egocentric videos of daily activities	location classes	egocentric localization, video indexing
Recognition of Personal Locations from Egocentric Videos	http://iplab.dmi.unict.it/PersonalLocations/	egocentric videos clips of daily activities	location classes	recognizing personal locations
EgoGesture	http://www.nlpr.ia.ac.cn/iva/yfzhang/datasets/egogesture.html	2k videos from 50 subjects performing 83 gestures	Gesture labels, depth	Gesture recognition
EgoHands	http://vision.soic.indiana.edu/projects/egohands/	48 videos of interactions between two people	Hand segmentation masks	Egocentric hand segmentation
DoMSEV	http://www.verlab.dcc.ufmg.br/semantic-hyperlapse/cvpr2018-dataset/	80 hours/different activities	Scene/Action labels with IMU, GPS mad depth	Summarization

Datasets (non-exhaustive)

Dataset	URL	Settings	Annotations	Goal
EGO-HPE	http://imagelab.ing.unimore.it/imagelab2015/researchactivity.asp?idAttivita=23	Egocentric videos for head pose estimation	Head pose of the subjects	Head-pose estimation
EGO-GROUP	http://imagelab.ing.unimore.it/imagelab2015/researchactivity.asp?idAttivita=23	18 videos of people engaging social relationships	Social relationships	Understanding social relationships
DR(eye)VE	http://aimagelab.ing.unimore.it/dreyeve	74 videos of people driving	Eye fixations	Autonomous and assisted driving
THU-READ	http://ivg.au.tsinghua.edu.cn/dataset/THU_READ.php	8 subjects performing 40 actions with a head-mounted RGBD camera	Action segments	RGBD egocentric action recognition
EGO-CH	https://iplab.dmi.unict.it/EGO-CH/	70 subjects visiting two cultural sites in Sicily, Italy.	Temporal segments, room-based localization, objects	Room-basd localization, Object detection, Behavioral analysis

Fundamental Tasks of a First Person Vision System



Localization in First Person Vision

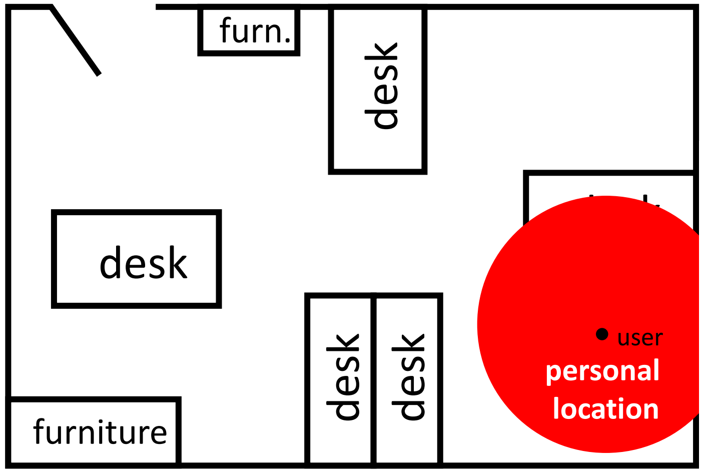
- Knowing the location of the user for a First Person Vision system is important to implement contextual awareness
  - Behave differently depending on the environment
    - Generate reminders when I get to a particular place
      - «remember to do the laundary when you get home»;
    - Turn notifications on or off when you are in given environments:
      - Put in silent mode when I am in a conference room;
  - Help localize/navigate the user
    - E.g., in a retail store or in a museum;
  - Implement augumented reality
    - Show location-specific information when I get to a place (e.g., a room in a museum)

Localization – Levels of Granularity

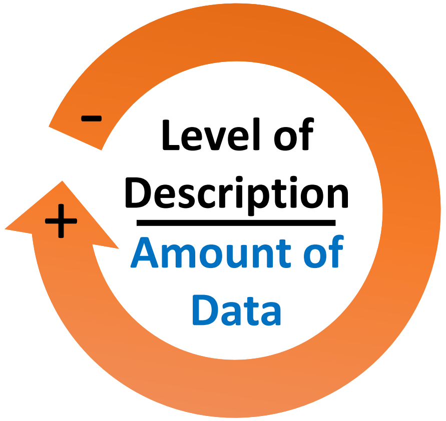
SCENE RECOGNITION

off-the-shelf detectors

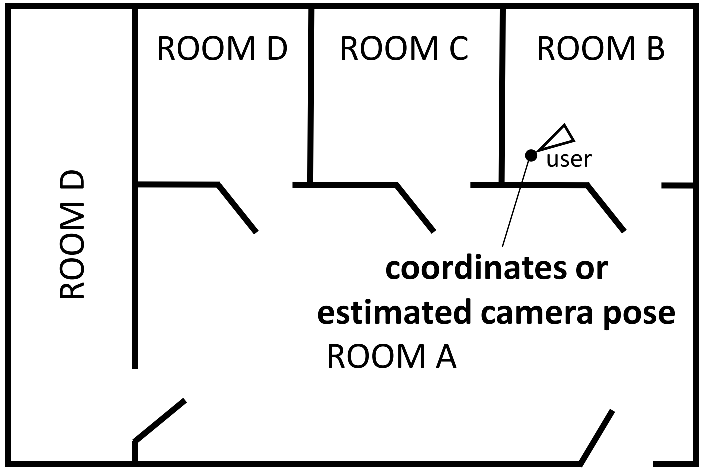
PERSONAL LOCATION RECOGNITION



few training data

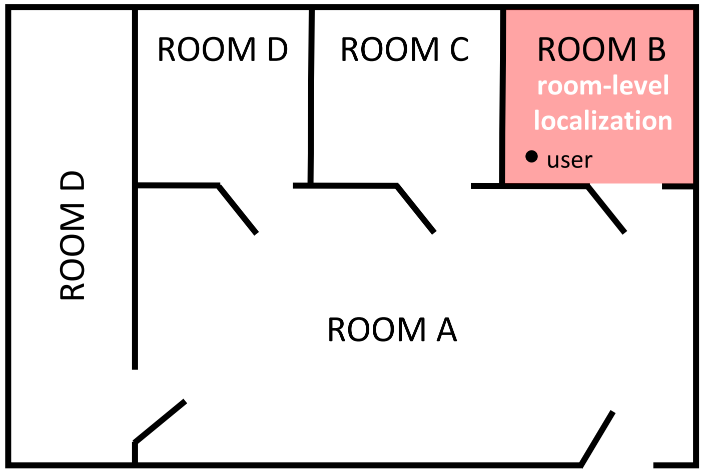


CAMERA POSE-ESTIMATION



3D reconstruction of the building

ROOM-LEVEL RECOGNITION



moderate amount of training data

Scene Recognition

?

Inside city

Street

Highway

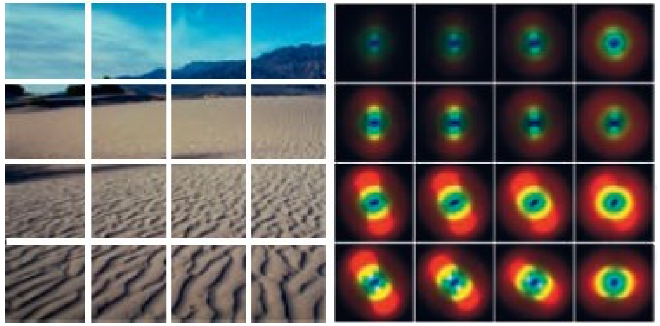
Coast

…

- The most basic form of localization;
- Tells what kind of scene the user is in;
- Useful to distinguish between (even for unseen places) :
  - indoor/outdoor
  - natural/artificial
  - conf. room
  - Office
- Can use off-the-shelf detections.

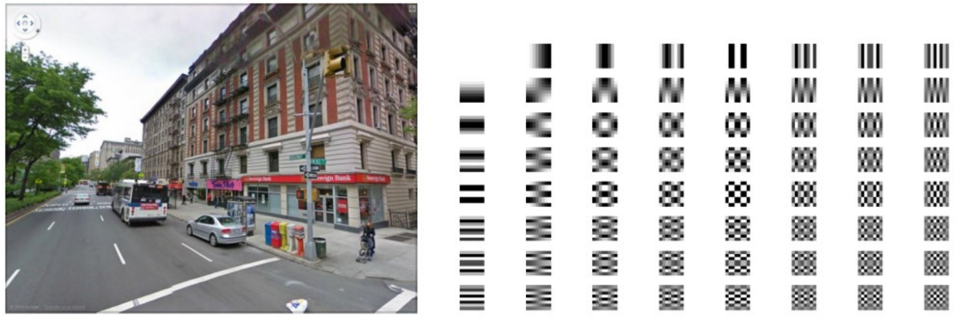
COMPUTATIONALLY INEXPENSIVE ALGORITHMS

GIST Descriptor



Oliva, Aude, and Antonio Torralba. "Modeling the shape of the scene: A holistic representation of the spatial envelope." International journal of computer vision 42.3 (2001): 145-175.

DCT-GIST (runs on the IGP pipeline)



G. M. Farinella, D. Ravì, V. Tomaselli, M. Guarnera, S. Battiato, *"Representing scenes for real-time context classification on mobile devices"*, Pattern Recognition, Elsevier, ISSN 0031-3203, Vol. 48, N. 4, pp. 1082-1096, doi: 10.1016/j.patcog.2014.05.014, 2015

DATA & CODE HERE -> http://places2.csail.mit.edu/

Scene Recognition – Places

GT: cafeteria
top-1: cafeteria (0.179)
top-2: restaurant (0.167)
top-3: dining hall (0.091)
top-4: coffee shop (0.086)
top-5: restaurant patio (0.080)

- Places is a large (10M images – 400+ classes) dataset for scene recognition;
- CNN models trained to recognize 365 scene classes available for download;
- Can be used off-the-shelf!

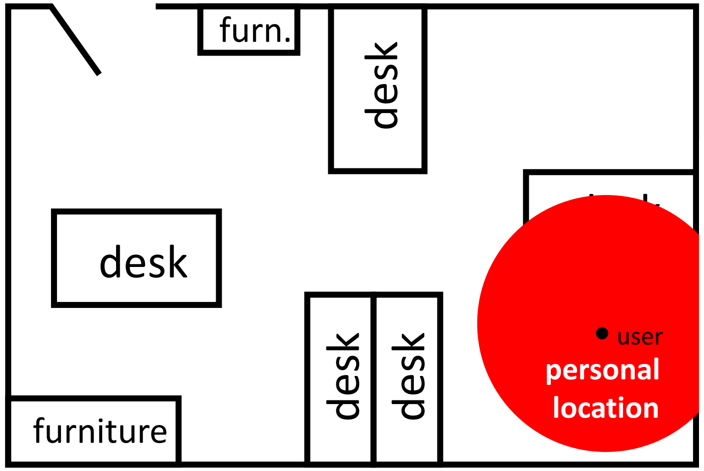
A 10 million Image Database for Scene Recognition B. Zhou, A. Lapedriza, A. Khosla, A. Oliva, and A. Torralba *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017*

Localization – Levels of Granularity

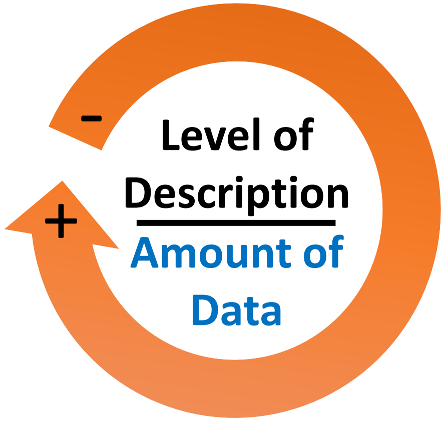
SCENE RECOGNITION

off-the-shelf detectors

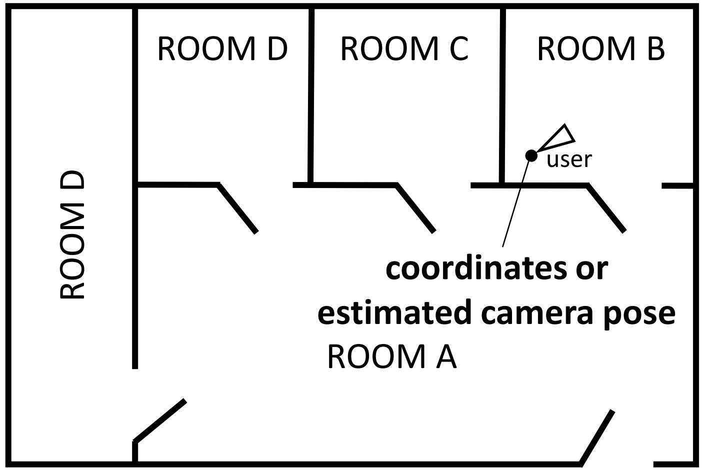
PERSONAL LOCATION RECOGNITION



few training data

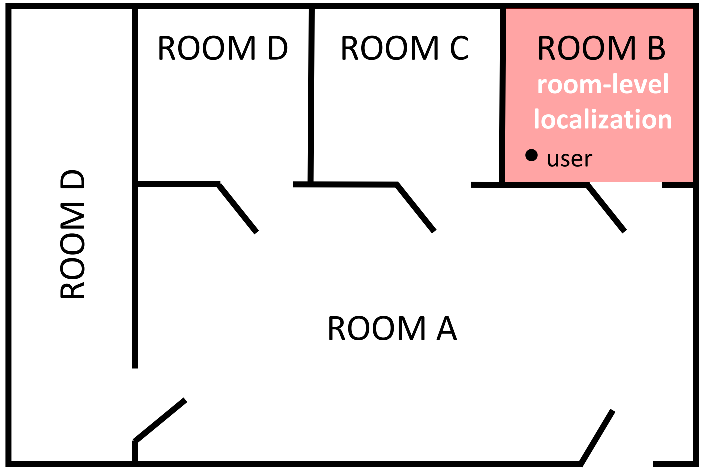


CAMERA POSE-ESTIMATION



3D reconstruction of the building

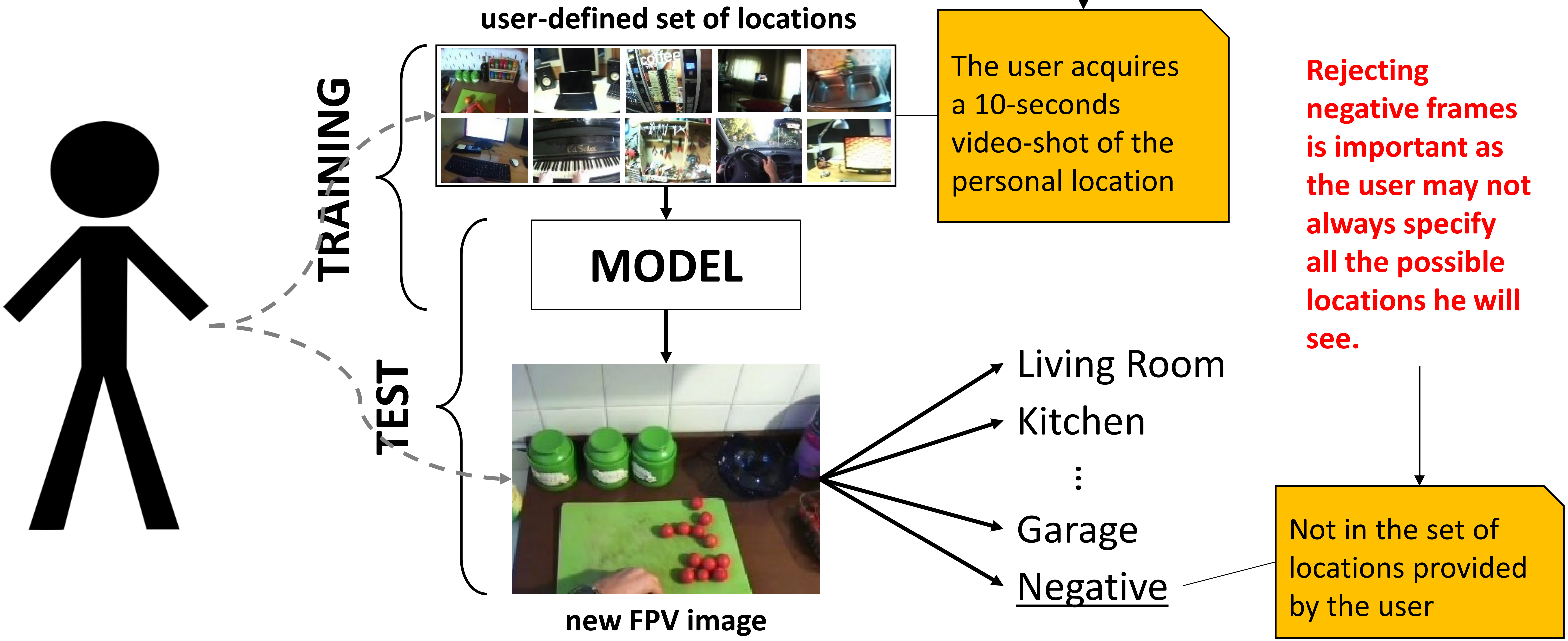
ROOM-LEVEL RECOGNITION



moderate amount of training data

Personal Locations

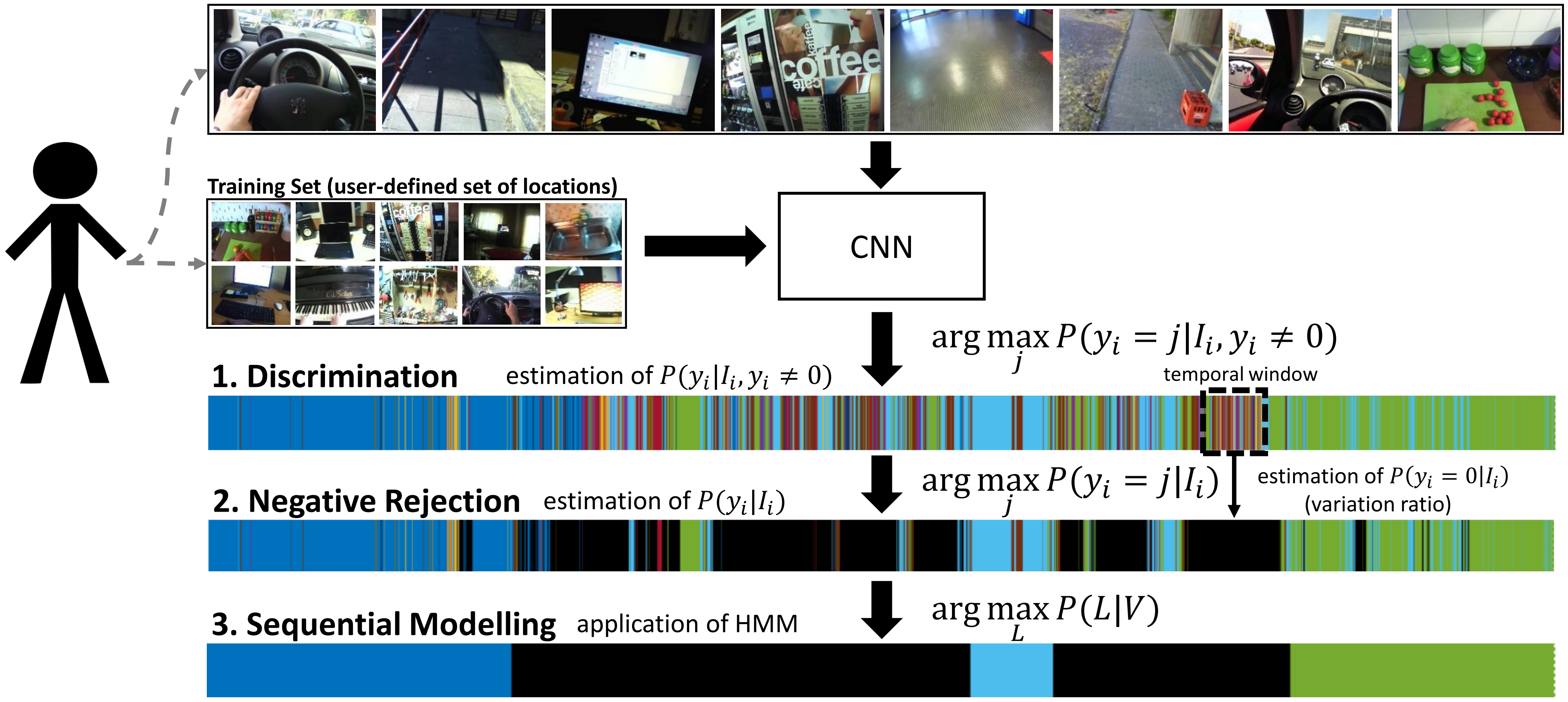
There is no training negatives!



A. Furnari, G. M. Farinella and S. Battiato. Recognizing Personal Locations from Egocentric Videos. IEEE Transactions on Human-Machine Systems, 2017.

CODE HERE -> http://iplab.dmi.unict.it/PersonalLocationSegmentation/

Personal Locations – Full Model



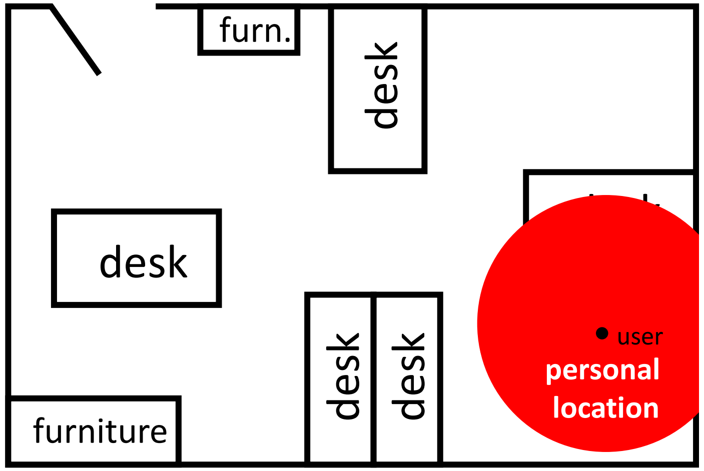
A. Furnari, G. M. Farinella, S. Battiato, Personal-Location-Based Temporal Segmentation of Egocentric Video for Lifelogging Applications, Journal of Visual Communication and Image Representation, 2017.

Localization – Levels of Granularity

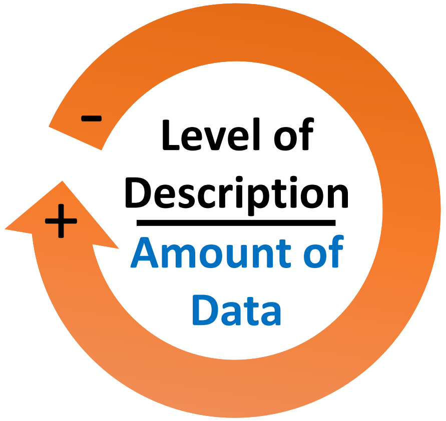
SCENE RECOGNITION

off-the-shelf detectors

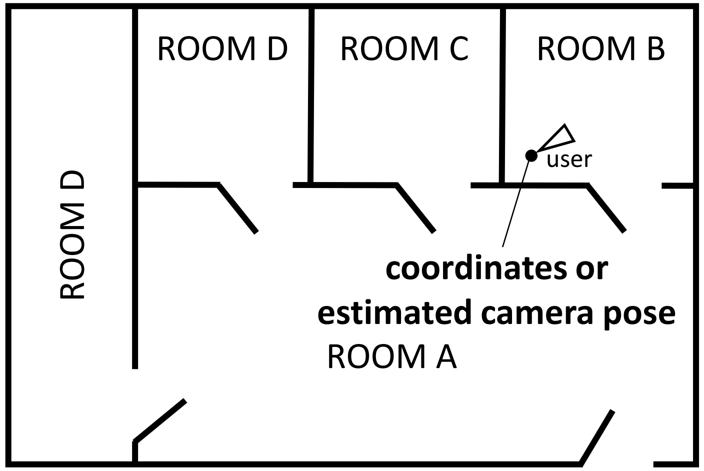
PERSONAL LOCATION RECOGNITION



few training data

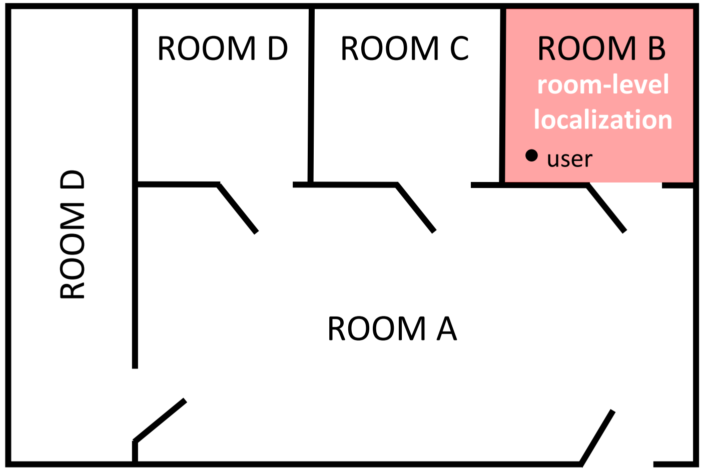


CAMERA POSE-ESTIMATION



3D reconstruction of the building

ROOM-LEVEL RECOGNITION



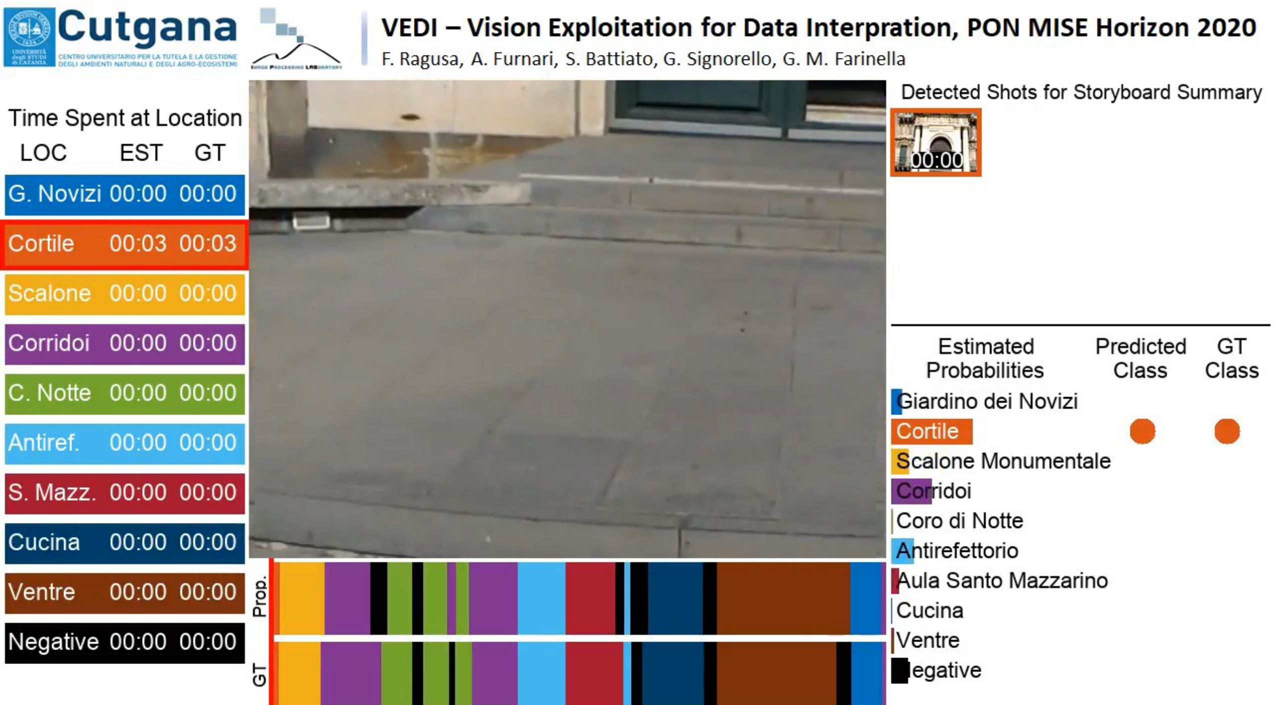
moderate amount of training data

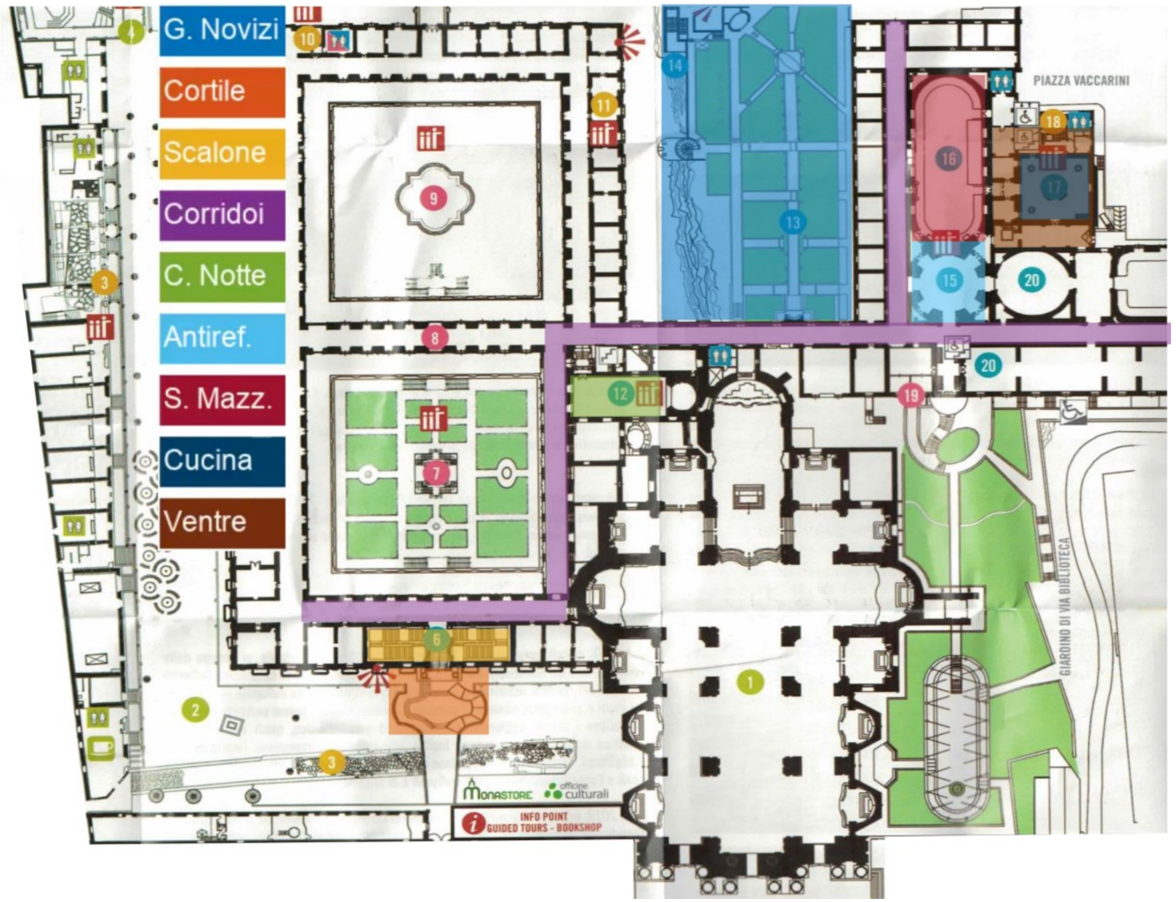
Room-Level Localization

<u>Localizing the user in a larger environment (e.g., a museum).</u>

Extending Personal Location Recognition to Room-Level Localization:

- Collect a longer training video for each room including different points of view;
- Same algorithm as before



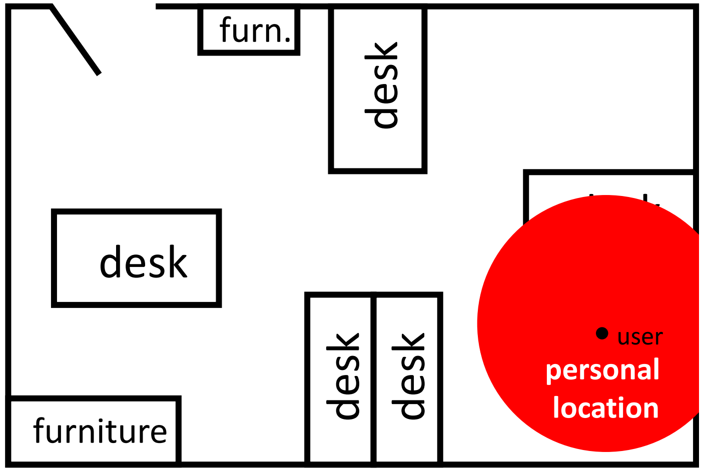


Localization – Levels of Granularity

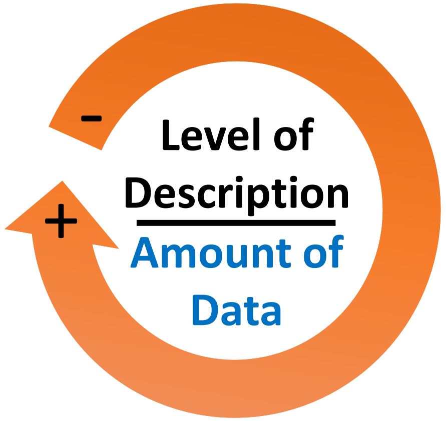
SCENE RECOGNITION

off-the-shelf detectors

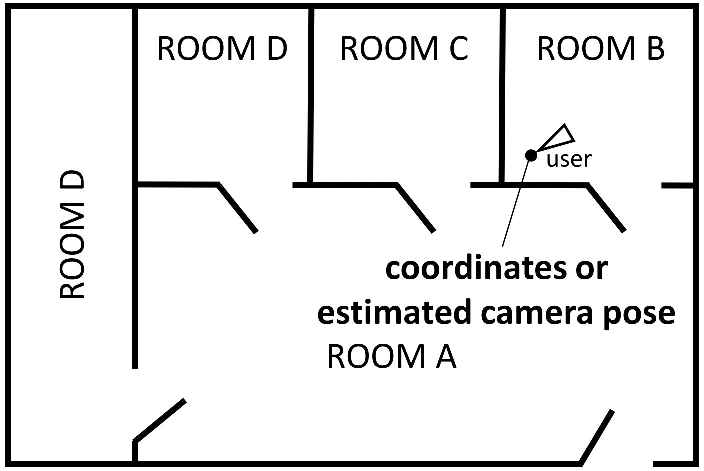
PERSONAL LOCATION RECOGNITION



few training data

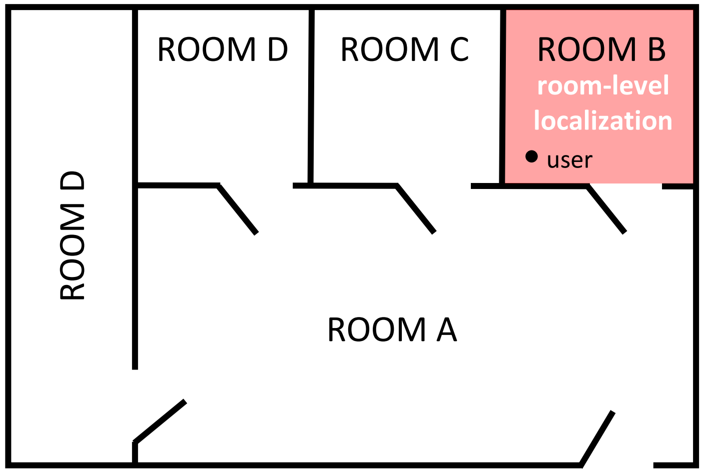


CAMERA POSE-ESTIMATION



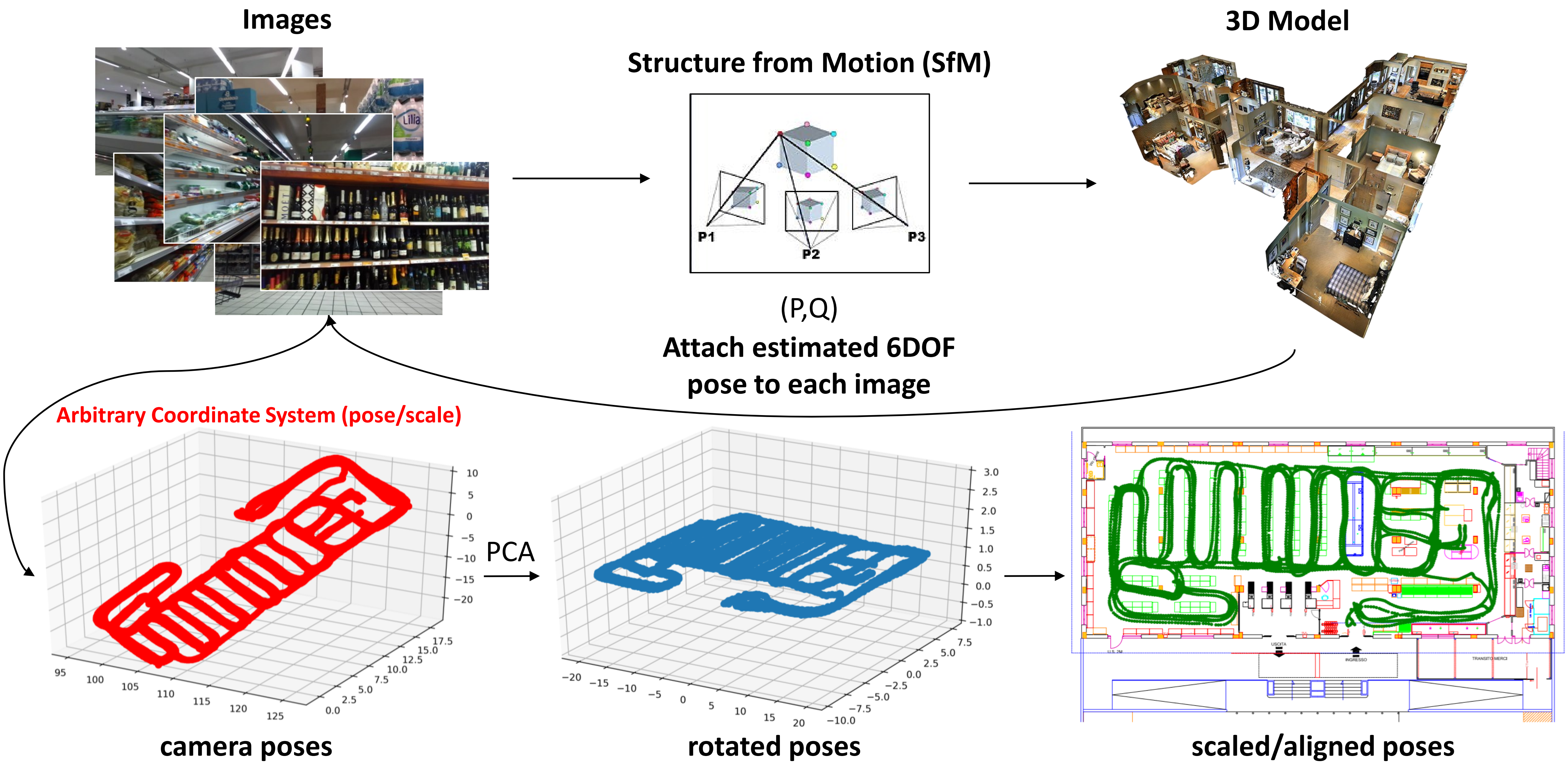
3D reconstruction of the building

ROOM-LEVEL RECOGNITION



moderate amount of training data

Camera Pose Estimation – Dataset Creation



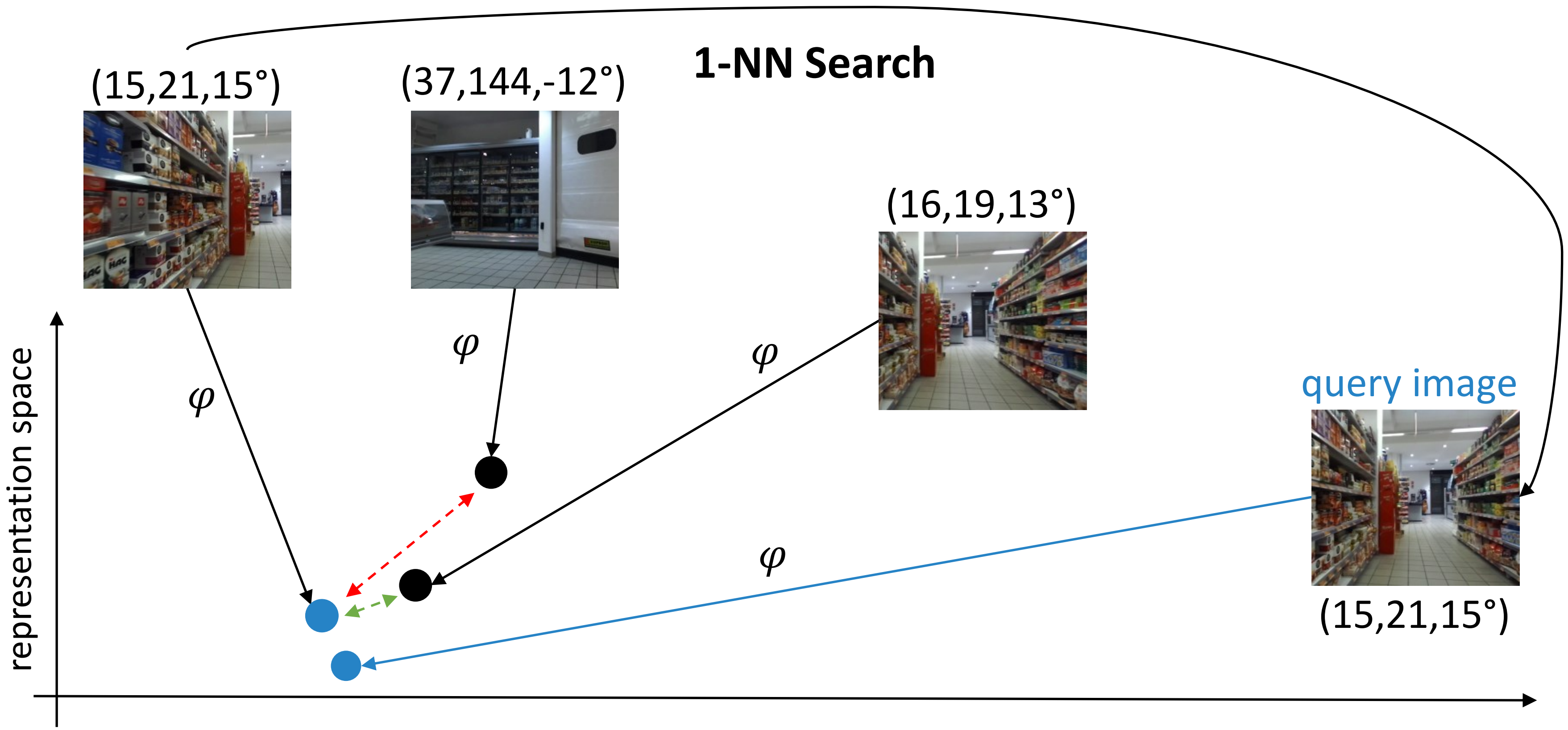
Structure from Motion (SfM) Softwares

Many options available:

- COLMAP (free)
  - https://colmap.github.io/
- Visual SFM (free)
  - http://ccwu.me/vsfm/
- 3D Zephir (paid)
  - https://www.3dflow.net/it/3df-zephyr-pro-3d-models-from-photos/

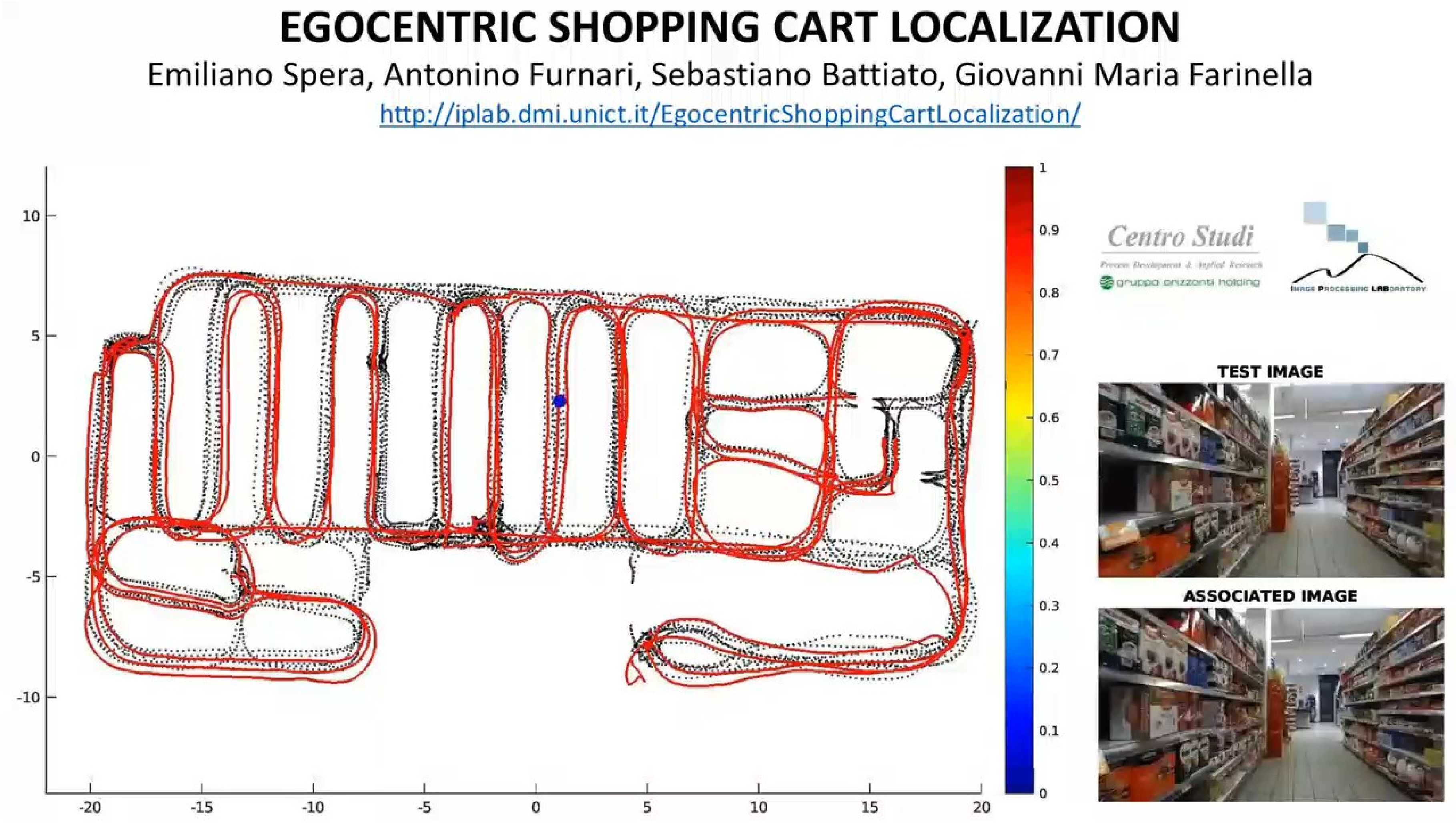
Camera Pose Estimation – Retrieval Approach

Use deep metric learning to <u>learn</u> a representation function φ which maps close to each other images of nearby locations



E. Spera, A. Furnari, S. Battiato, G. M. Farinella, Egocentric Shopping Cart Localization, International Conference on Pattern Recognition (ICPR), 2018
S. A. Orlando, A. Furnari, S. Battiato, G. M. Farinella. Image-Based Localization with Simulated Egocentric Navigations. VISAPP 2019

Camera Pose Estimation – Demo



E. Spera, A. Furnari, S. Battiato, G. M. Farinella, Egocentric Shopping Cart Localization, International Conference on Pattern Recognition (ICPR), 2018

Other approaches to visual localization

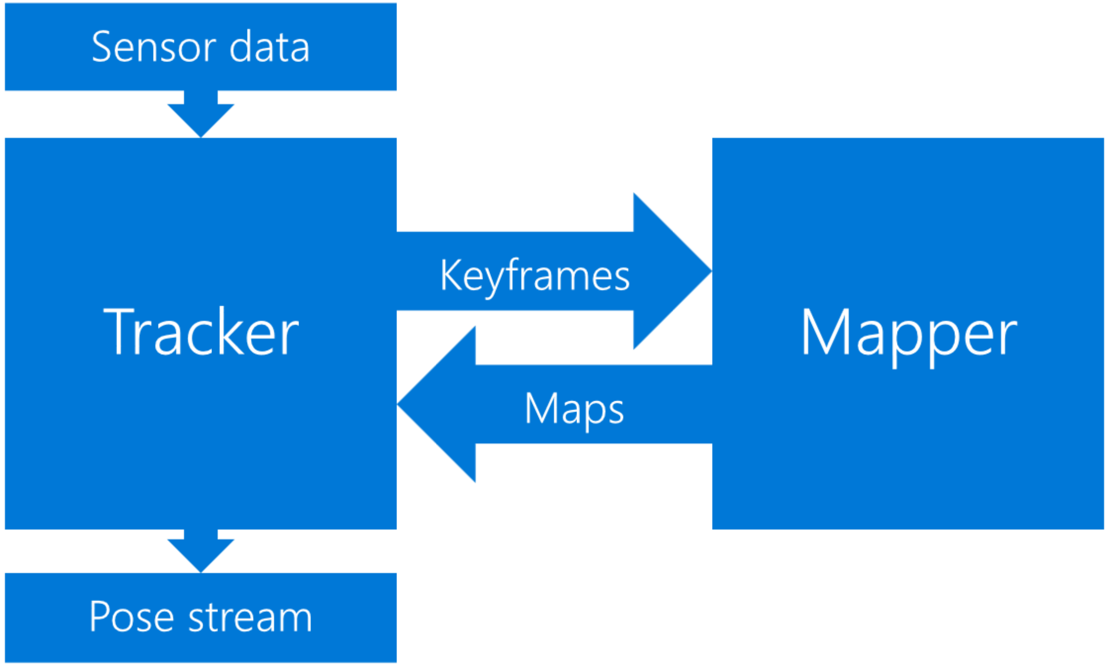
Literature is rich. See here -> https://sites.google.com/view/lsvpr2019/home

Camera Pose Estimation - HoloLens

Microsoft HoloLens implements a localization system which is used for augmented reality.

- The mapper continuously refines a map of the environment;
- The tracker sees small local sub-maps;

Microsoft provides an API to access HoloLens's location information, which can be used by other apps.



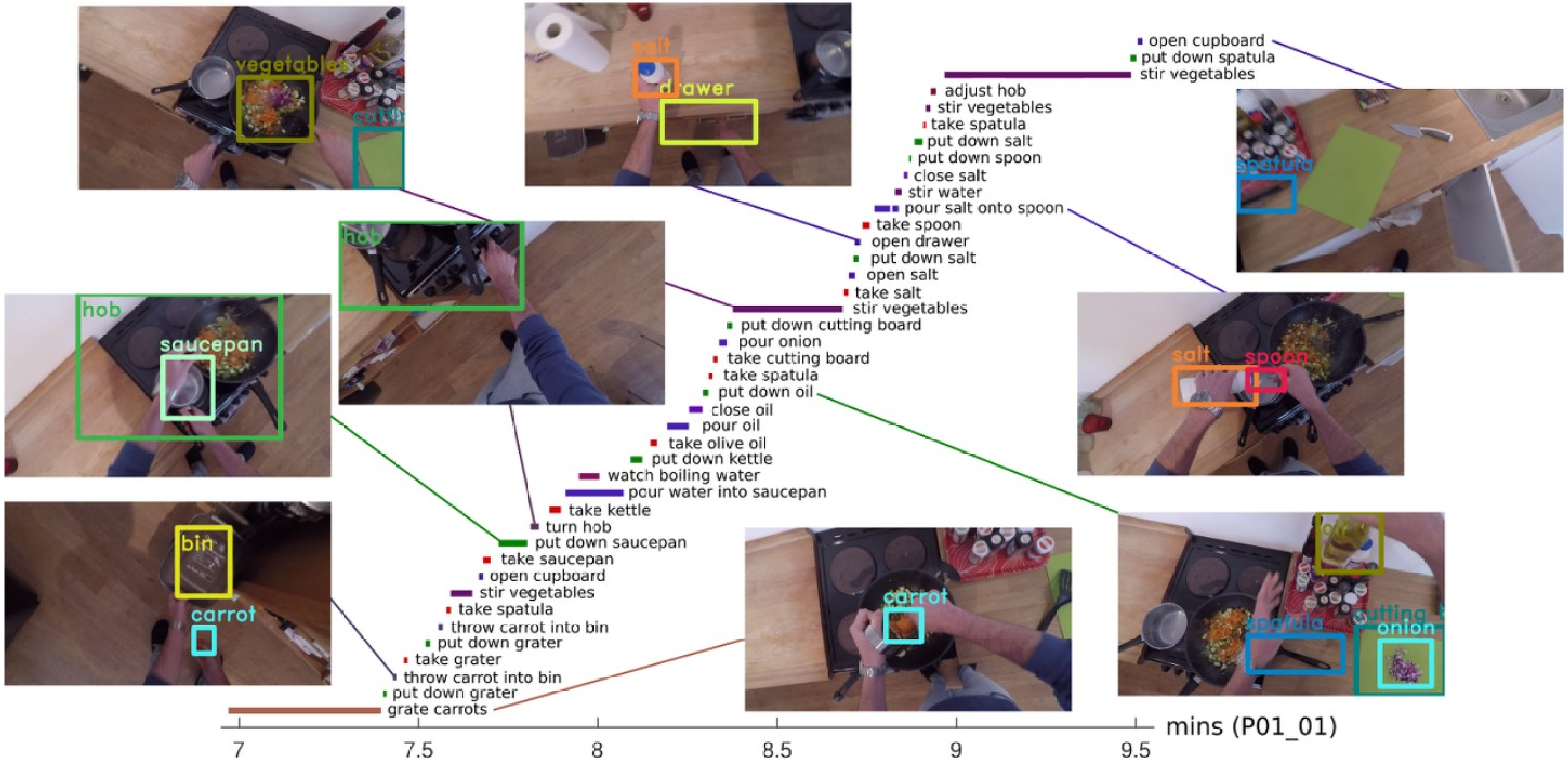
https://docs.microsoft.com/en-us/windows/mixed-reality/coordinate-systems

Objects and Actions are tight!

Useful to know what is in the scene

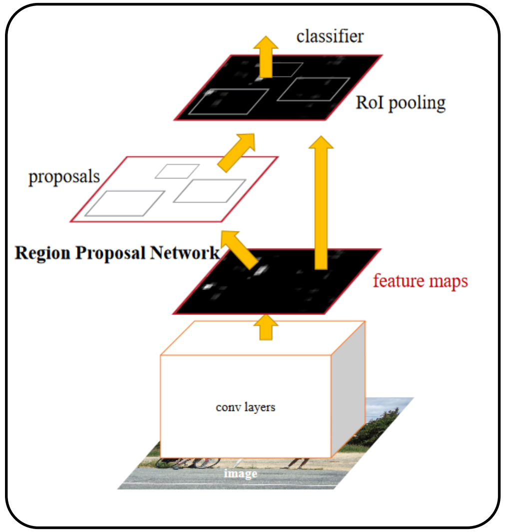
Useful to know what actions can be performed

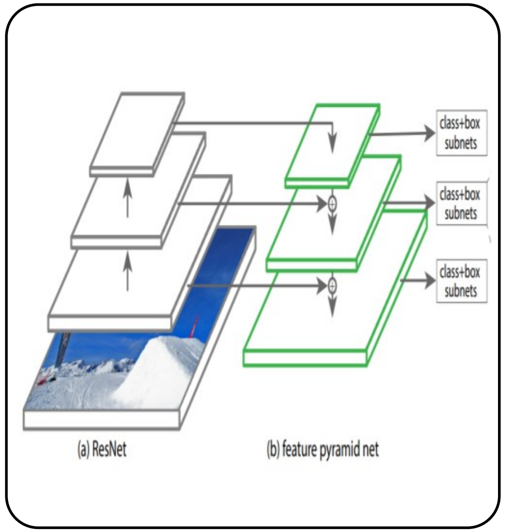
Object Detection

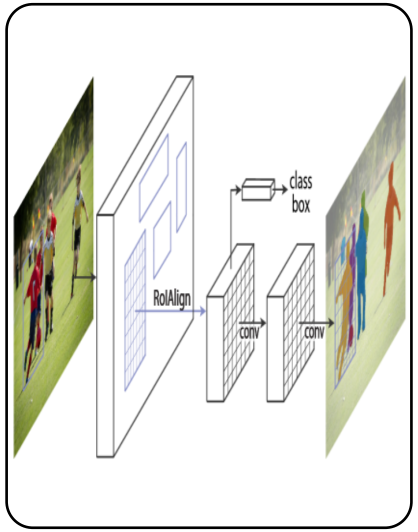


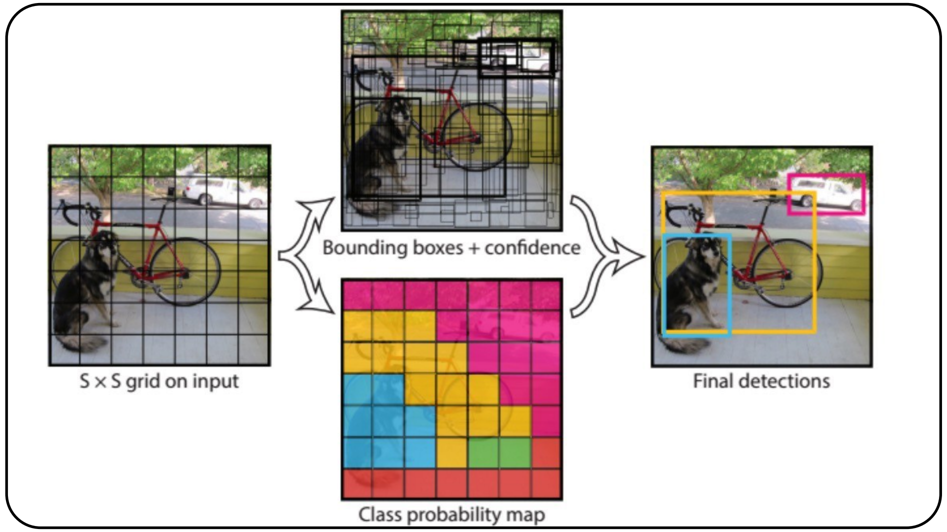
D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro and T. Perrett, W. Price, M. Wray (2018). Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In European Conference on Computer Vision.

Off-the-shelf object detectors









Faster-RCNN
(bounding boxes)

RetinaNet
(bounding boxes - faster)

Mask-RCNN
(boxes + segments)

YOLO
(much faster, but less accurate)

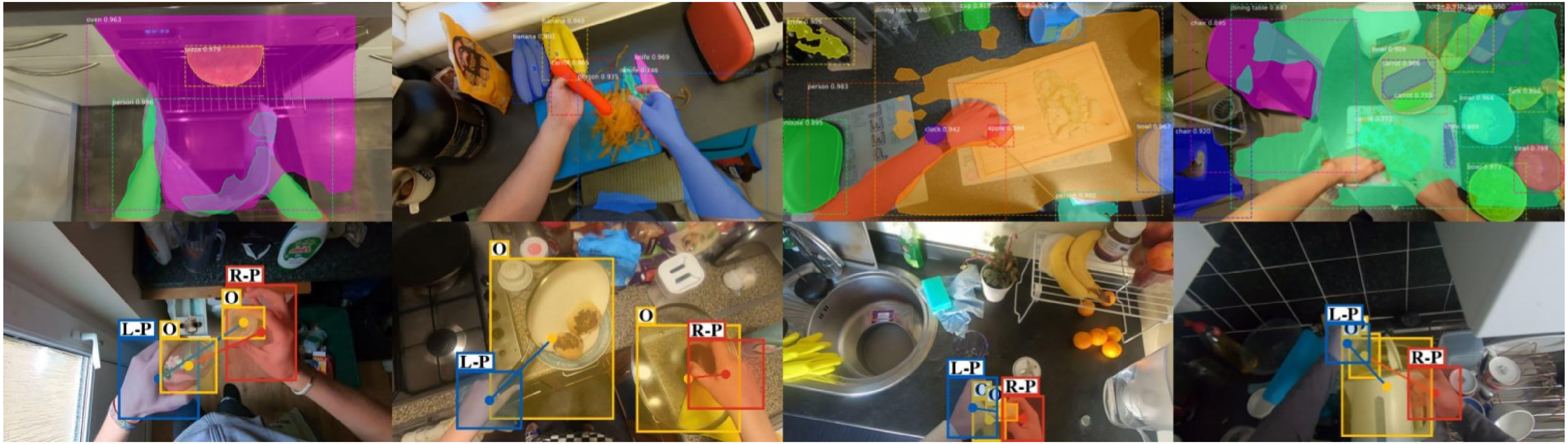
https://github.com/facebookresearch/detectron2

https://pjreddie.com/darknet/yolo/

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*.
Joseph Redmon, Ali Farhadi, YOLO9000: Better, Faster, Stronger, The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017
He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017, October). Mask r-cnn. In *Computer Vision (ICCV), 2017* (pp. 2980-2988). IEEE.

Off-the-shelf detectors on EPIC-KITCHENS

Depending on the scenario, off-the-shelf detectors can be a starting point, but they are not always accurate.



Damen, Doughty, Farinella, Furnari, Kazakos, Moltisanti, Munro, Price, Wray (2020). Rescaling Egocentric Vision. *arXiv preprint arXiv:2006.13256* (2020).

Train/Finetune your own object detector

https://www.csee.umbc.edu/~hpirsiav/papers/ADLdataset/

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

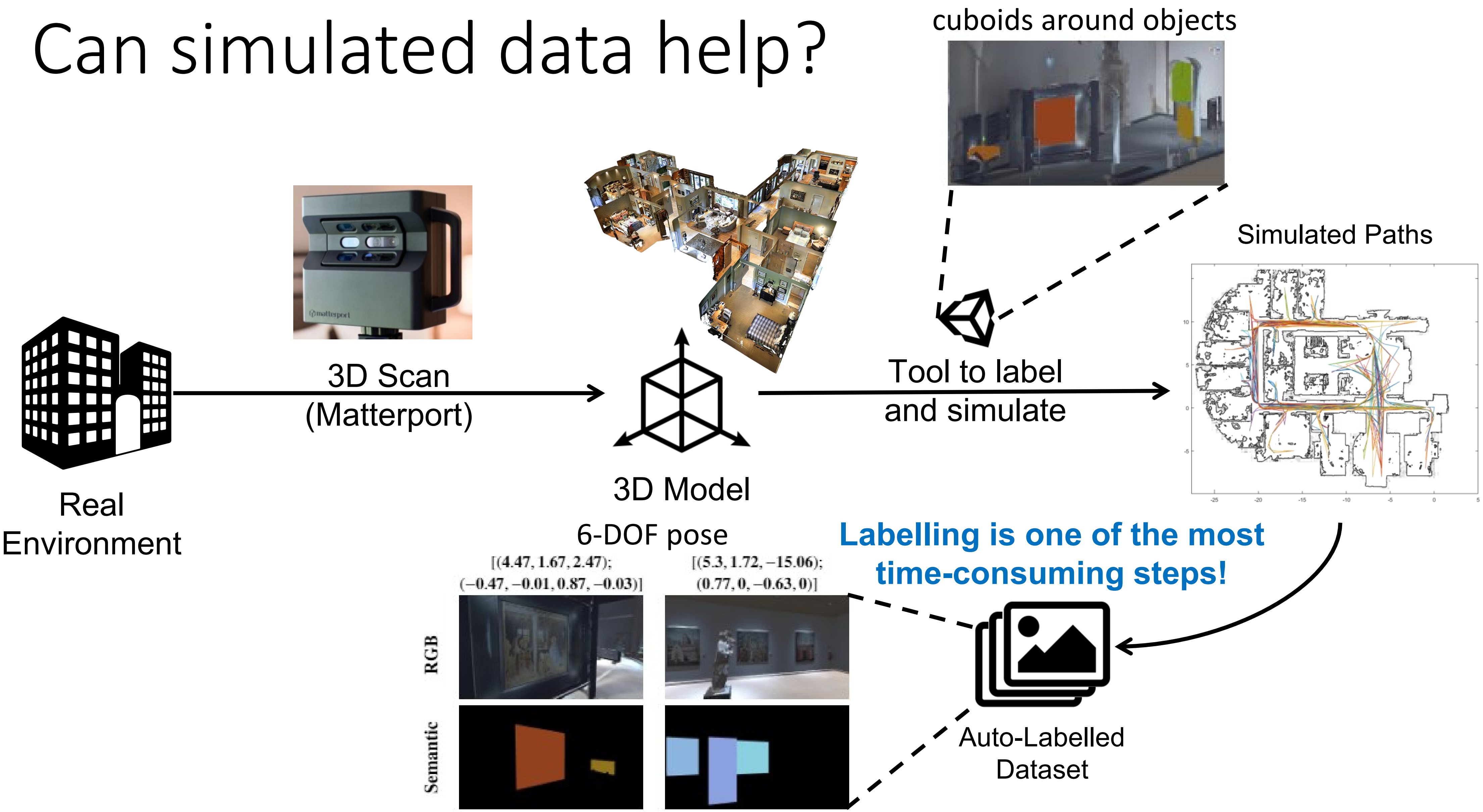
https://iplab.dmi.unict.it/EGO-CH/

https://iplab.dmi.unict.it/MECCANO/

- In some scenario, it could be necessary to fine-tune an object-detector with application-specific data.
- On the left: main egocentric datasets providing bounding box annotations.
- Recently, EGO4D has been released and it has been annotated with bounding boxes.

DATA & CODE HERE -> http://iplab.dmi.unict.it/SimulatedEgocentricNavigations/

Can simulated data help?



S. Orlando, A. Furnari, G. M. Farinella (2020). Egocentric Visitor Localization and Artwork Detection in Cultural Sites Using Synthetic Data . Pattern Recognition Letters - Special Issue on Pattern Recognition and Artificial Intelligence Techniques for Cultural Heritage.

Domain Gap Between Real and Synthetic Images

- Synthetic Images are Easier to collect.
- However, there is significant domain GAP between real and synthetic images;
- As a result, methods trained only on synthetic images will not work directly with real ones.

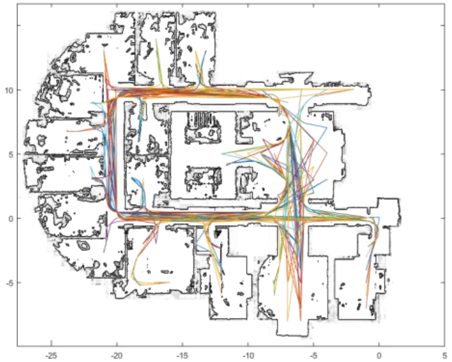
An object detection method trained on synthetic images (left), does not perform well on real images (right)

DATA & CODE HERE -> http://iplab.dmi.unict.it/SimulatedEgocentricNavigations/

Can simulated data help?

cuboids around objects

Simulated Paths



3D Scan
(Matterport)

Tool to label
and simulate

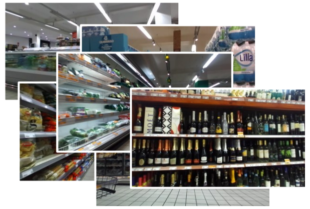
Real
Environment

3D Model

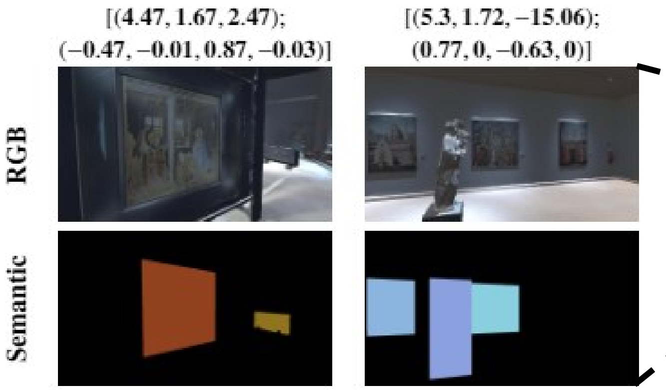
6-DOF pose

Labelling is one of the most time-consuming steps!

real images



domain
adaptation

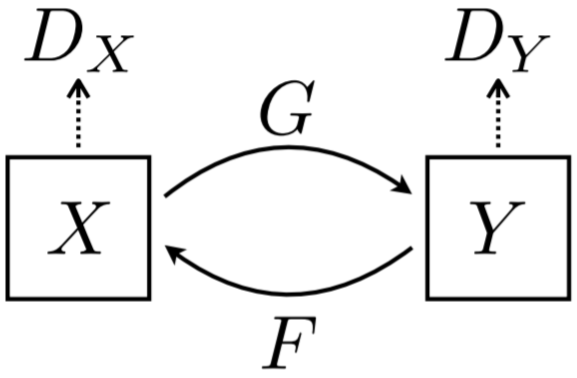


Auto-Labelled
Dataset

S. Orlando, A. Furnari, G. M. Farinella (2020). Egocentric Visitor Localization and Artwork Detection in Cultural Sites Using Synthetic Data . Pattern Recognition Letters - Special Issue on Pattern Recognition and Artificial Intelligence Techniques for Cultural Heritage.

Domain Adaptation Through Image to Image Translation

- One way to reduce the domain gap is by making the synthetic and real images look similar.
- This can be done using image to image translation.
- CycleGAN uses two discriminators to learn two mappings (from synthetic to real and vice versa) which make transformed images undistinguishable.



Zhu, J. Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 2223-2232).

DATA HERE -> https://iplab.dmi.unict.it/EGO-CH-OBJ-SEG/

Domain Adaptation for Semantic Object Segmentation Dataset

Synthetic Images

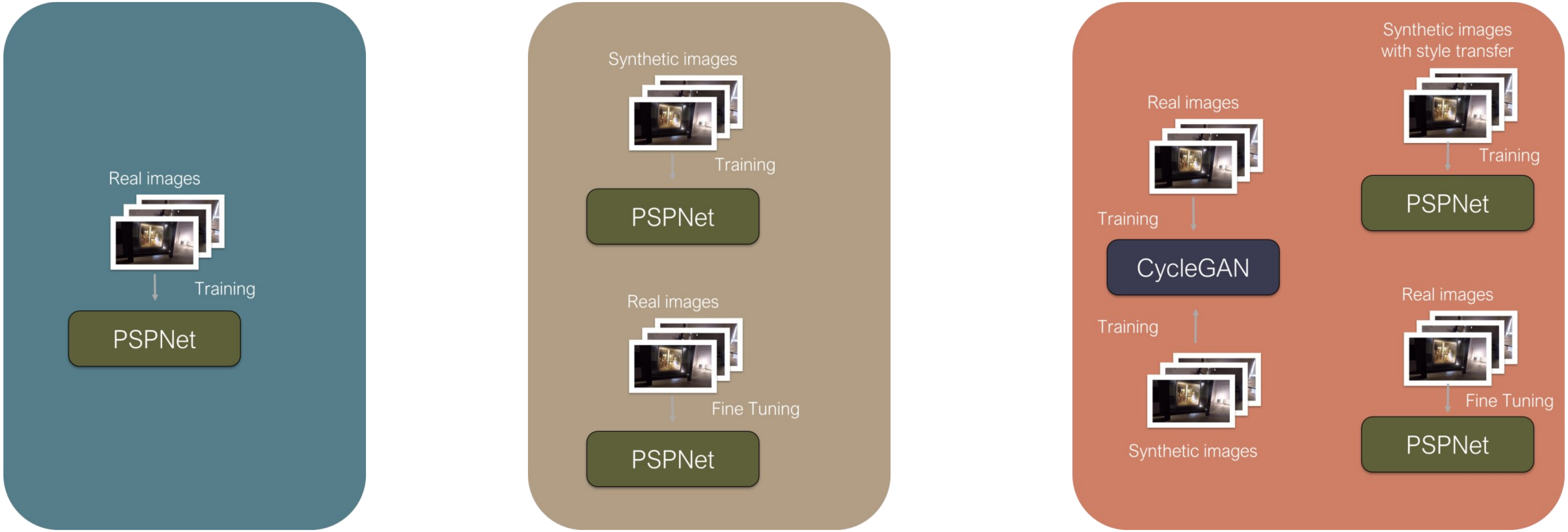
Real Images

24 objects, ~25k synthetic images, ~5k real labeled images, semantic segmentations masks

Francesco Ragusa, Daniele DiMauro, Alfio Palermo, Antonino Furnari, Giovanni Maria Farinella (2020). Semantic Object Segmentation in Cultural Sites using Real and Synthetic Data. International Conference on Pattern Recognition (ICPR).

DATA HERE -> https://iplab.dmi.unict.it/EGO-CH-OBJ-SEG/

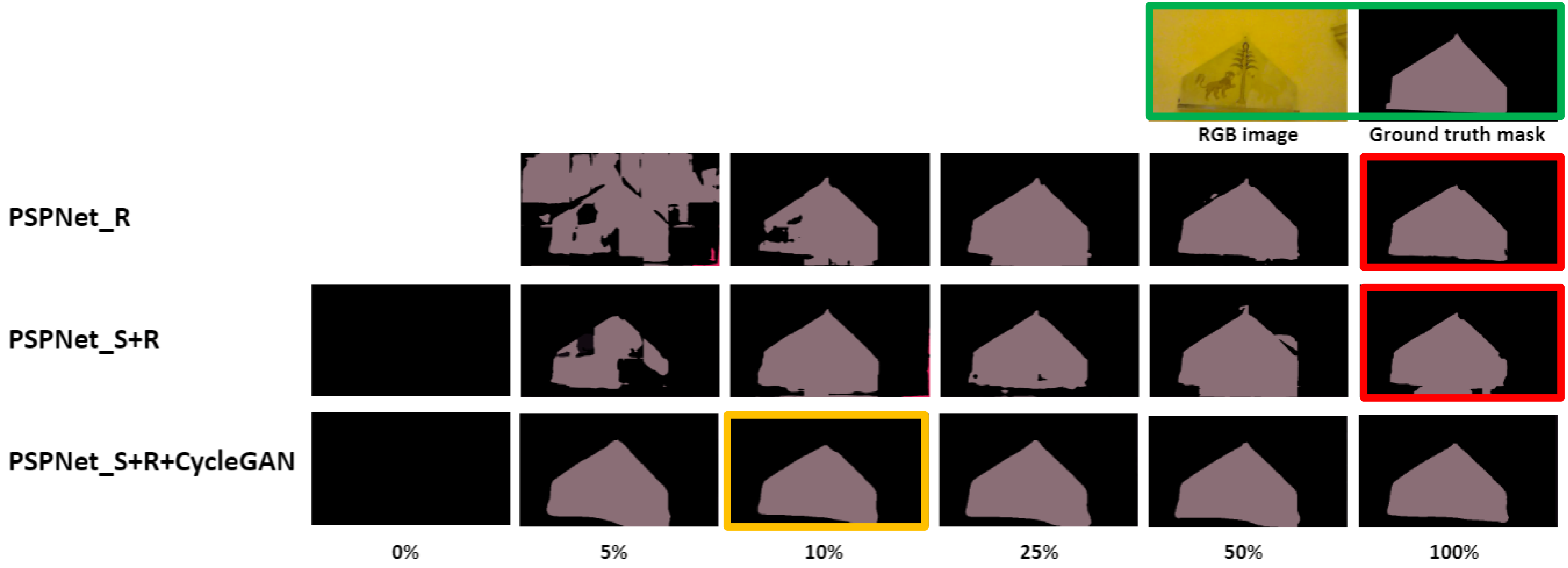
Domain Adaptation for Semantic Object Segmentation Dataset



Francesco Ragusa, Daniele DiMauro, Alfio Palermo, Antonino Furnari, Giovanni Maria Farinella (2020). Semantic Object Segmentation in Cultural Sites using Real and Synthetic Data. International Conference on Pattern Recognition (ICPR).

DATA HERE -> https://iplab.dmi.unict.it/EGO-CH-OBJ-SEG/

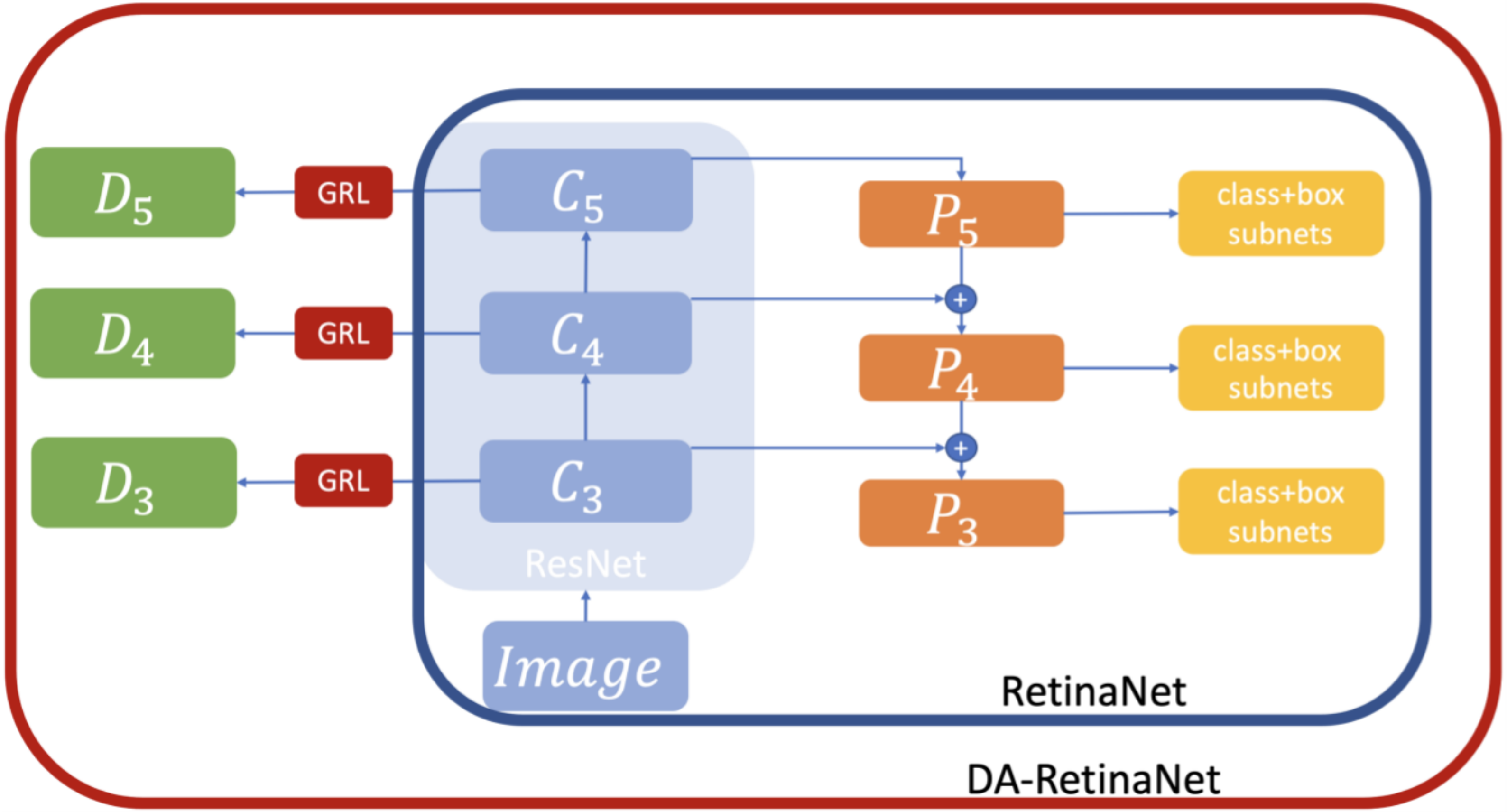
Domain Adaptation for Semantic Object Segmentation Dataset



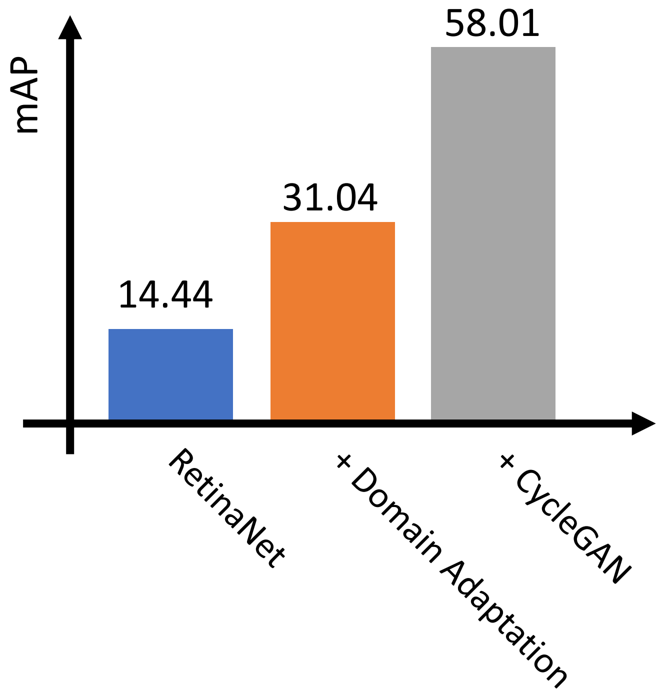
Francesco Ragusa, Daniele DiMauro, Alfio Palermo, Antonino Furnari, Giovanni Maria Farinella (2020). Semantic Object Segmentation in Cultural Sites using Real and Synthetic Data. International Conference on Pattern Recognition (ICPR).

CODE HERE -> https://iplab.dmi.unict.it/EGO-CH-OBJ-UDA

Unsupervised Domain Adaptation for Object Detection

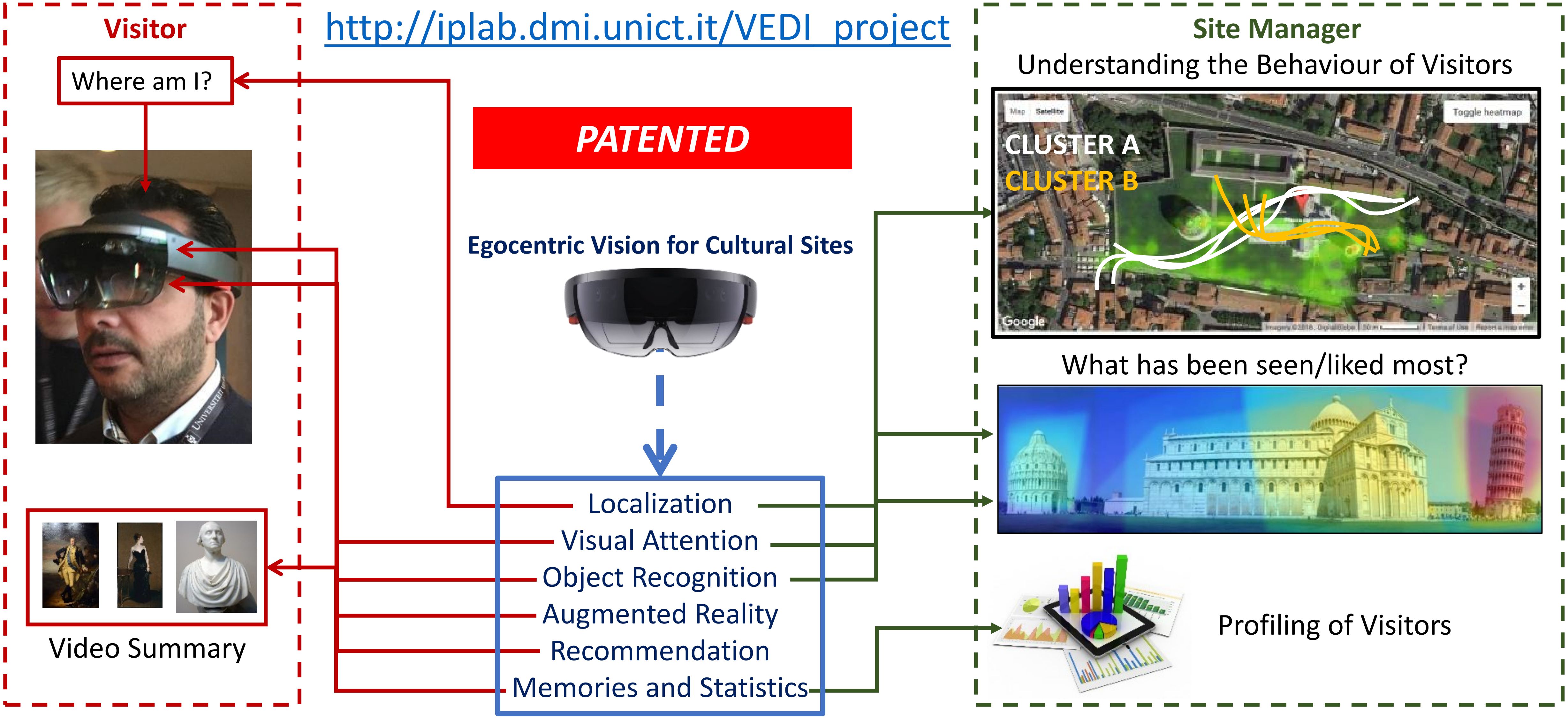


We applied domain adaptation through adversarial learning to the RetinaNet object detector and combined it with image-to-image translation.



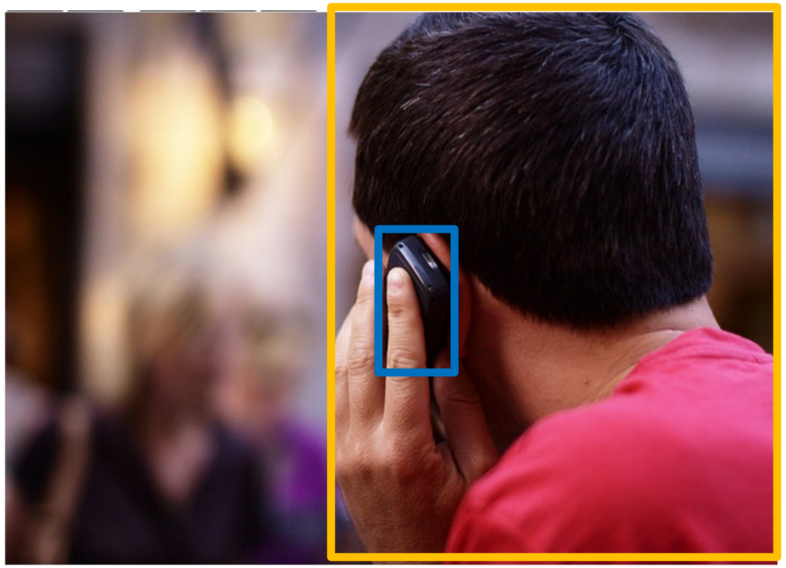
Giovanni Pasqualino, Antonino Furnari, Giovanni Signorello, Giovanni Maria Farinella (2021). An Unsupervised Domain Adaptation Scheme for Single-Stage Artwork Recognition in Cultural Sites. Image and Vision Computing

Vision Exploitation for Data Interpretation (VEDI)

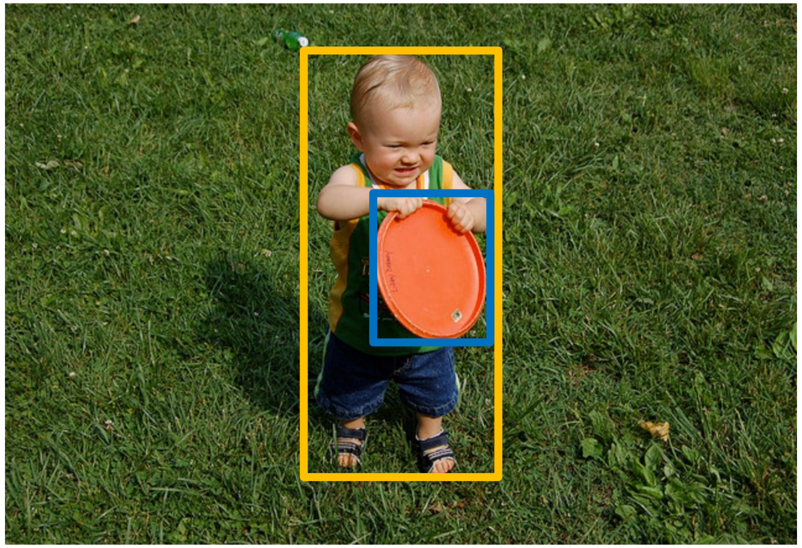


G. M. Farinella, G. Signorello, S. Battiato, A. Furnari, F. Ragusa, R. Leonardi, E. Ragusa, E. Scuderi, A. Lopes, L. Santo, M. Samarotto. VEDI: Vision Exploitation for Data Interpretation. In 20th International Conference on Image Analysis and Processing (ICIAP), 2019

Human-Object Interaction



<human, talks, cellphone>



<human, holds, freesbe>

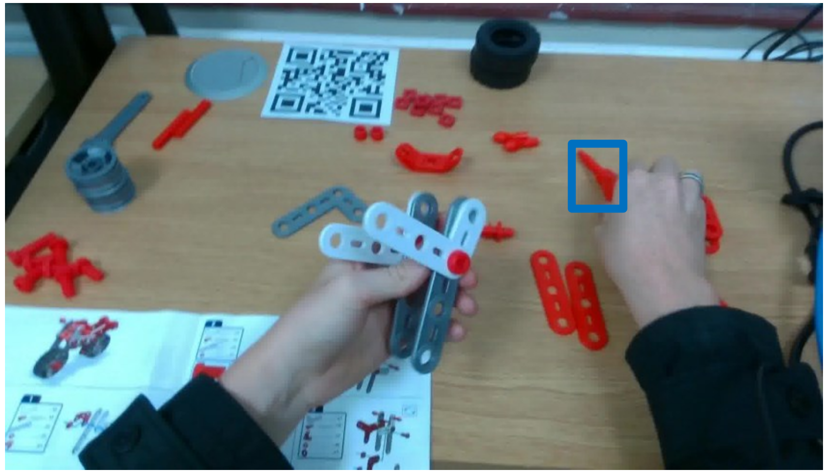
Georgia Gkioxari, Ross Girshick, Piotr Dollàr, Kaiming He. (2018). Detecting Human-Object Interactions. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Egocentric Human-Object Interaction

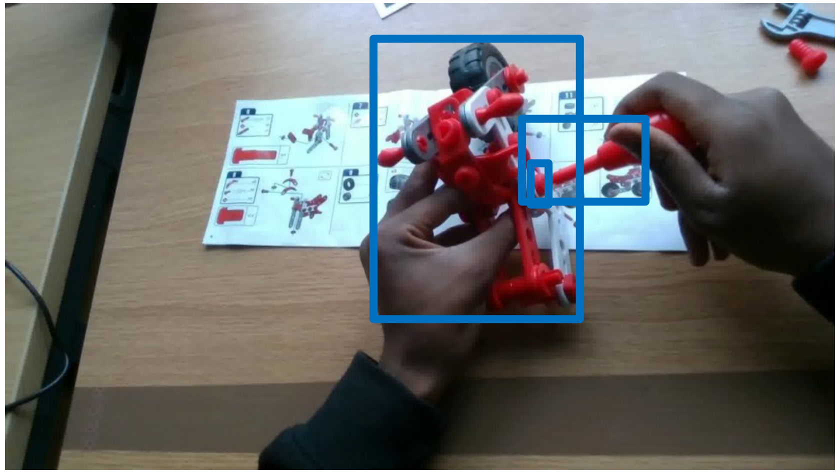
$$O = \{o_1, o_2, \ldots, o_n\}$$

$$V = \{v_1, v_2, \ldots, v_m\}$$

$$e = (v_h, \{o_1, o_2, \ldots, o_i\})$$



<take, screwdriver>

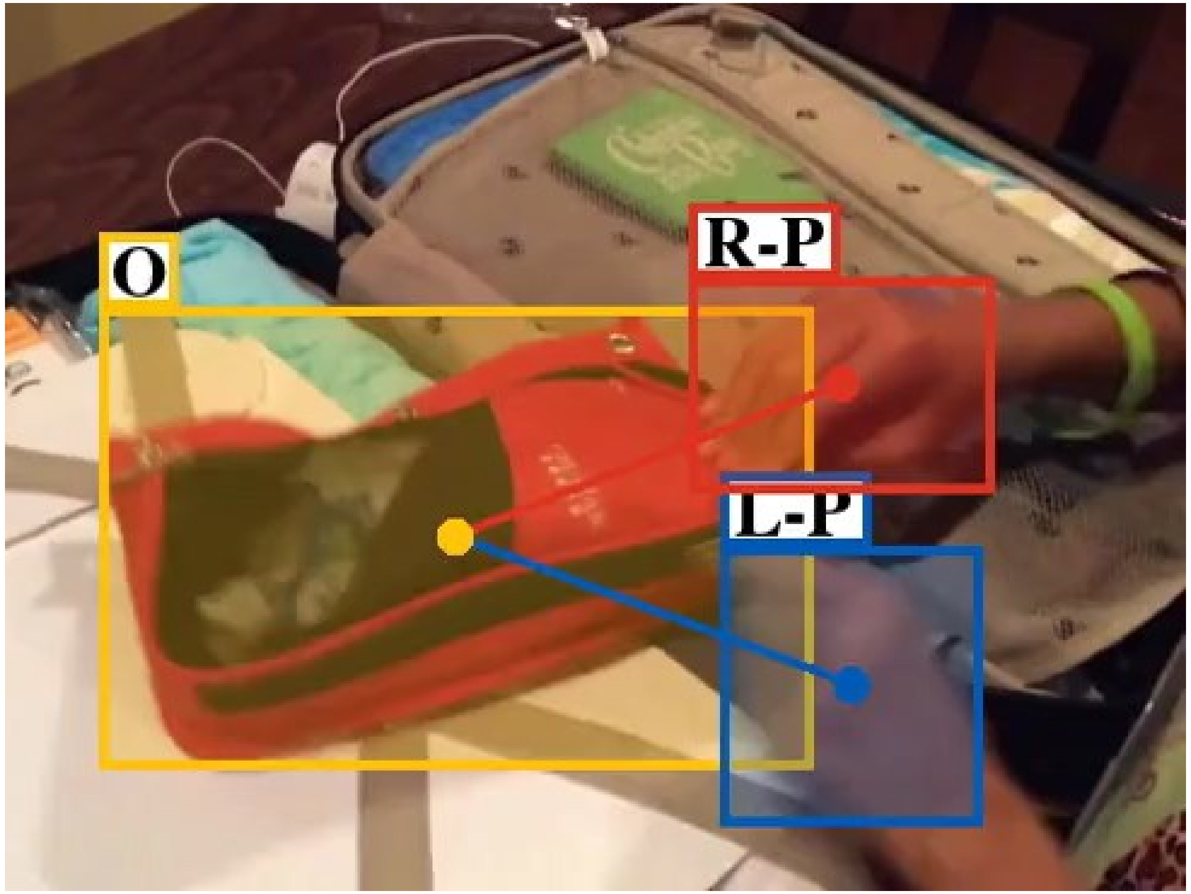


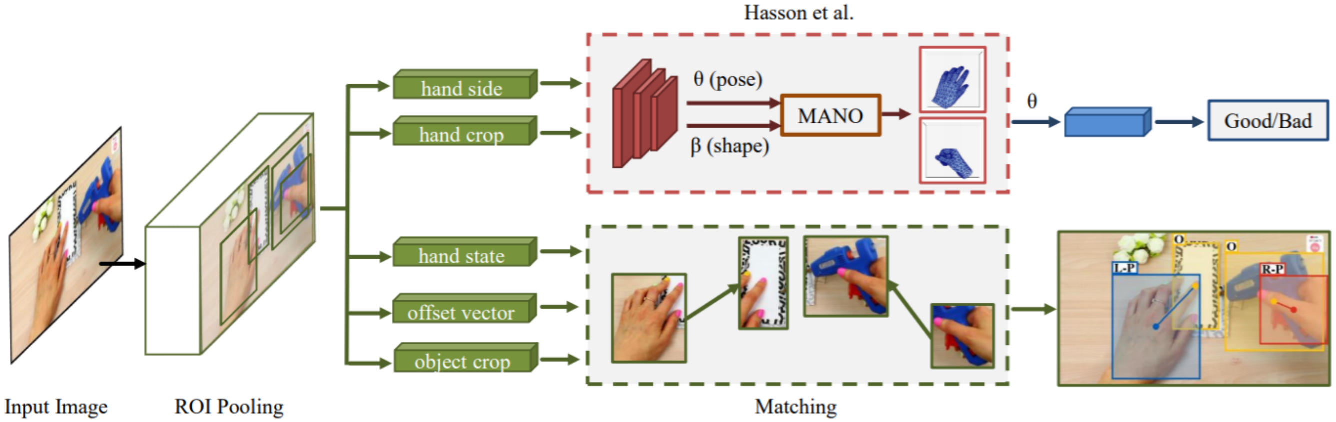
<screw, {screwdriver, screw, partial_model}>

F. Ragusa, A. Furnari, S. Livatino, G. M. Farinella. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. In IEEE Winter Conference on Application of Computer Vision (WACV), 2021. **ORAL**

CODE & DATA HERE -> https://fouheylab.eecs.umich.edu/~dandans/projects/100DOH/

Hands in Contact – Hands + Objects





An «augmented» detector which recognizes:

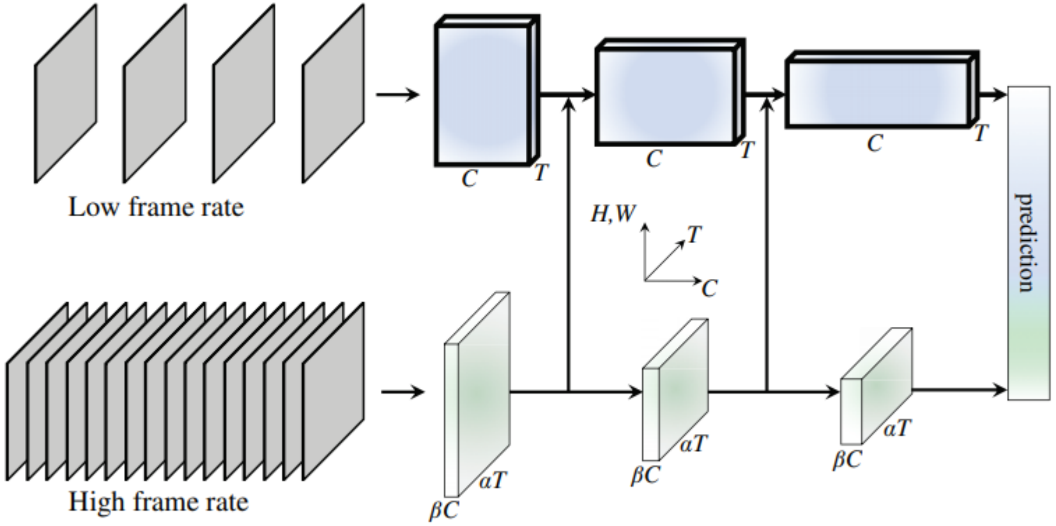
- The left hand;
- The right hand;
- The interacted object.

Shan, D., Geng, J., Shu, M., & Fouhey, D. F. (2020). Understanding human hands in contact at internet scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9869-9878).

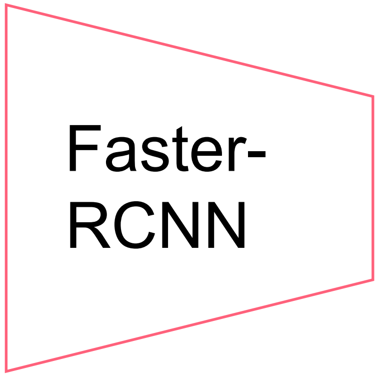
CODE & DATA HERE -> https://github.com/fpv-iplab/MECCANO

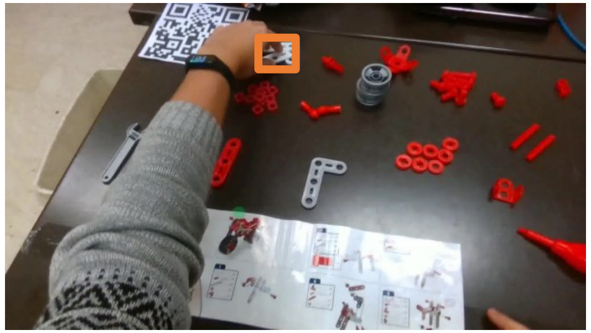
Egocentric Human-Object Interaction

Video Based



<take>



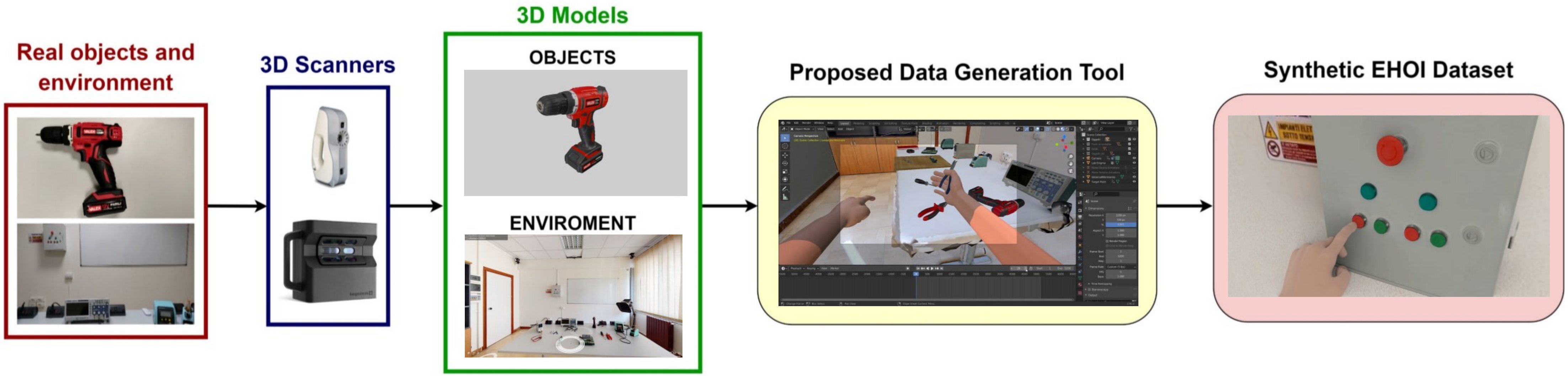


<white_bar>

F. Ragusa, A. Furnari, S. Livatino, G. M. Farinella. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. In IEEE Winter Conference on Application of Computer Vision (WACV), 2021. **ORAL**

DATA HERE -> https://iplab.dmi.unict.it/EHOI_SYNTH/

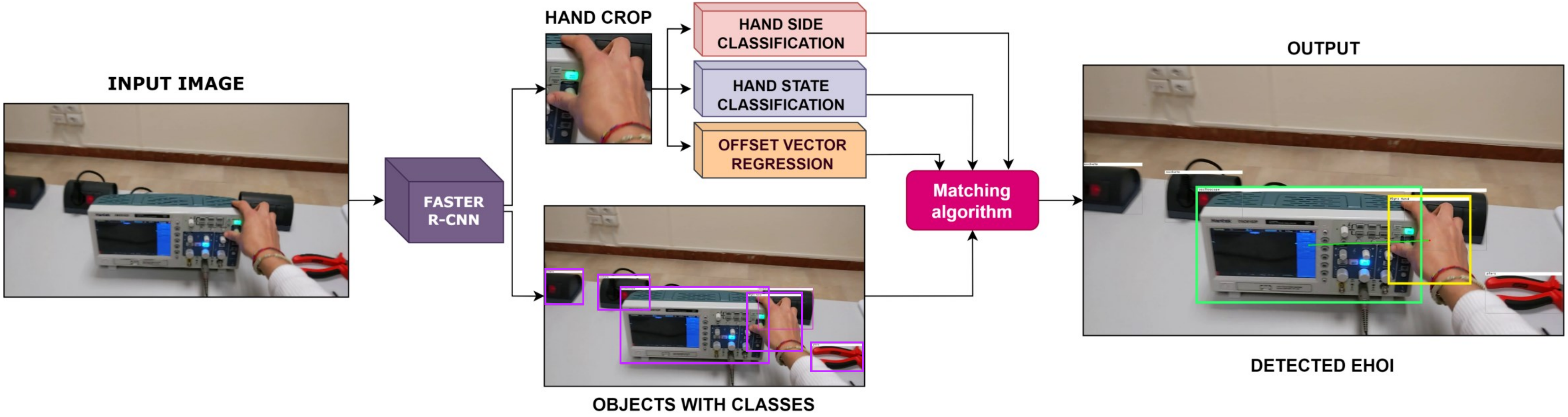
Can simulated data help?



R. Leonardi, F. Ragusa, A. Furnari, G. M. Farinella (2022). Egocentric Human-Object Interaction Detection Exploiting Synthetic Data. In 21st International Conference on Image Analysis and Processing.

DATA HERE -> https://iplab.dmi.unict.it/EHOI_SYNTH/

Can simulated data help?



R. Leonardi, F. Ragusa, A. Furnari, G. M. Farinella (2022). Egocentric Human-Object Interaction Detection Exploiting Synthetic Data. In 21st International Conference on Image Analysis and Processing.

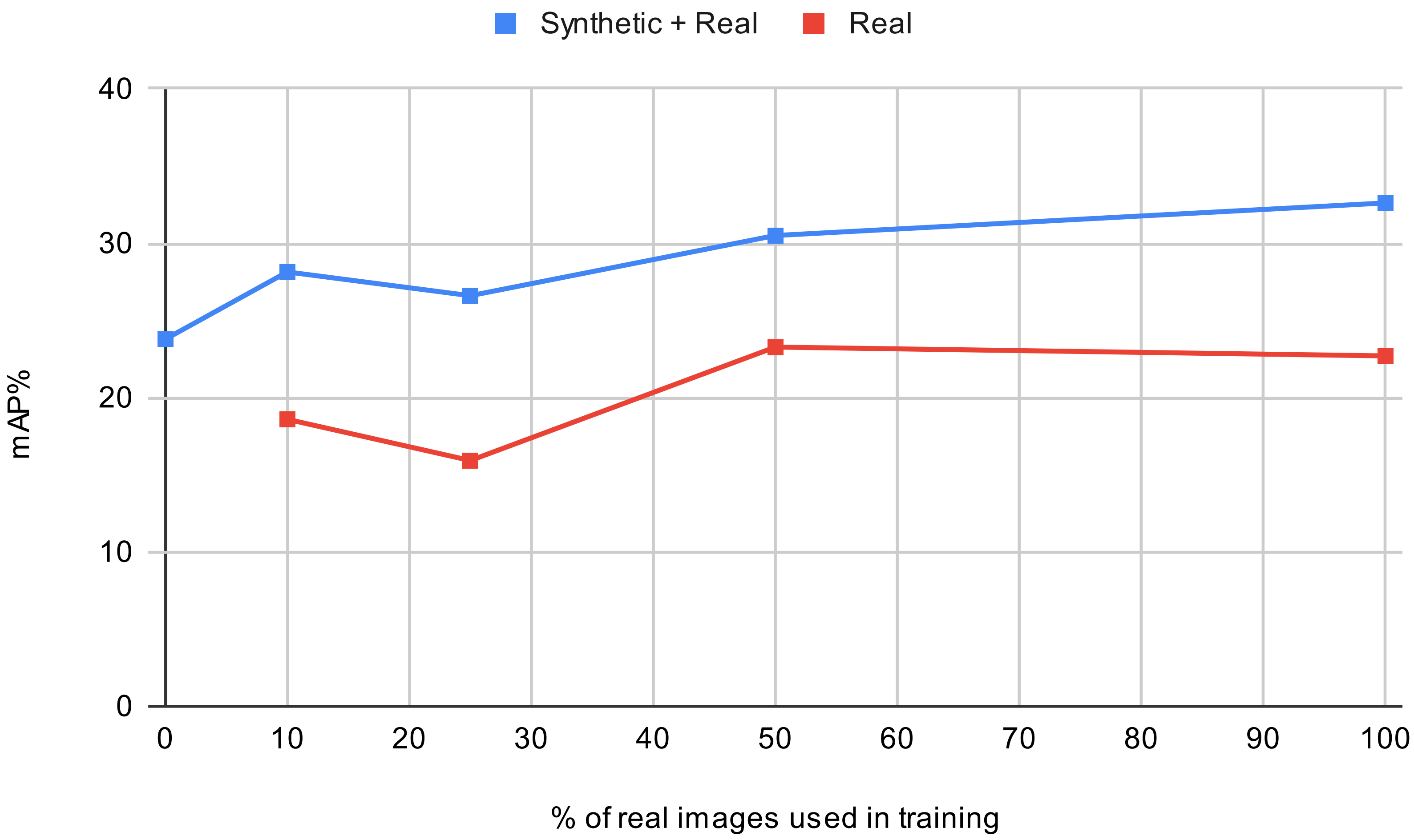
DATA HERE -> https://iplab.dmi.unict.it/EHOI_SYNTH/

IN THIS CONFERENCE

ID 221: Poster

Tomorrow 15.00 – 16.00

Can simulated data help?



R. Leonardi, F. Ragusa, A. Furnari, G. M. Farinella (2022). Egocentric Human-Object Interaction Detection Exploiting Synthetic Data. In 21st International Conference on Image Analysis and Processing.

Understanding Actions

- Recognizing and detecting the actions performed by user allows to understand what happens in the the video;
- This can be useful to:
  - Segment the video into coherent temporal units for:
    - Summarization;
    - Video understanding;
  - Understand the user’s goals to assist them;

Relation between Action and Interaction

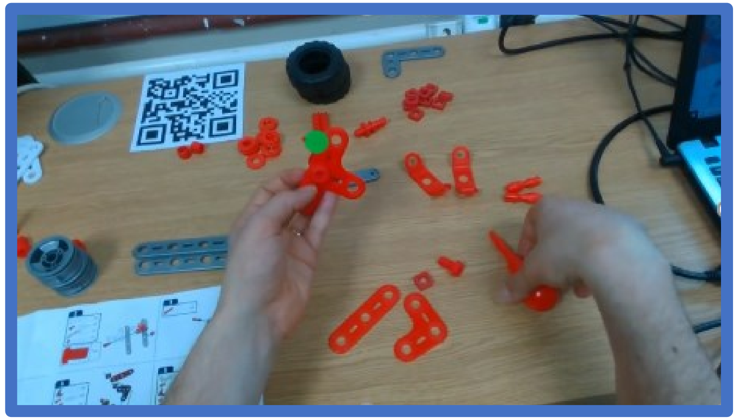
TAKE SCREWDRIVER

Relation between Action and Interaction

TAKE SCREWDRIVER

Start Action

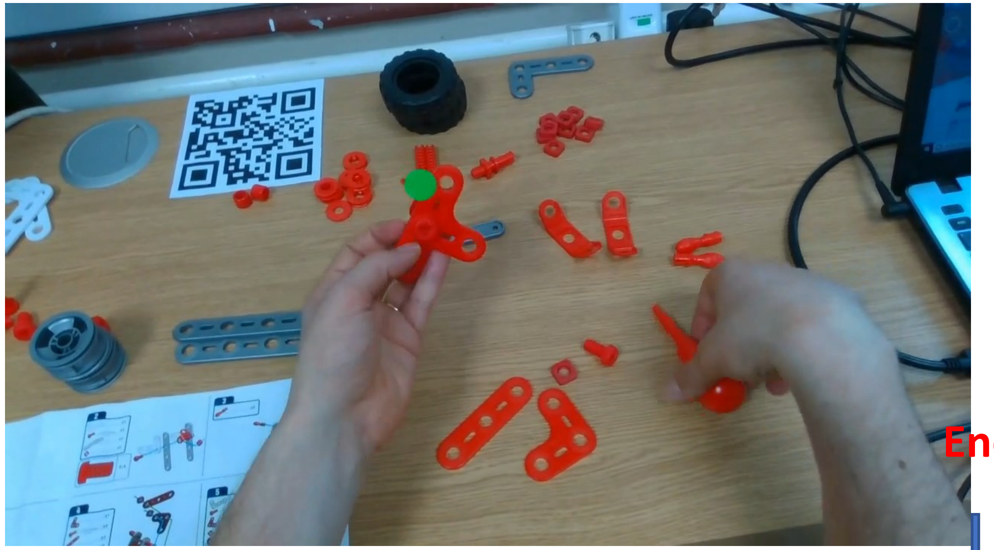
Start Interaction (H-O)



Frame of Contact

Relation between Action and Interaction

TAKE SCREWDRIVER

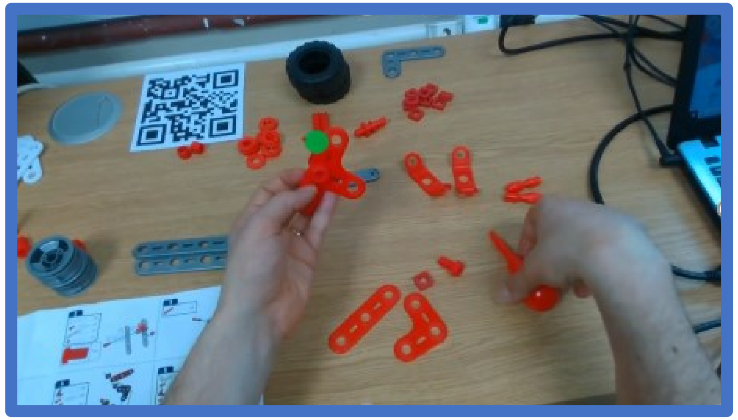


End Interaction

Start Action

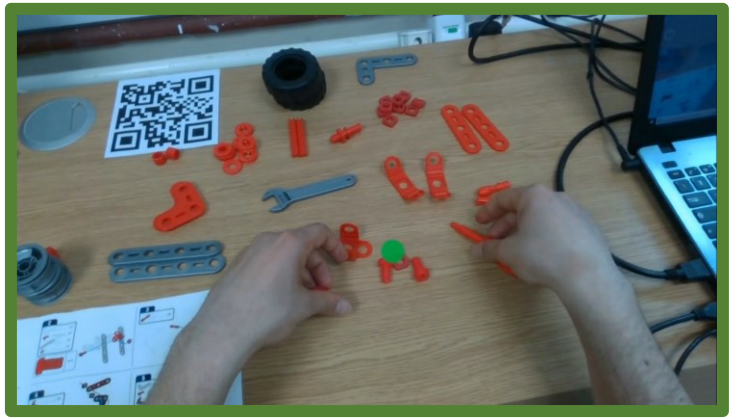
Start Interaction (H-O)

End Action



Frame of Contact

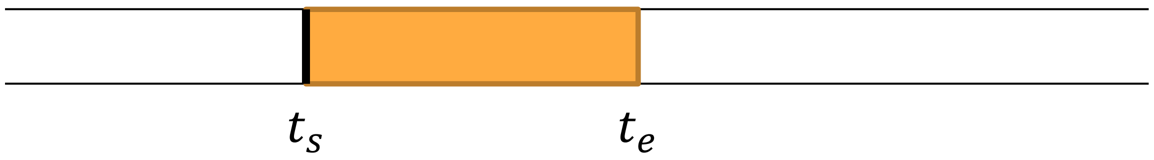
Frame of Decontact

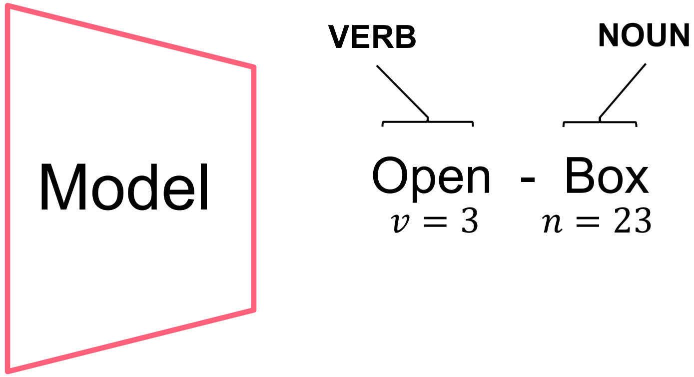


Relation between Action and Interaction

Relation	Verbs	MECCANO verbs
A_s → I_s → I_e → A_e	pat, hit, kick	//
A_s → I_s → A_e → I_e	pick up	take, fit, align, plug, pull
A_s → I_s → A_e, I_e	close, open, turn on, press, push	browse
A_s → A_e	walk, jump, run	//
I_s → A_s → A_e → I_e	wring out, wash, cut, mix	pull
I_s → A_s → I_e → A_e	throw, leave, place	put
I_s → A_s → I_e, A_e	move	browse
I_s, A_s → A_e → I_e	twist, rip	screw, unscrew, tighten, loosen
I_s, A_s → I_e, A_e	stretch, knead, write, watch	check

Action Recognition: Task



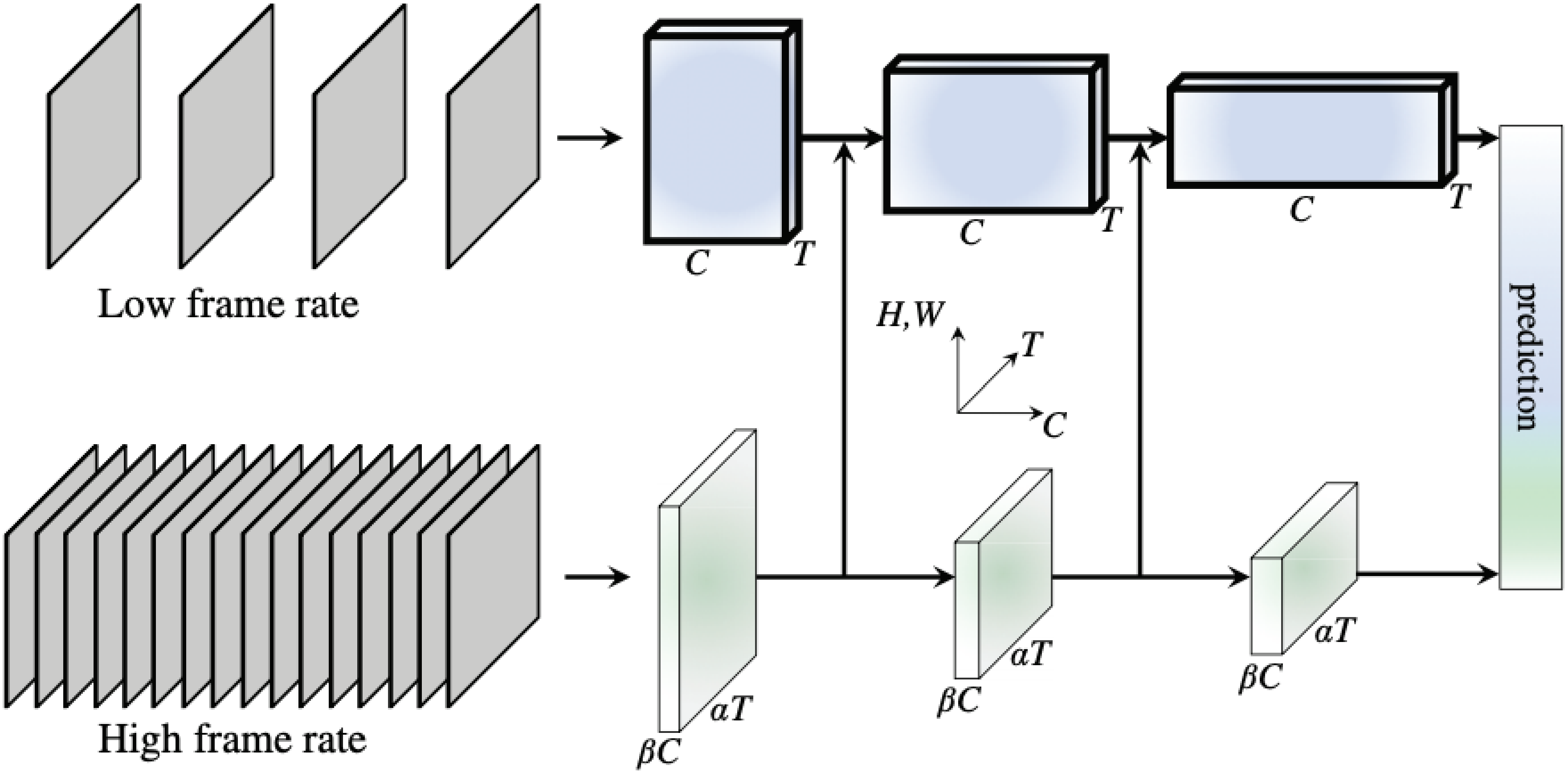


"observe a trimmed segment denoted by start and end time and classify the action present in the clip"

As defined in EPIC-KITCHENS-2020

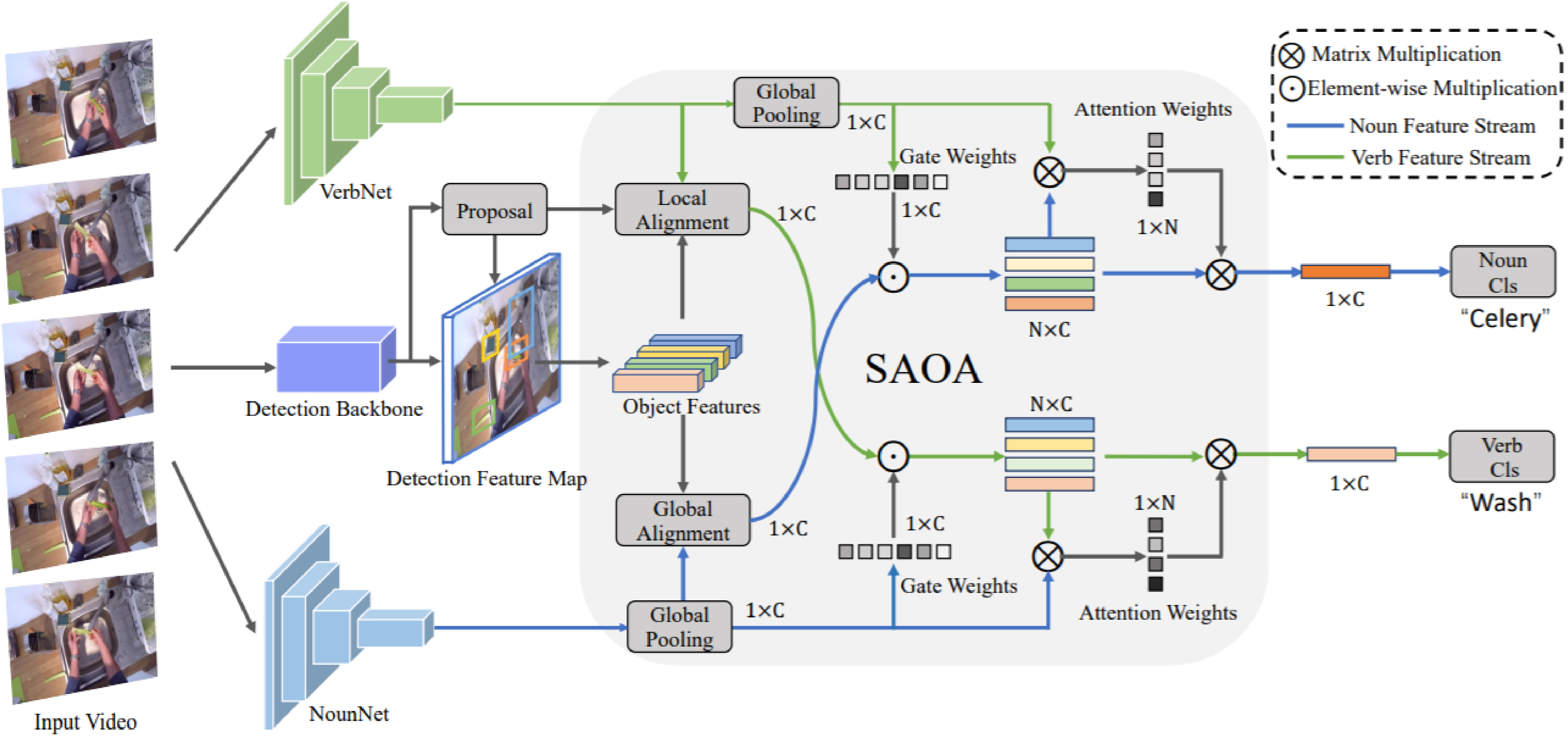
CODE HERE -> https://github.com/facebookresearch/SlowFast

SlowFast Networks for Video Recognition



Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 6202-6211).

Object-centric Egocentric Action Recognition



Wang, X., Zhu, L., Wu, Y., & Yang, Y. (2020). Symbiotic attention for egocentric action recognition with object-centric alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Personal assistants and Future Predictions

Intelligent assistants should be able to understand what are the user’s goals and what is going to happen in the future.

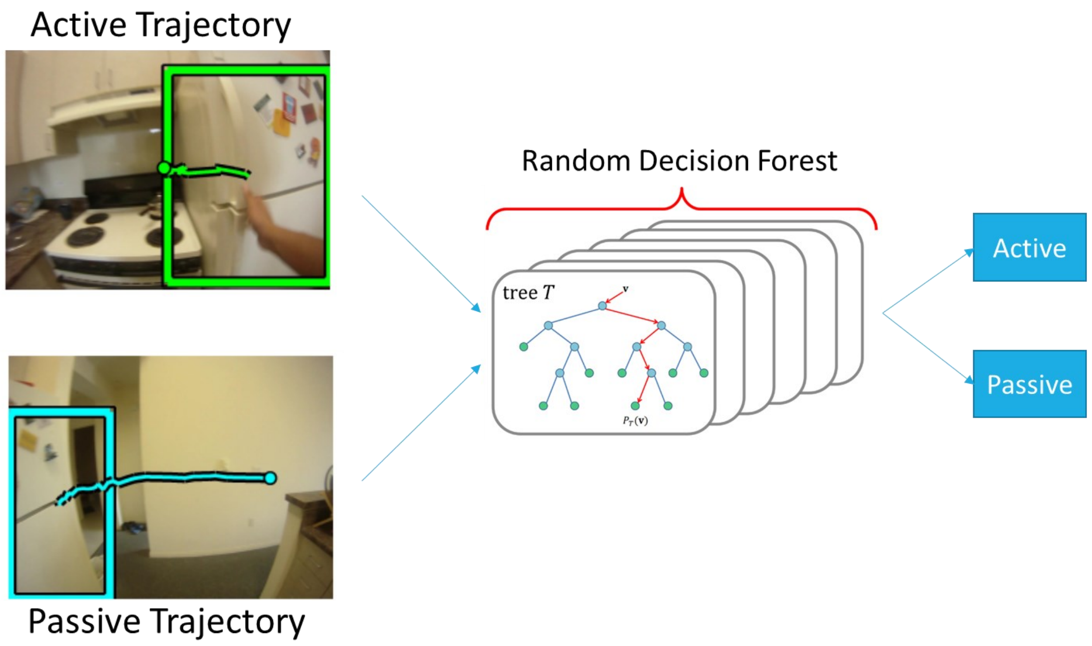
Next-active-object: **LOCKER**

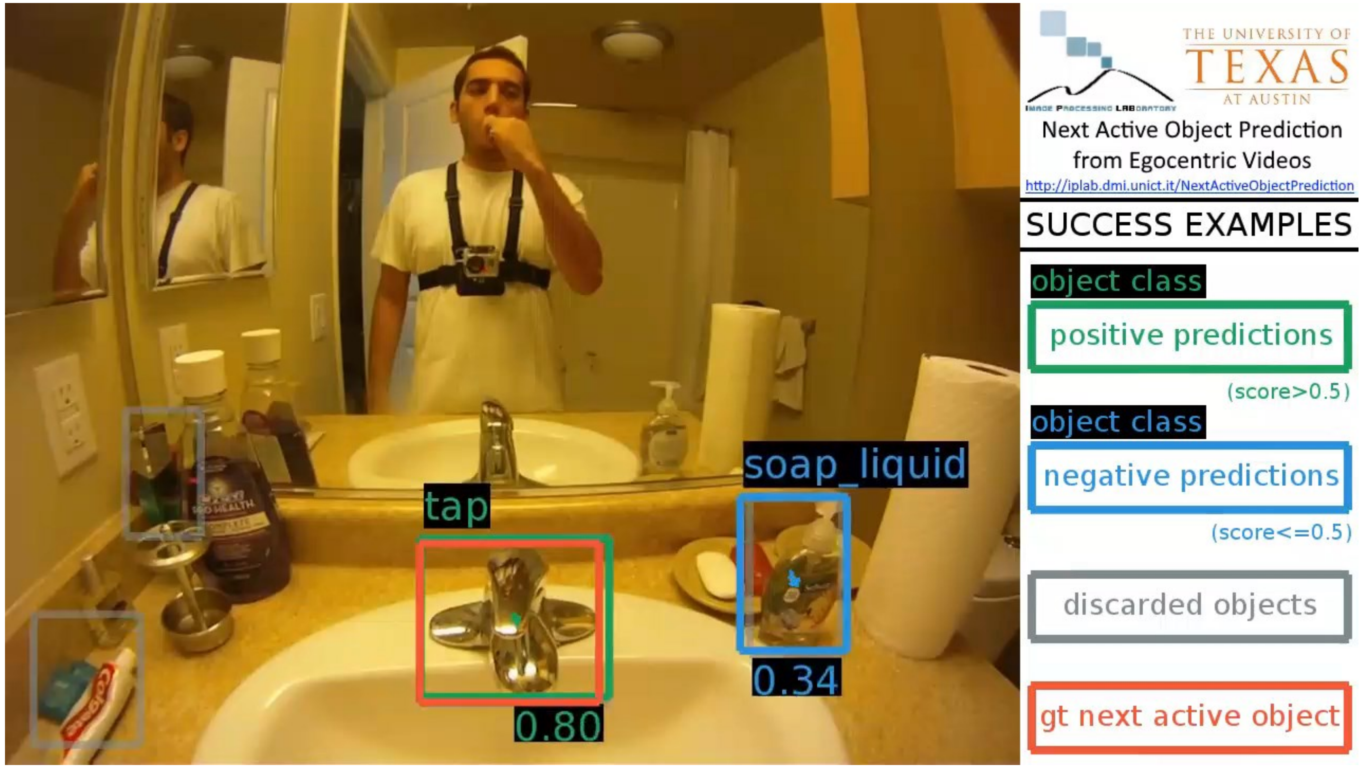
Next action: **OPEN LOCKER**

http://iplab.dmi.unict.it/NextActiveObjectPrediction/

Anticipation – Next-Active-Objects

Use egocentric object trajectories to distinguish passive from next-active-objects (i.e., those which will be used soon by the user).





A. Furnari, S. Battiato, K. Grauman, G. M. Farinella, Next-Active-Object Prediction from Egocentric Videos, Journal of Visual Communication and Image Representation, 2017

Action Anticipation Task - Definition

(observed video)

Model

Take - Plate

(unobserved)

$t_s - \tau_a - \tau_o$ $t_s - \tau_a$ t_s t_e

τ_o arbitrary $\tau_a = 1s;$

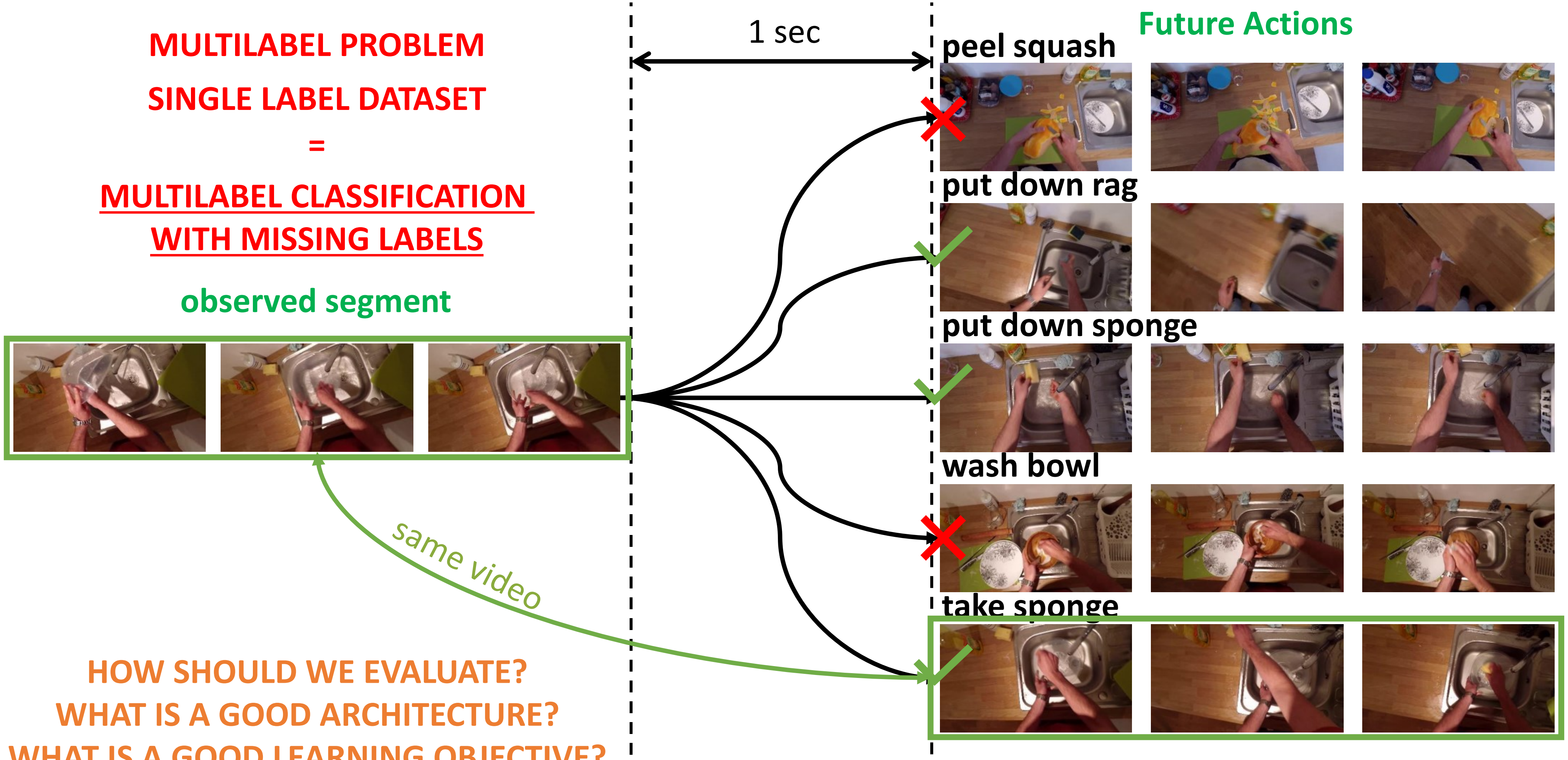
Action Anticipation

MULTILABEL PROBLEM

SINGLE LABEL DATASET

=

MULTILABEL CLASSIFICATION WITH MISSING LABELS

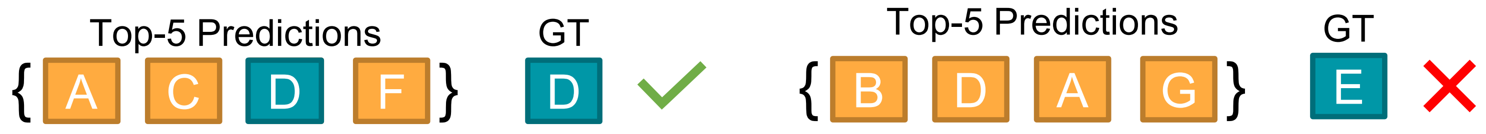


HOW SHOULD WE EVALUATE?
WHAT IS A GOOD ARCHITECTURE?
WHAT IS A GOOD LEARNING OBJECTIVE?

A. Furnari, S. Battiato, G. M. Farinella, Leveraging Uncertainty to Rethink Loss Functions and Evaluation Measures for Egocentric Action Anticipation, EPIC Workshop in conjunction with ECCV, 2018

Mean Top-5 Recall

- Since multiple futures are possible, TOP-1 accuracy is not a good measure;
- TOP-5 Accuracy still suffers from a major problem due to the unbalanced nature of the dataset:
  - A high Top-5 accuarcy is possible by ranking higher the most frequent classes;
- To overcome this issue, we introduced Mean Top-5 Recall:
  - An action anticipation is correct if the ground truth label is among the Top-5 predicted actions:



  - Results are computed **per-class** then averaged to obtain a single measure.

A. Furnari, S. Battiato, G. M. Farinella, Leveraging Uncertainty to Rethink Loss Functions and Evaluation Measures for Egocentric Action Anticipation, EPIC Workshop in conjunction with ECCV, 2018

Action Anticipation: Encoding vs Inference

what will happen in 1 second?

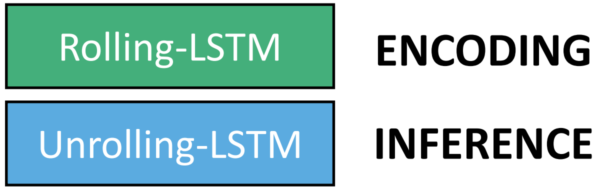
ENCODING { took container, opened tap, filled container } → «put down container» } **INFERENCE**

The two tasks are usually performed jointly

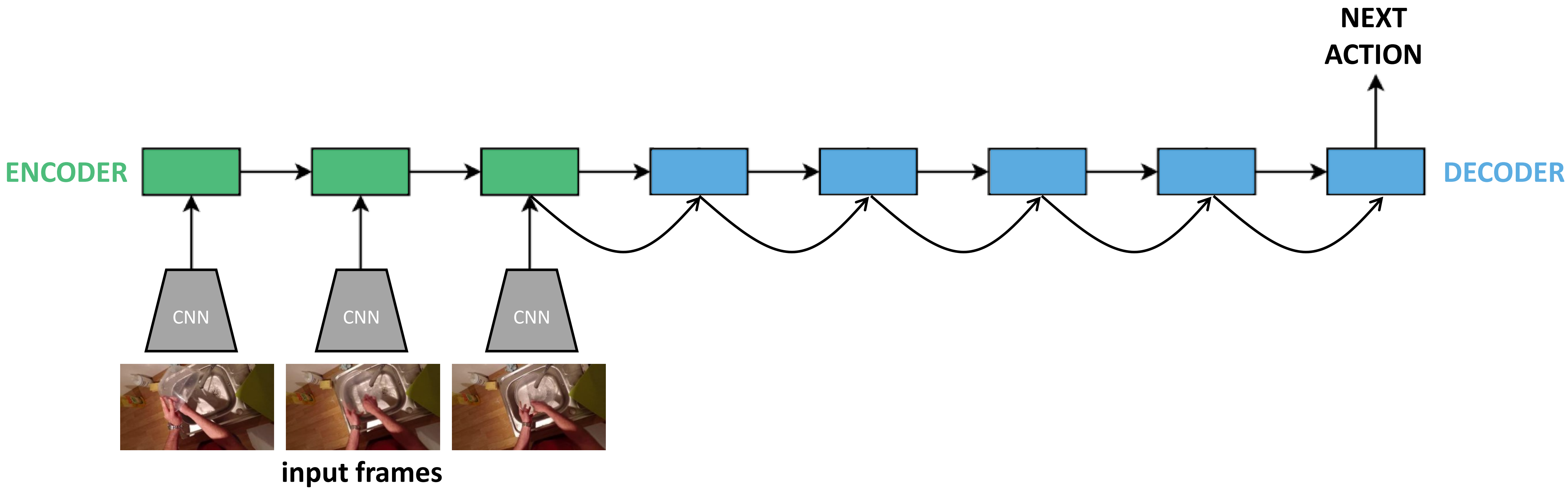
We propose to disentangle them by using <u>two separate LSTMs</u>

A. Furnari, G. M. Farinella, What Would You Expect? Anticipating Egocentric Actions with Rolling-Unrolling LSTMs and Modality Attention. ICCV 2019 (ORAL).
A. Furnari, G. M. Farinella. Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video. TPAMI 2020. http://iplab.dmi.unict.it/rulstm

Sequence to Sequence Models



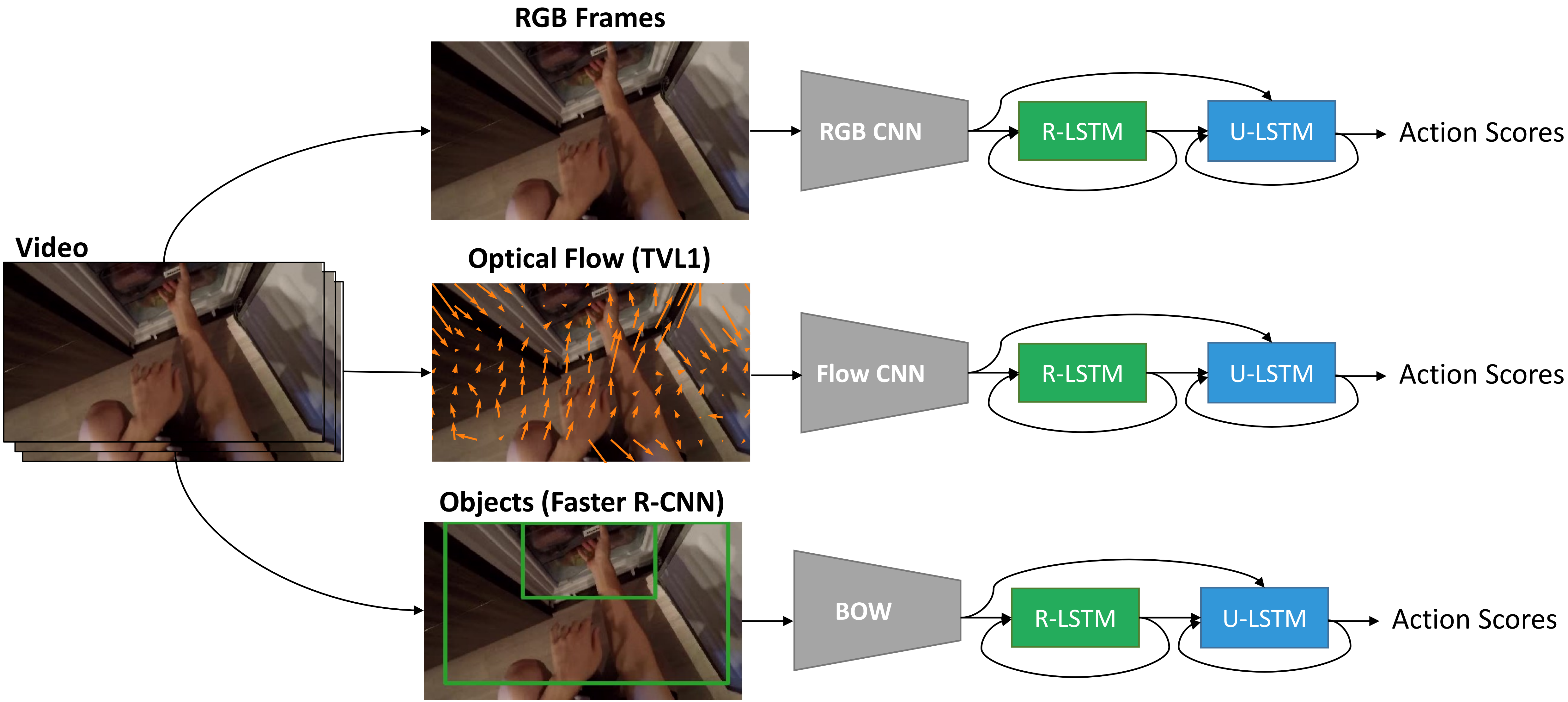
We take inspiration from sequence to sequence models.



A. Furnari, G. M. Farinella, What Would You Expect? Anticipating Egocentric Actions with Rolling-Unrolling LSTMs and Modality Attention. ICCV 2019 (ORAL).
A. Furnari, G. M. Farinella. Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video. TPAMI 2020. http://iplab.dmi.unict.it/rulstm

Three Modalities

How to fuse?

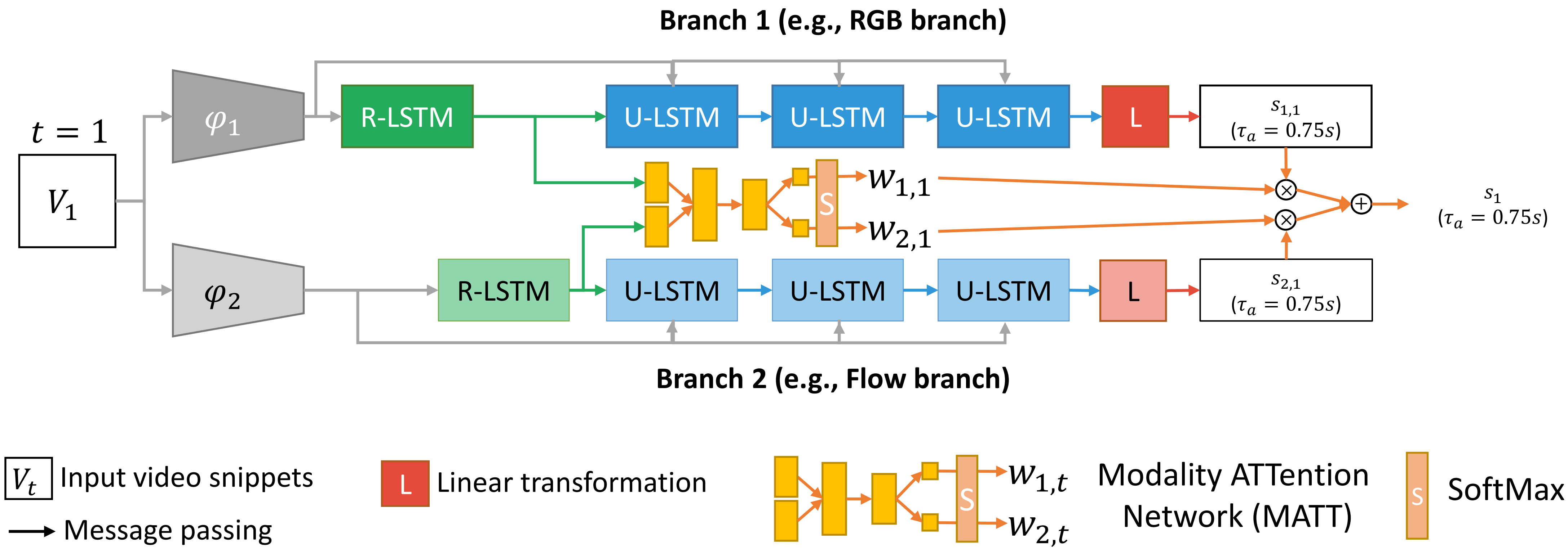


A. Furnari, G. M. Farinella, What Would You Expect? Anticipating Egocentric Actions with Rolling-Unrolling LSTMs and Modality Attention. ICCV 2019 (ORAL).

A. Furnari, G. M. Farinella. Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video. TPAMI 2020. http://iplab.dmi.unict.it/rulstm

Modality ATTention (MATT)

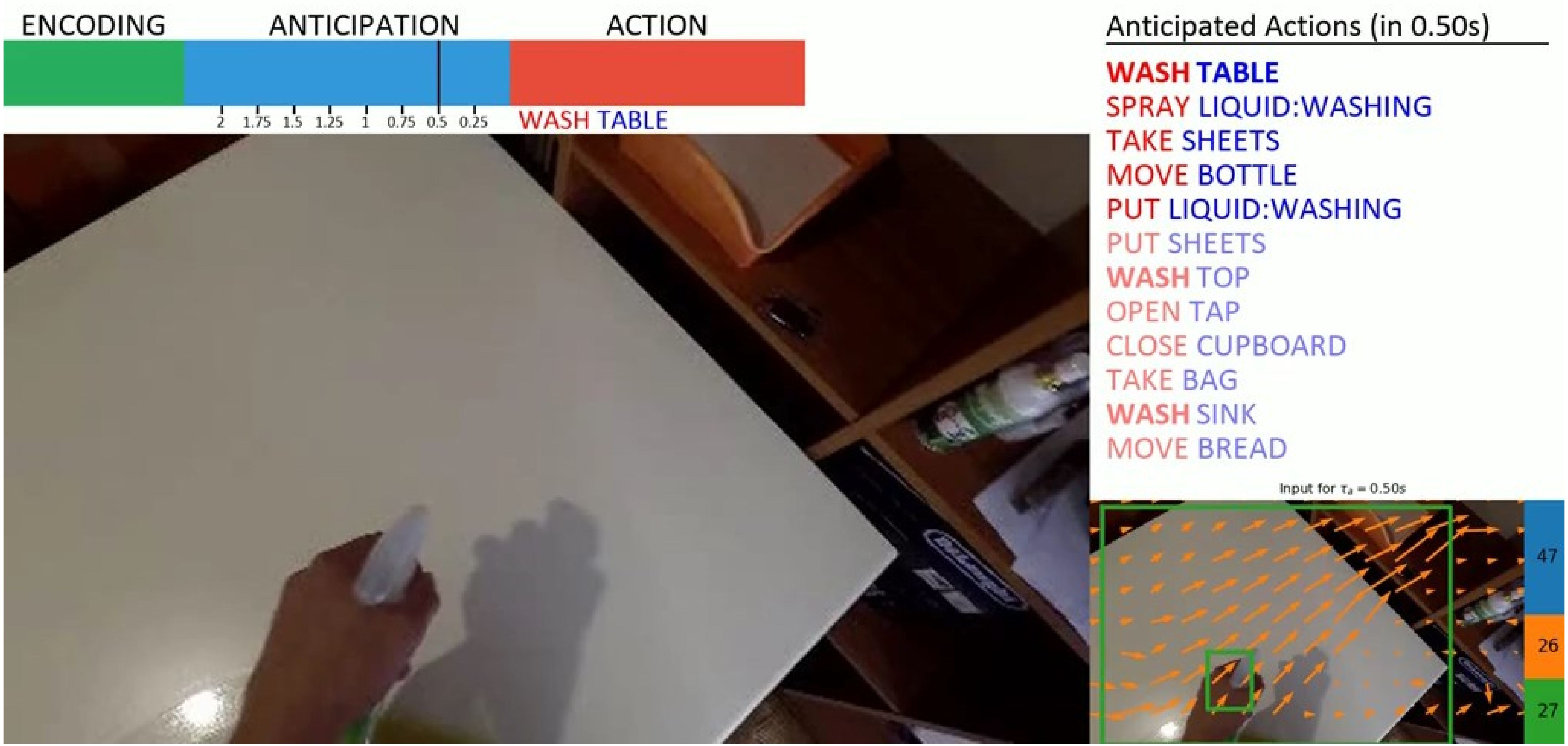
The relative importance of each modality may depend on the observed sample.



A. Furnari, G. M. Farinella, What Would You Expect? Anticipating Egocentric Actions with Rolling-Unrolling LSTMs and Modality Attention. ICCV 2019 (ORAL).

A. Furnari, G. M. Farinella. Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video. TPAMI 2020. http://iplab.dmi.unict.it/rulstm

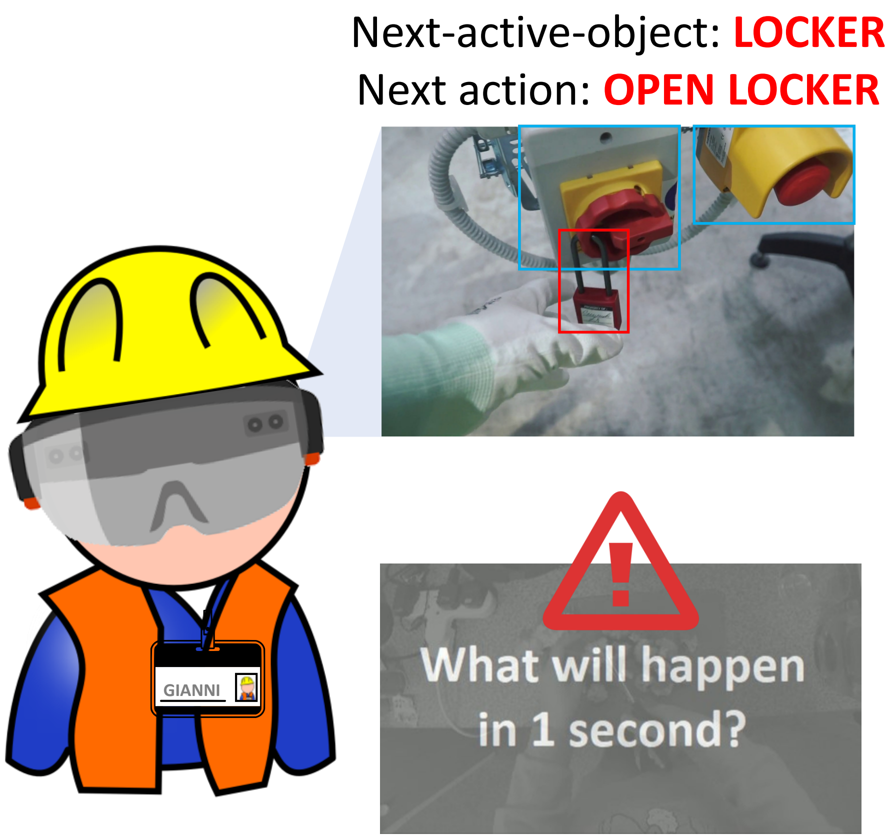
Demo Video: Egocentric Action Anticipation



A. Furnari, G. M. Farinella, What Would You Expect? Anticipating Egocentric Actions with Rolling-Unrolling LSTMs and Modality Attention. ICCV 2019 (ORAL).

A. Furnari, G. M. Farinella. Rolling-Unrolling LSTMs for Action Anticipation from First-Person Video. TPAMI 2020. http://iplab.dmi.unict.it/rulstm

Can we bring egocentric vision to industry?



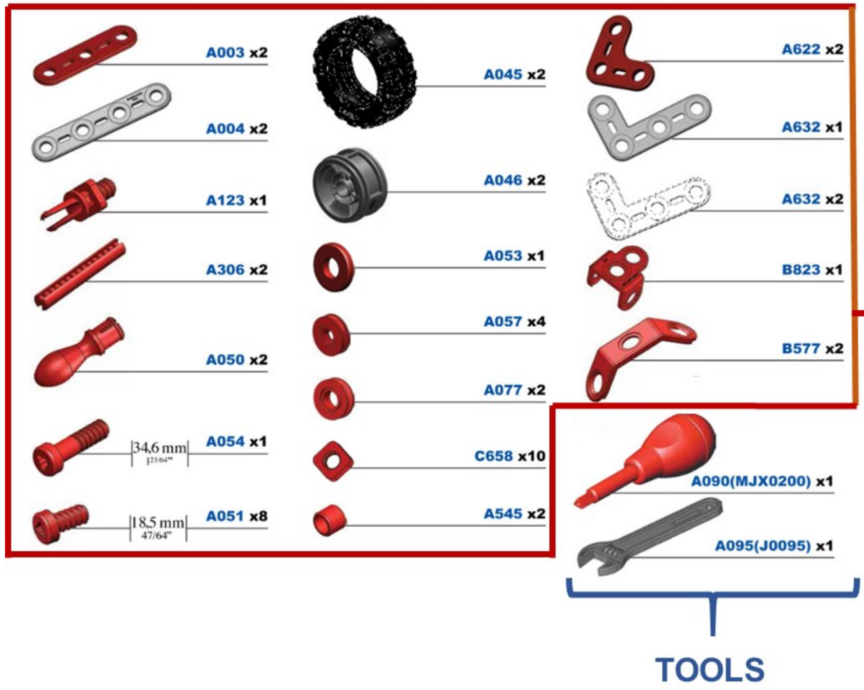
- The factory is a natural place for a wearable assistant;
- Closed-world assumption;
- Current research has considered different scenarios;
- No datasets in industrial-like scenarios;

The MECCANO Dataset

Data HERE -> https://iplab.dmi.unict.it/MECCANO/

We asked subjects to record egocentric videos while assembling a toy motorbike.

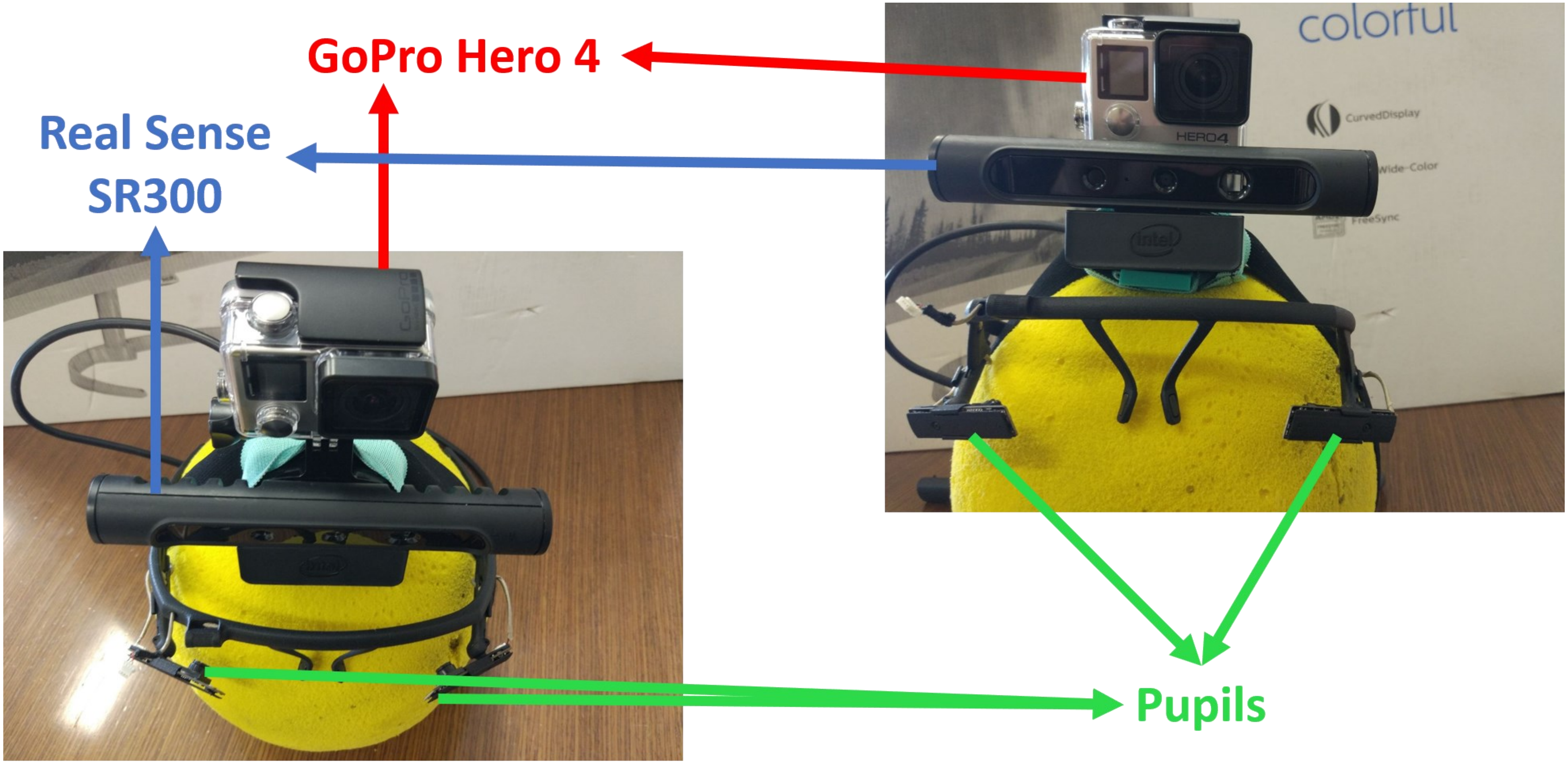
The assembly required to interact with several parts and two tools.



The scenario is industrial-like, with subjects undertaking interactions with tiny objects and tools in a sequential fashion to reach a goal.

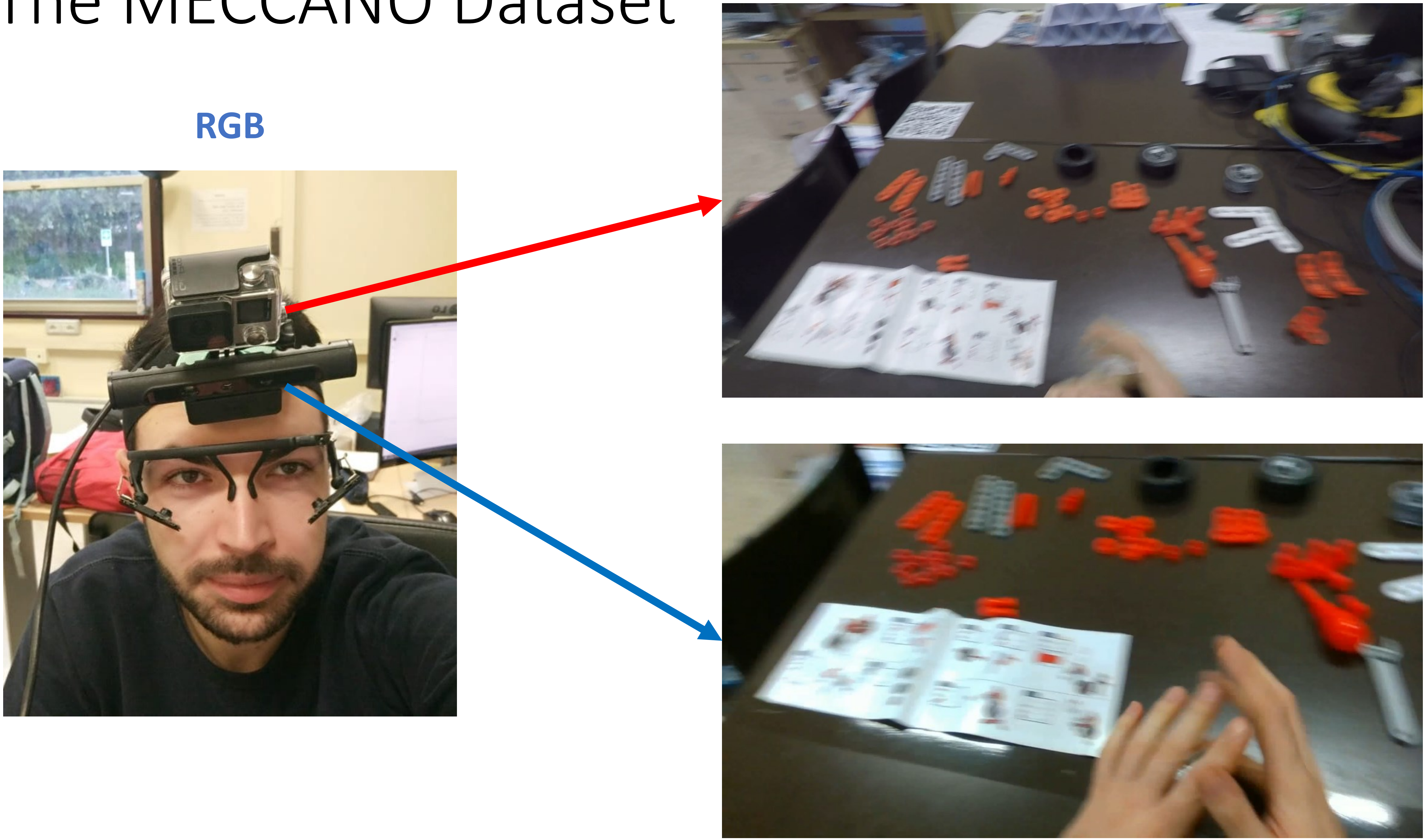
F. Ragusa, A. Furnari, S. Livatino, G. M. Farinella. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. WACV, 2021 (https://arxiv.org/abs/2010.05654). ORAL.

Data Collection

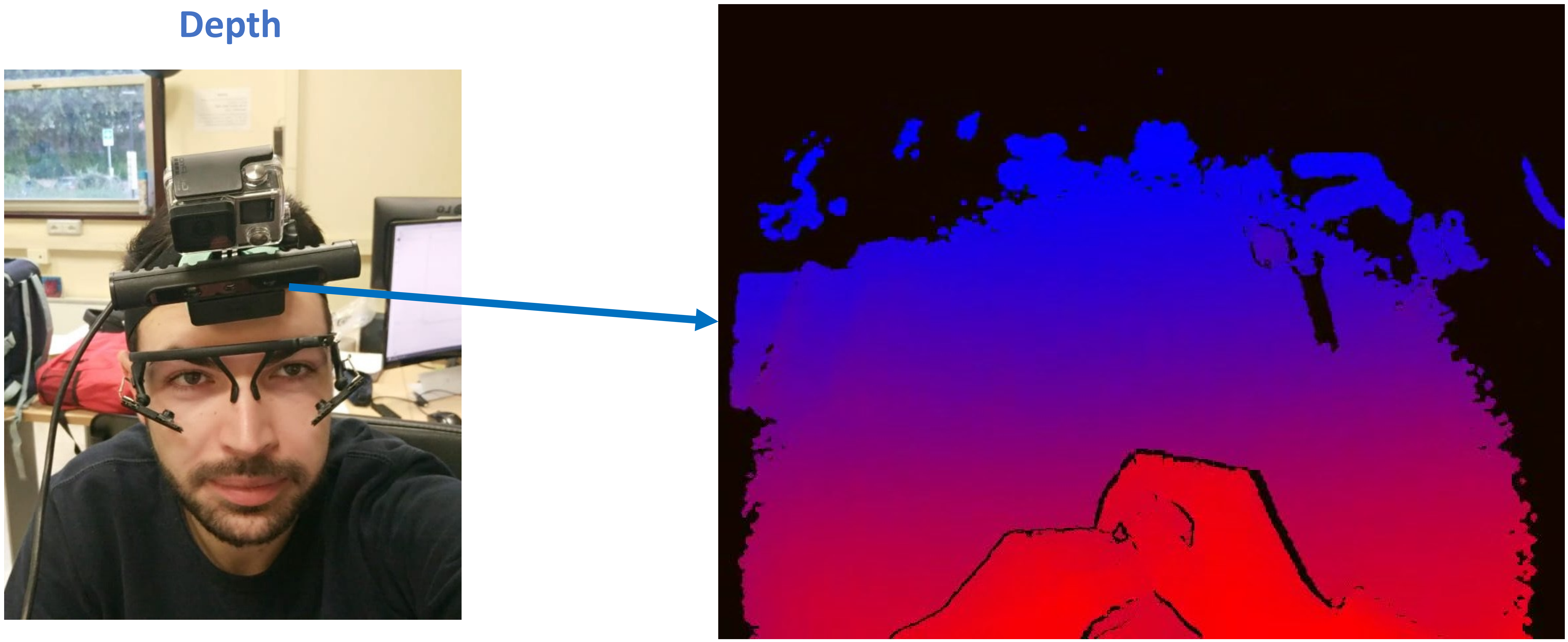


The MECCANO Dataset

RGB

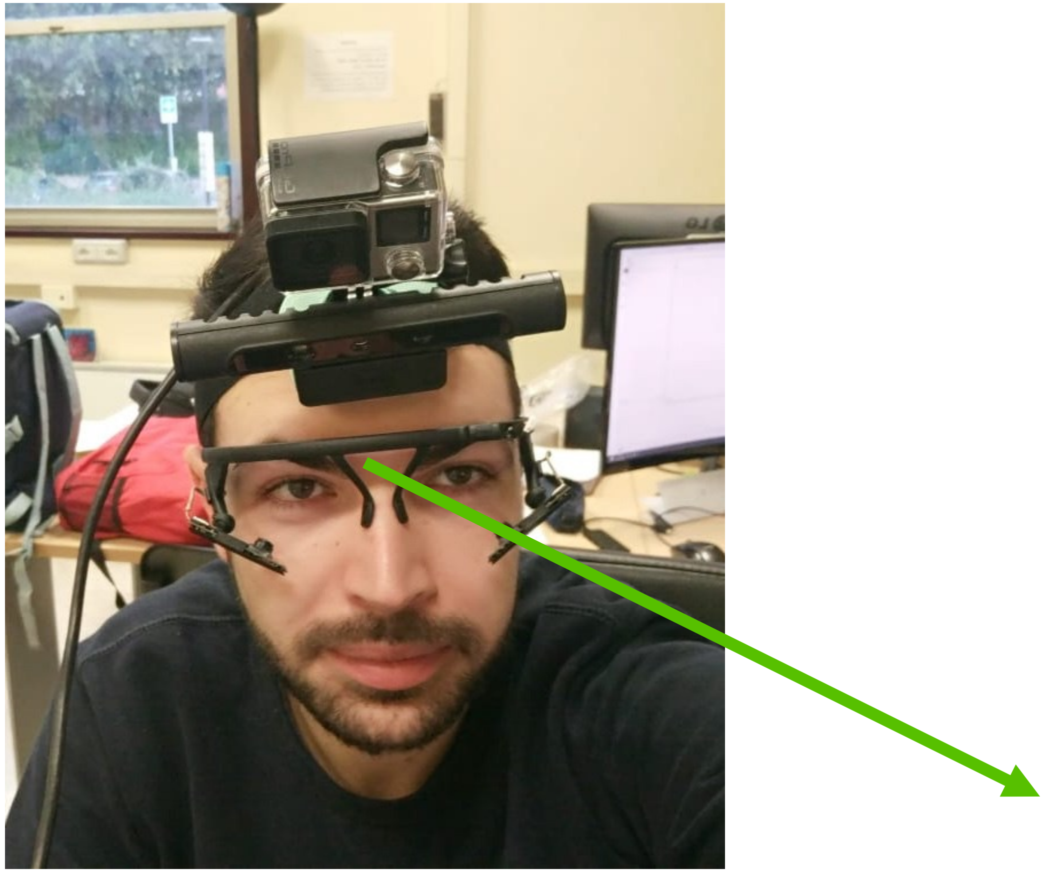


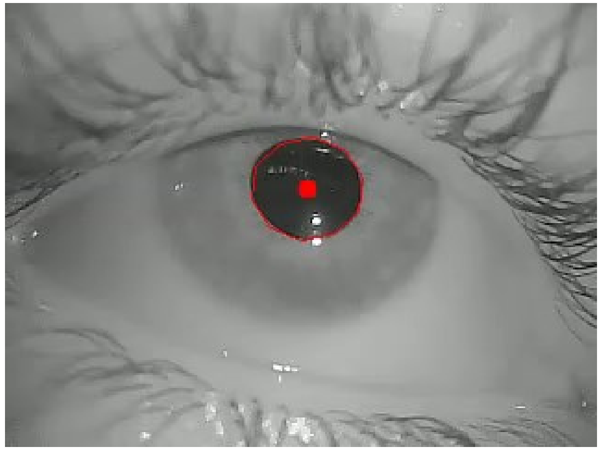
The MECCANO Dataset

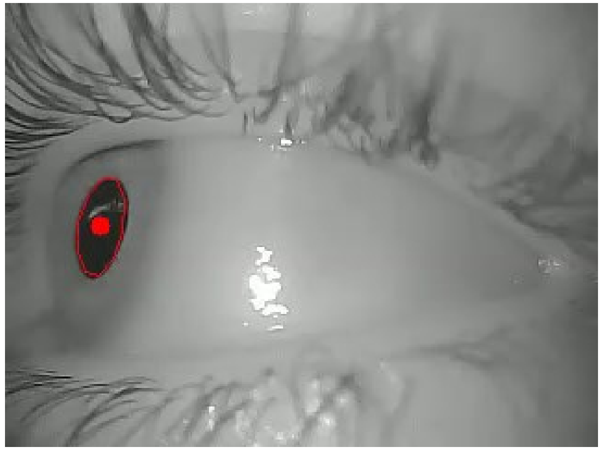


The MECCANO Dataset

Gaze







The MECCANO Dataset: Statistics

20 Subjects

3 Modalities

20 min. avg. Video length

5 Tasks

8858 Segments

64349 Boxes

20 Objects

12 Verbs

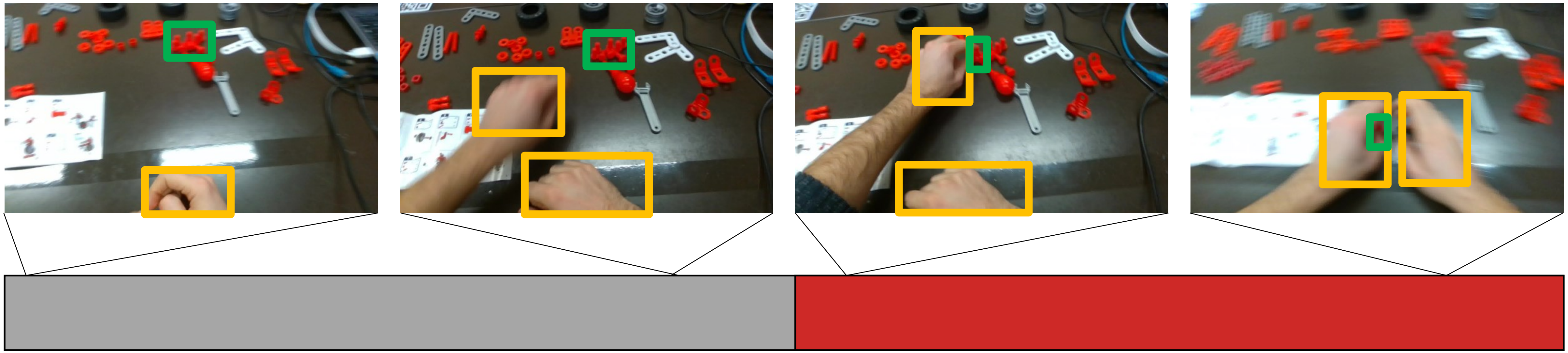
61 Actions

F. Ragusa, A. Furnari, S. Livatino, G. M. Farinella. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. WACV, 2021 (https://arxiv.org/abs/2010.05654). ORAL.

The MECCANO Dataset: Hands and Future Objects

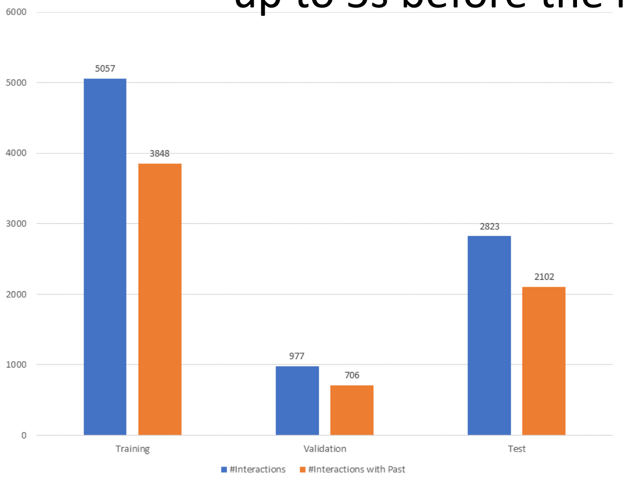
Hands + Next Active Objects

Action Boundaries + Active Objects + Hands

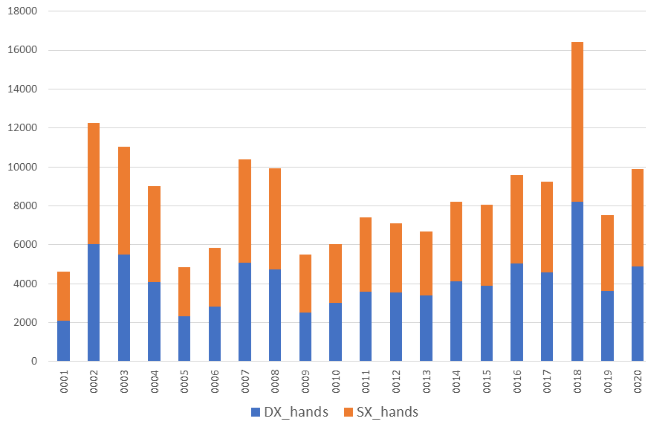


up to 5s before the interaction

Human-Object Interaction



Video	Interactions	Interactions with past
0001	319	257
0002	586	452
0003	573	429
0004	485	372
0005	251	200
0006	307	234
0007	493	367
0008	550	384
0009	289	289
0010	304	194
0011	400	310
0012	384	258
0013	313	244
0014	434	297
0015	425	324
0016	576	436
0017	484	339
0018	788	603
0019	400	294
0020	496	373
Total	8857	6656



F. Ragusa, A. Furnari, S. Livatino, G. M. Farinella. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. WACV, 2021 (https://arxiv.org/abs/2010.05654). ORAL.

The MECCANO Dataset: Tasks

1) Action Recognition

2) Active Object Detection and Recognition

3) EHOI Interaction Detection

PRESENT

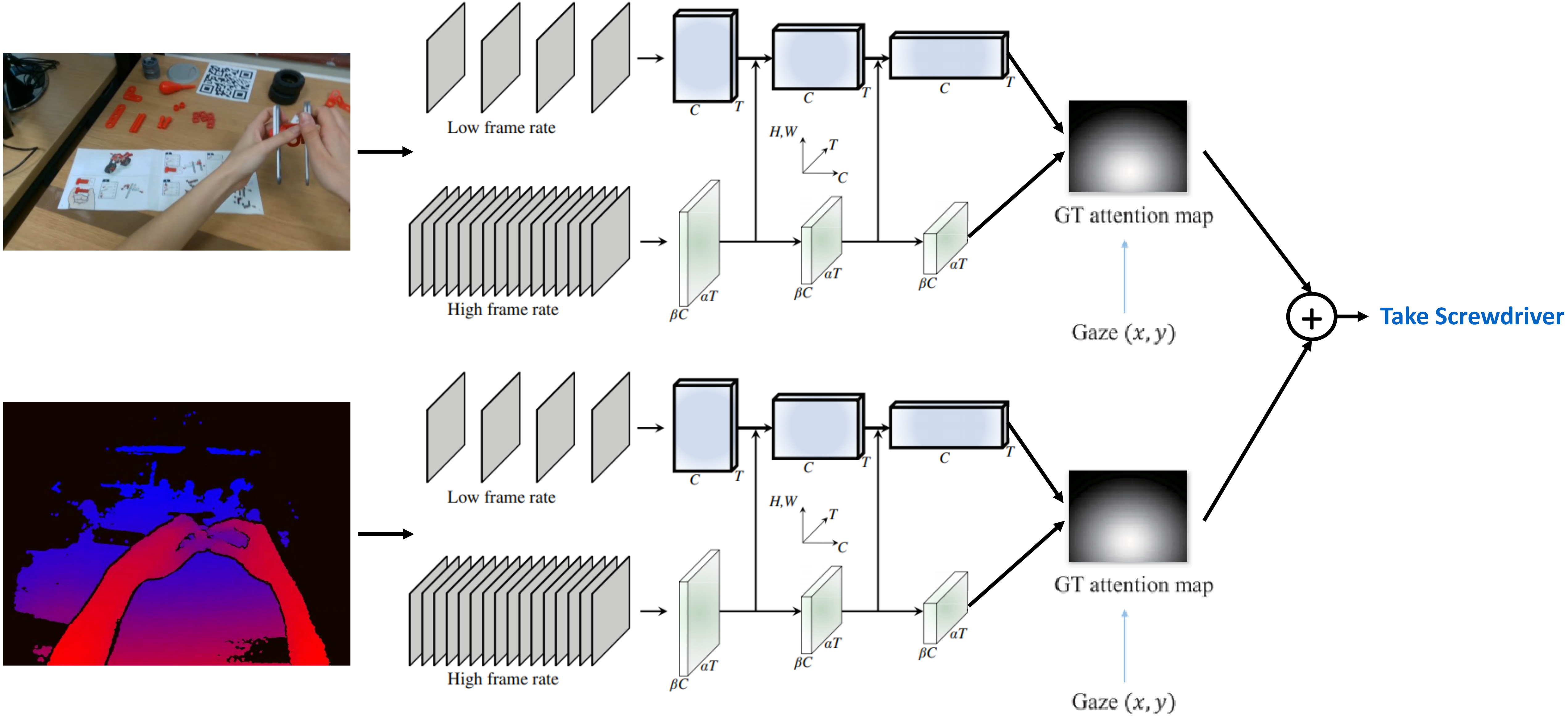
4) Action Anticipation

5) Next-Active Object Detection

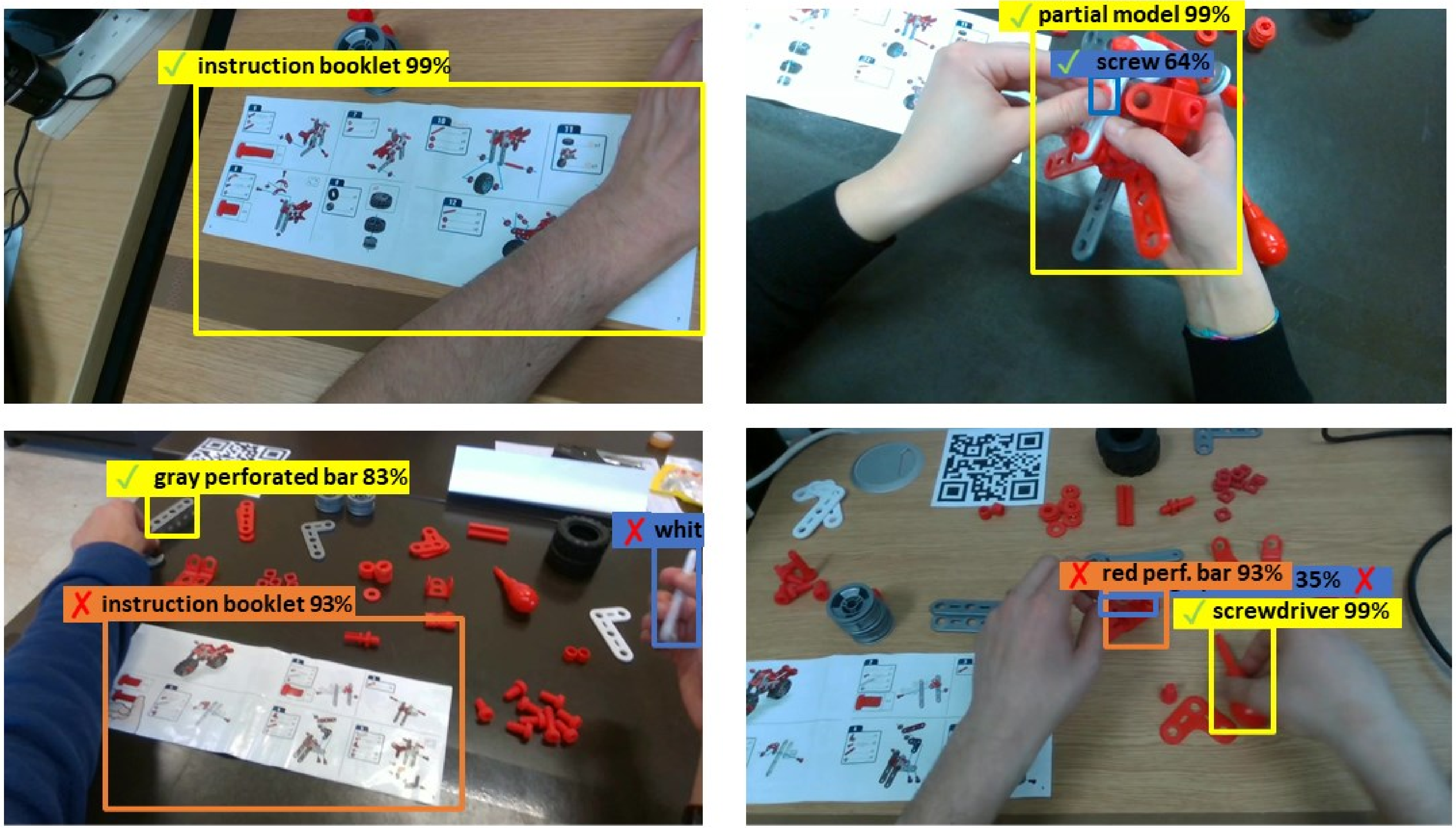
FUTURE

F. Ragusa, A. Furnari, S. Livatino, G. M. Farinella. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. WACV, 2021 (https://arxiv.org/abs/2010.05654). ORAL.

Action Recognition

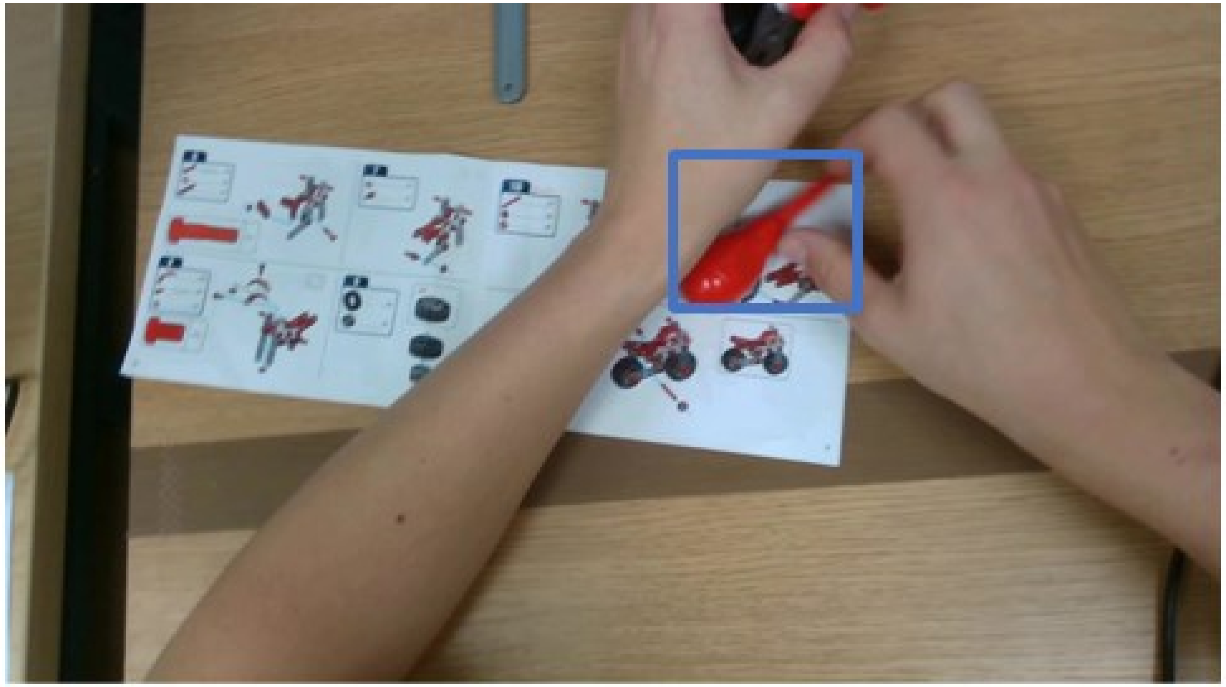


Active Object Detection and Recognition

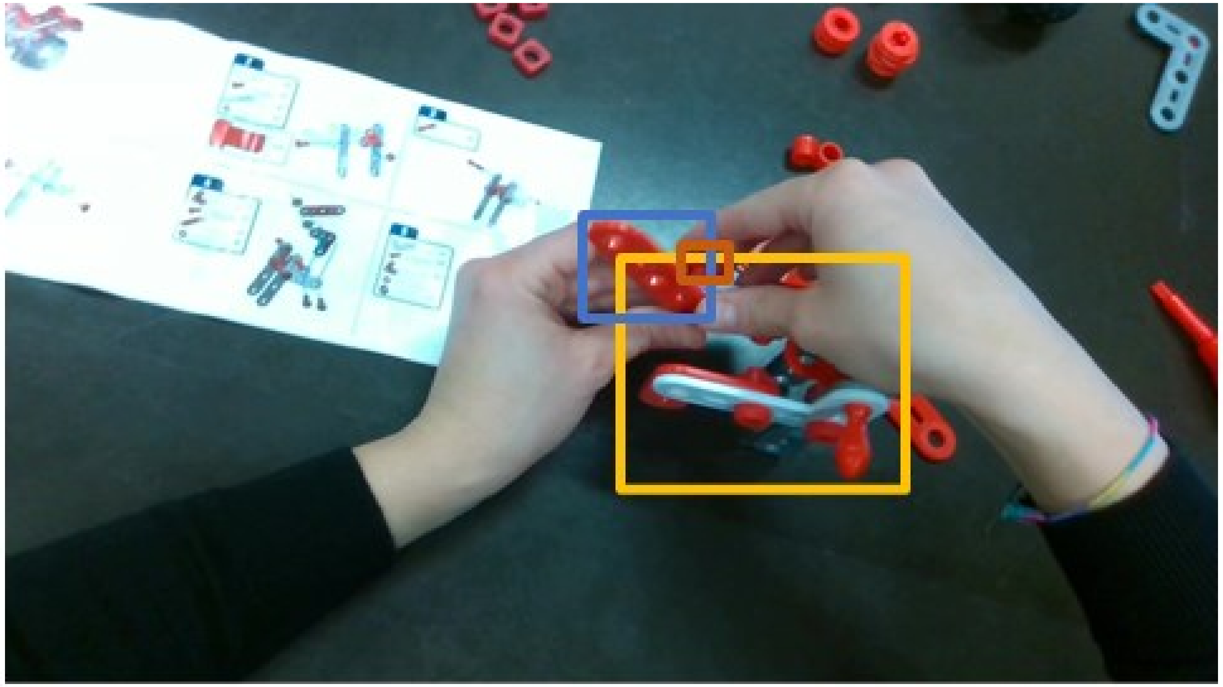


F. Ragusa, A. Furnari, S. Livatino, G. M. Farinella. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. In IEEE Winter Conference on Application of Computer Vision (WACV), 2021. **ORAL**

EHOI Detection



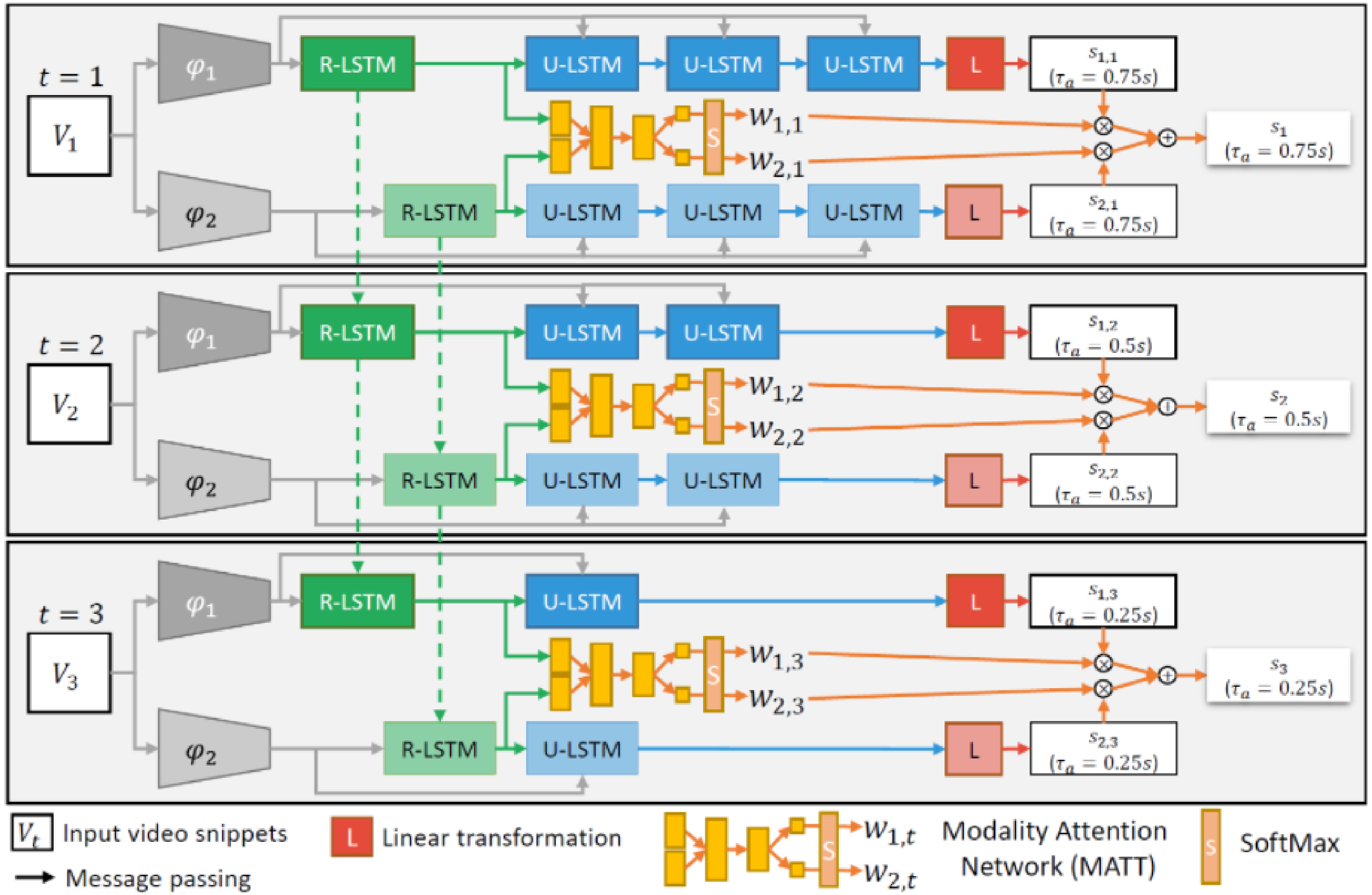
<put, screwdriver>



<plug, {red_perforated_bar, screw, partial_model }>

F. Ragusa, A. Furnari, S. Livatino, G. M. Farinella. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. In IEEE Winter Conference on Application of Computer Vision (WACV), 2021. **ORAL**

Action Anticipation



Modalities:

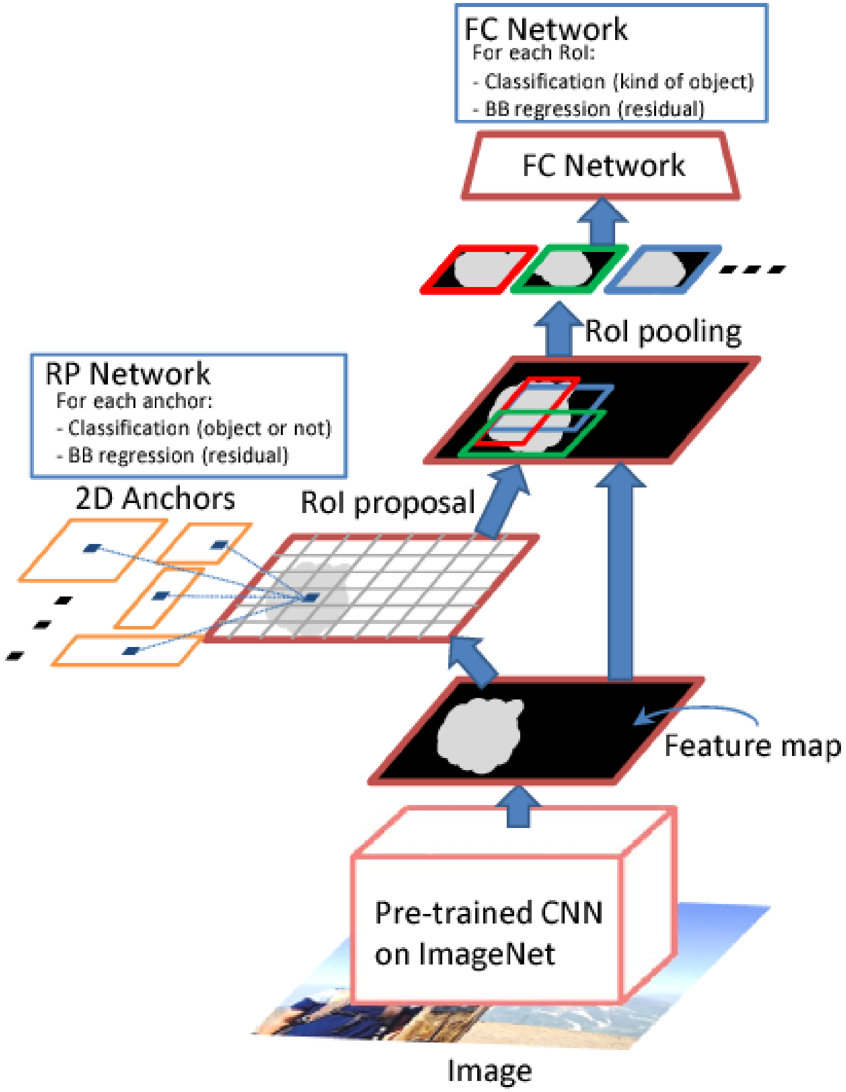
- RGB
- Optical Flow
- Objects

Our Modalities:

- RGB + Flow
- Depth
- Objects
- Hands
- Gaze

F. Ragusa, A. Furnari, S. Livatino, G. M. Farinella. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. In IEEE Winter Conference on Application of Computer Vision (WACV), 2021. **ORAL**

Next-Active Objects Detection



F. Ragusa, A. Furnari, S. Livatino, G. M. Farinella. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. In IEEE Winter Conference on Application of Computer Vision (WACV), 2021. **ORAL**

NEXT VISION

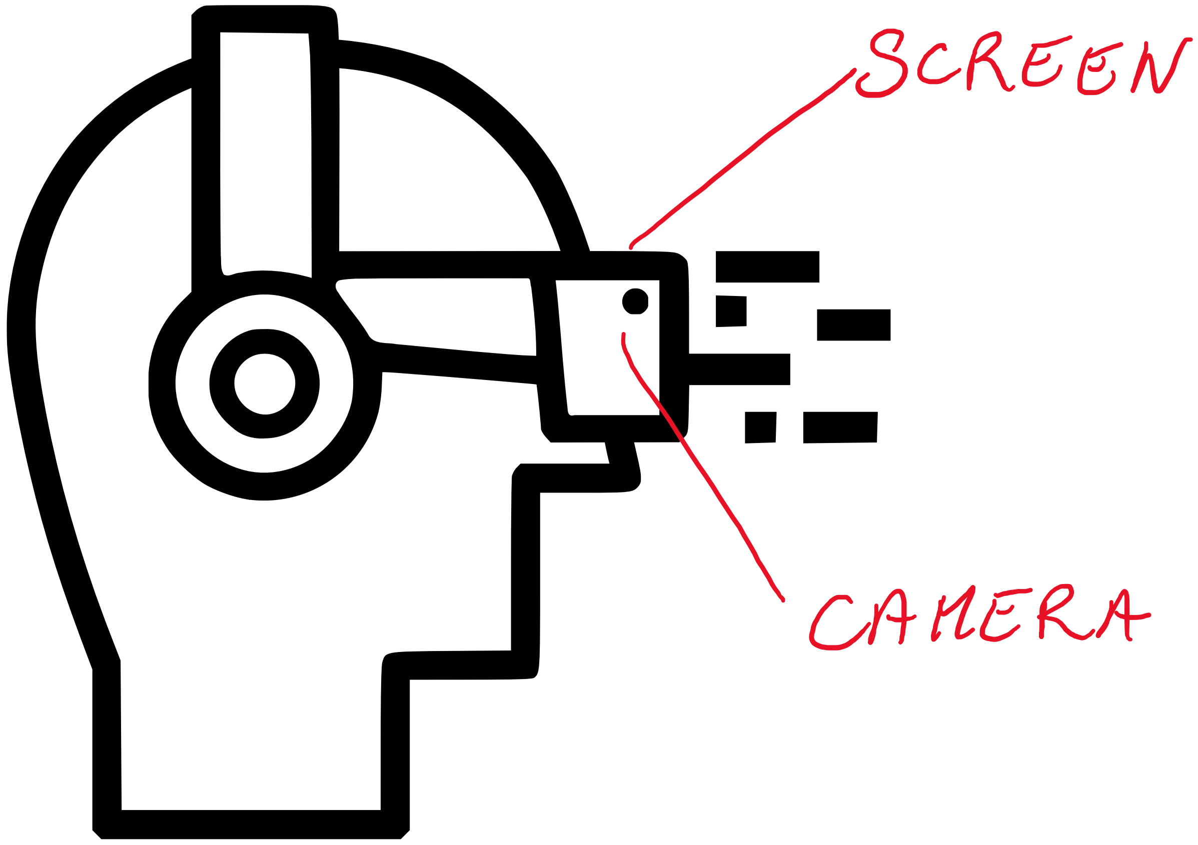
Spin-off of the University of Catania

https://www.nextvisionlab.it/

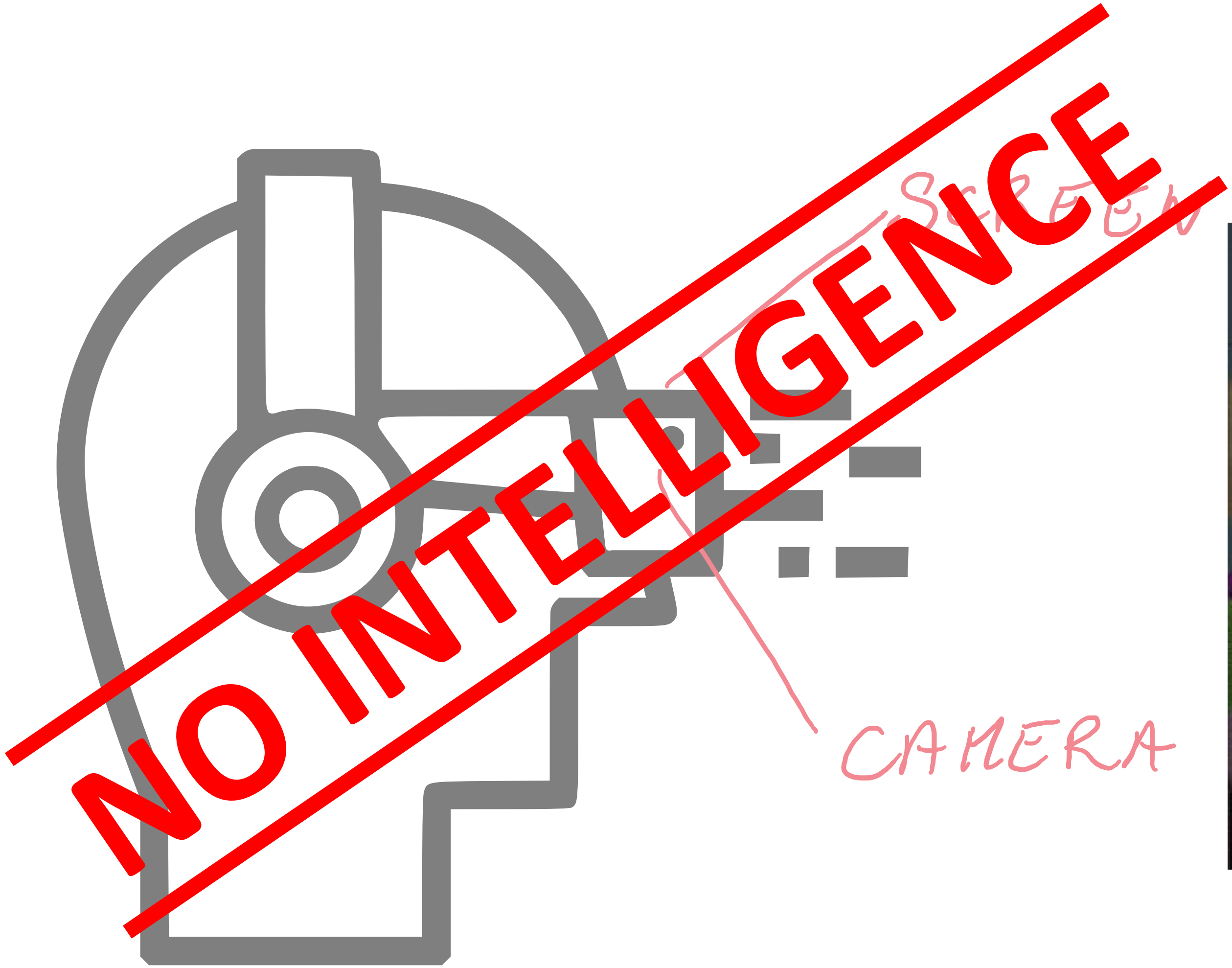
UNIVERSITÀ degli STUDI di CATANIA

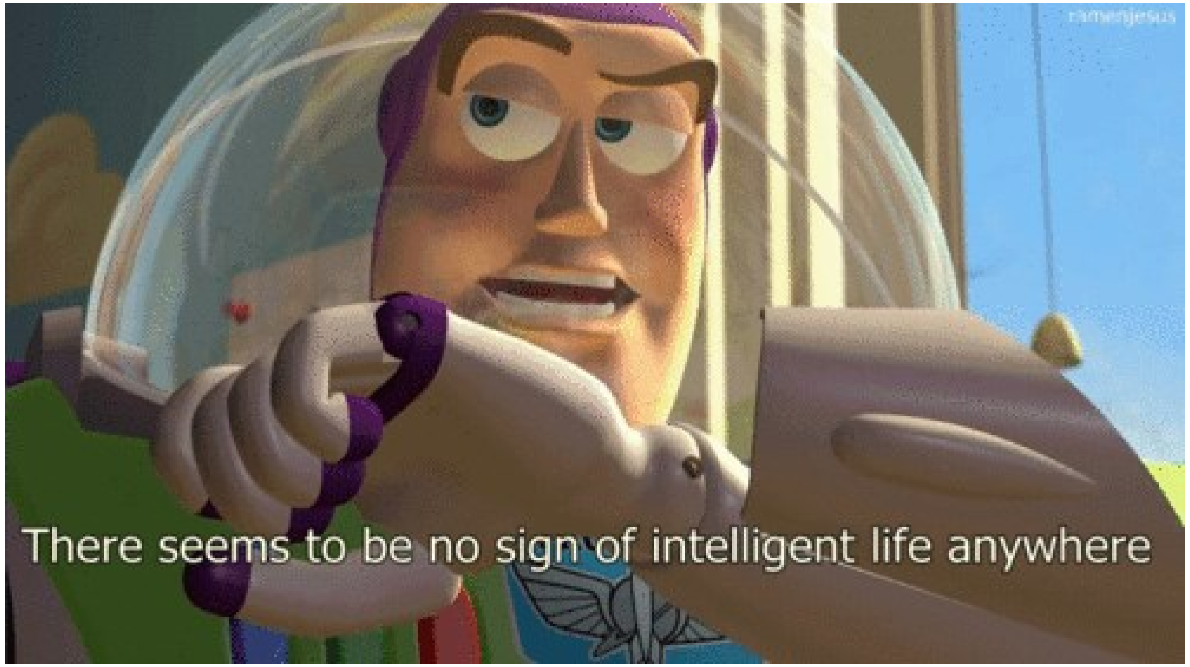
IMAGE PROCESSING LABORATORY

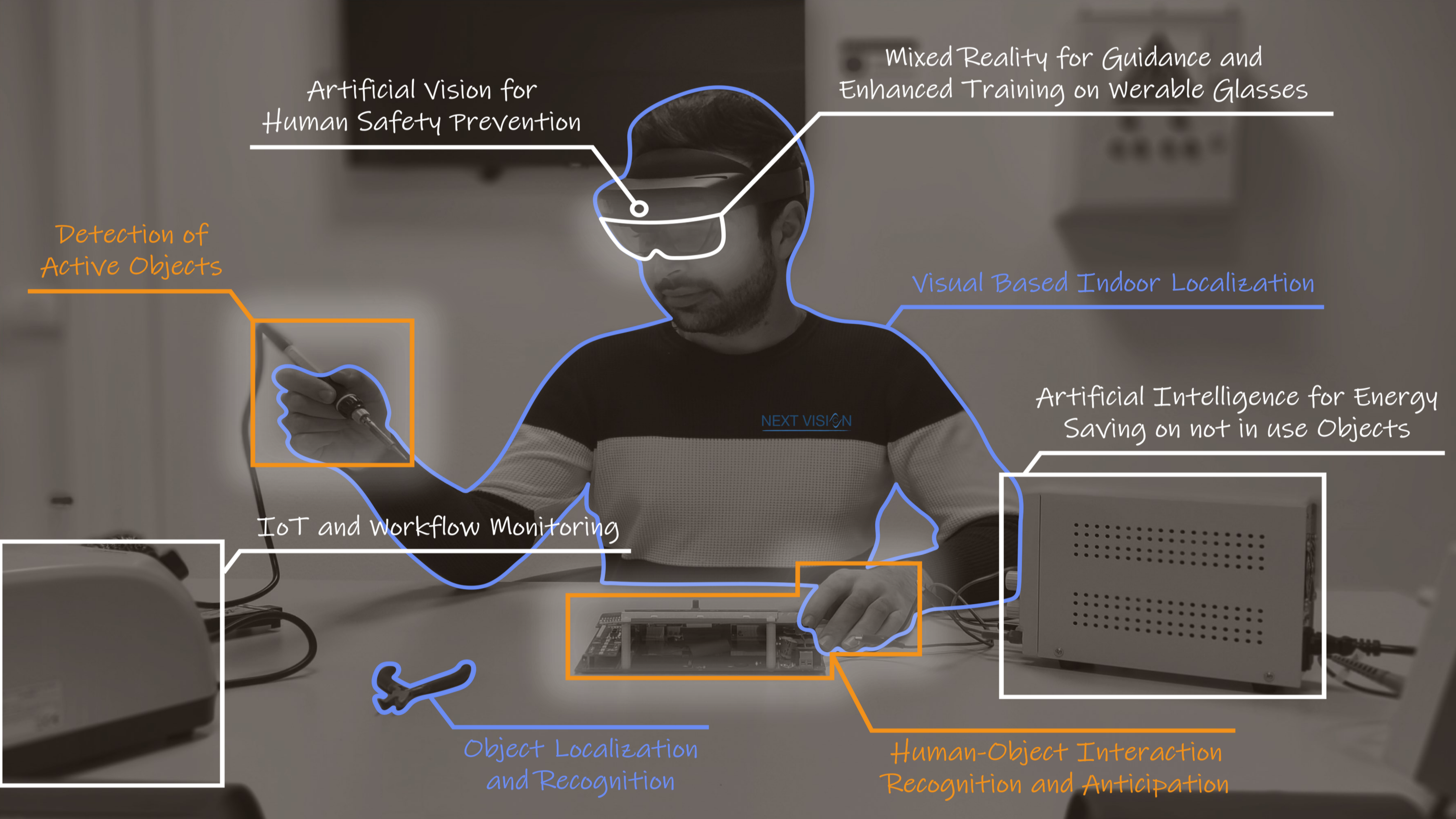
Wearable Devices



Wearable Devices







Artificial Vision for
Human Safety Prevention
Mixed Reality for Guidance and
Enhanced Training on Werable Glasses
Detection of
Active Objects
Visual Based Indoor Localization
NEXT VISION
Artificial Intelligence for Energy
Saving on not in use Objects
IoT and Workflow Monitoring
Object Localization
and Recognition
Human-Object Interaction
Recognition and Anticipation

Solutions: NAIROBI

Solutions: Human-Object Interaction Recognition

Conclusion

- First Person Vision paves the way to a variety of user-centric applications;
- However, we are still missing solid building blocks related to fundamental problems of First Person Vision such as action recognition, object detection, action anticipation and human-object interaction detection;
- Consumer devices are starting to appear, but the near future of First Person Vision is in focused applications such as the ones in industrial scenarios.

Look for us

- **25 May 15:00-16:00 Poster Session**
  - Egocentric Human-Object Interaction Detection Exploiting Synthetic Data
  - Weakly Supervised Attended Object Detection Using Gaze Data as Annotations
  - Panoptic Segmentation in Industrial Environments using Synthetic and Real Data
- **26 May 11:45 - 12:00 Oral session**
  - Unsupervised Multi-Camera Domain Adaptation for Object Detection
- **26 May 15:30 - 16:30 Poster Session**
  - Untrimmed Action Anticipation

Before we leave…

The slides of this tutorial are available online at:

http://www.antoninofurnari.it/talks/iciap2022

Thank you!

Antonino Furnari

Francesco Ragusa

First Person (Egocentric) Vision for Human-Centric Assistance: History, Building Blocks, and Applications

Antonino Furnari, Francesco Ragusa

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

Department of Mathematics and Computer Science - University of Catania

Next Vision s.r.l., Italy

furnari@dmi.unict.it - http://www.antoninofurnari.it/

francesco.ragusa@unict.it - https://iplab.dmi.unict.it/ragusa/

http://iplab.dmi.unict.it/fpv - https://www.nextvisionlab.it/