Building Wearable Assistants with First Person (Egocentric Vision): History, Challenges, Opportunities 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

antonino.furnarni@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/

Agenda

1) Part I: Definitions, motivations, history and research trends [09.00 - 10.30] – 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.

   Coffee Break [10.30 – 11.00]

2) **Part II: Building Blocks for First Person Vision Systems [11.00 – 12.30] – Francesco Ragusa**
   a) EgoData Acquisition & Datasets;
   b) Fundamental Task in First Person Vision:
      a) Localization;
      b) Object Detection and Recognition;
      c) Egocentric Human-Object Interaction;
      d) Action/Activities;
      e) Anticipation.
   c) Real Application Examples developed at Next Vision;
   d) Conclusion.

Part 2

Building Blocks for First Person Vision Systems

Data Acquisition

Four things to pay attention to when collecting first person visual data

Video Quality

Field of View

Wearing Modality

Other Modalities

Data Acquisition – Video Quality

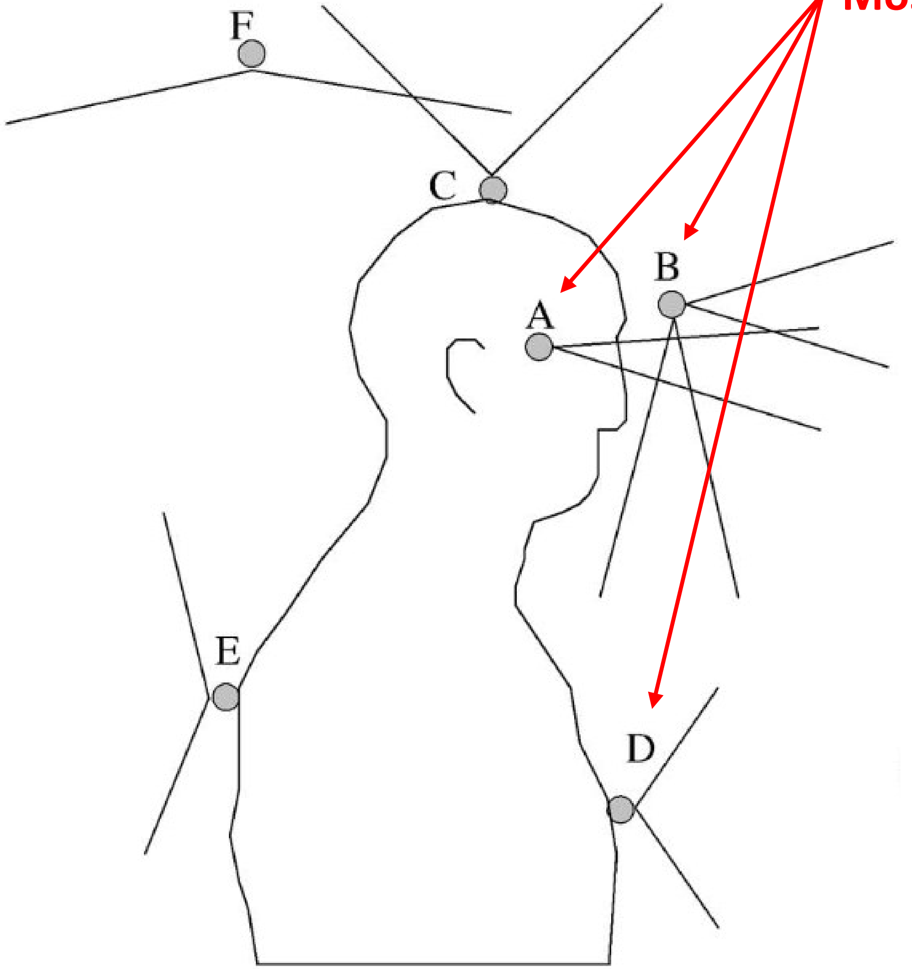
- Try to get a high quality camera to get high quality images!
- Egocentric video is subject to motion blur and exposure issues.

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

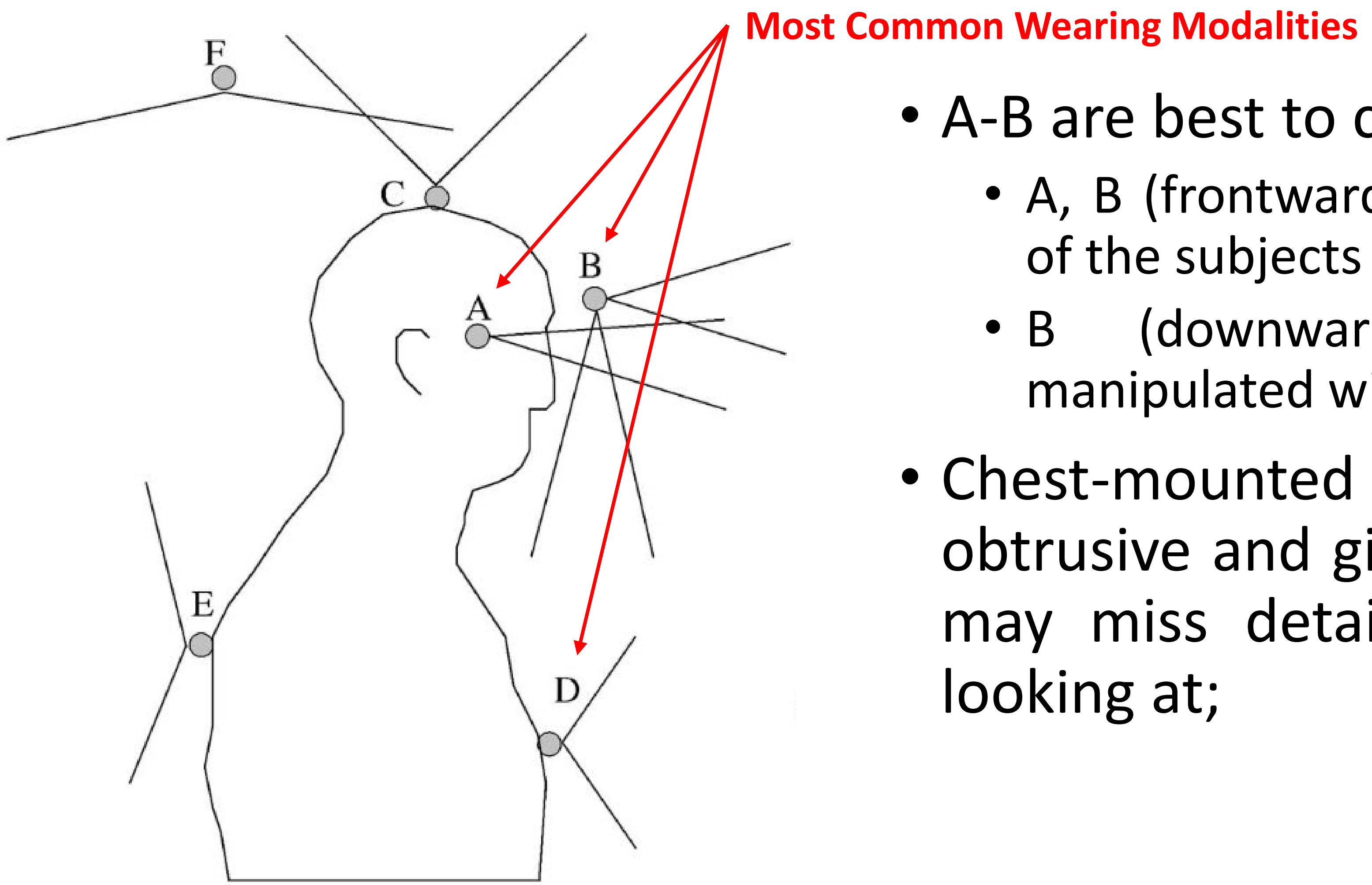
B (frontward)

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;

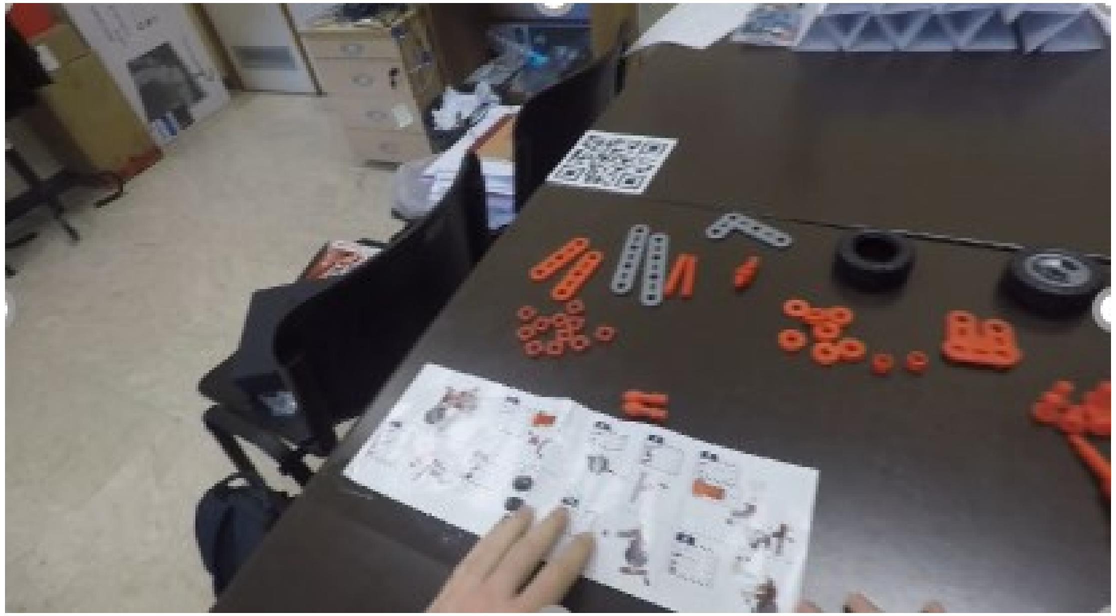
Mayol-Cuevas, W. W., Tordoff, B. J., & Murray, D. W. (2009). On the choice and placement of wearable vision sensors. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, *39*(2), 414-425.

Data Acquisition – Field of View (FOV)

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

Narrow Angle

Wide Angle



Data Acquisition – Other Modalities – Depth

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

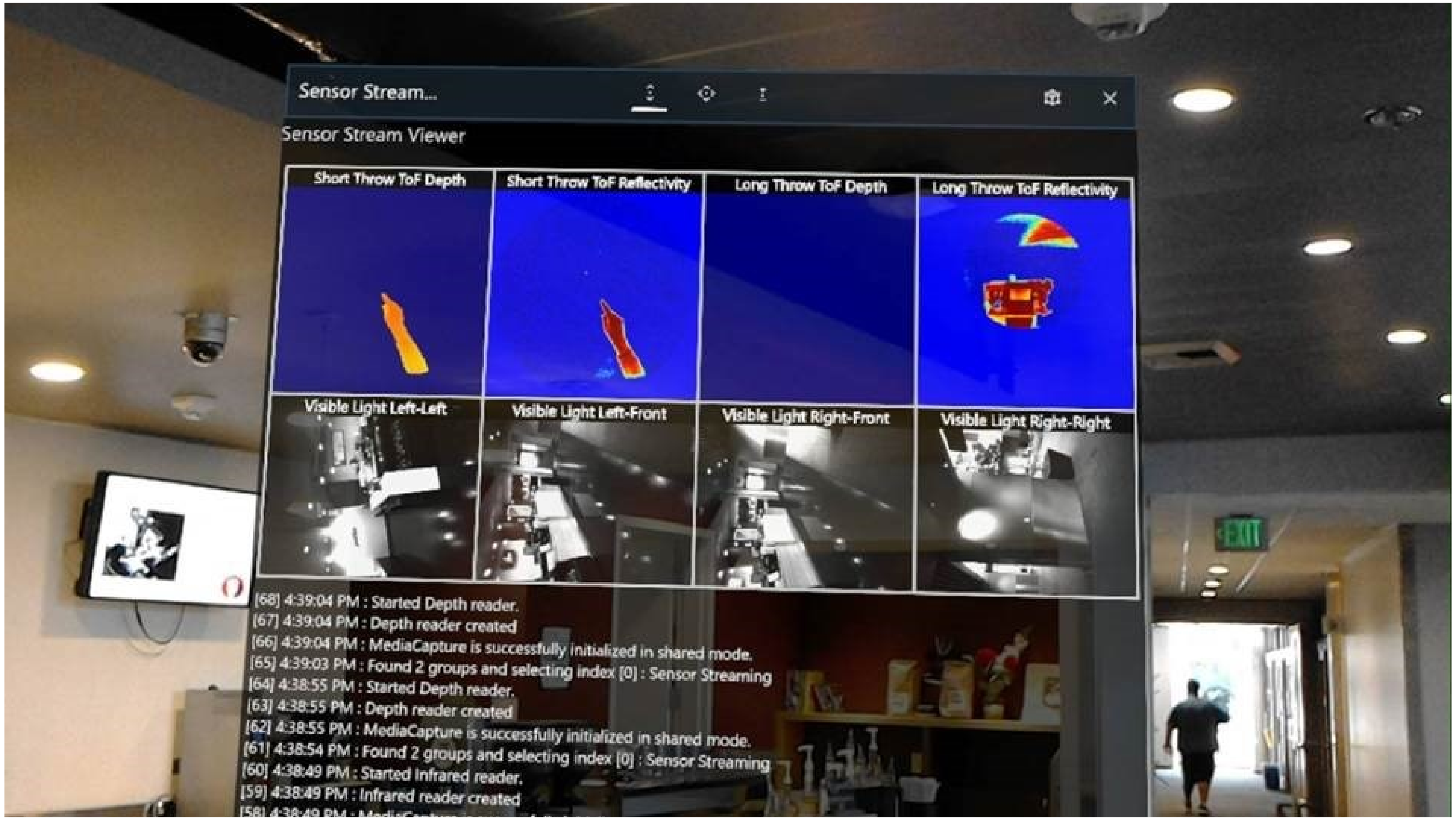
Wan, S., & Aggarwal, J. K. (2015). Mining discriminative states of hands and objects to recognize egocentric actions with a wearable RGBD camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 36-43).

https://github.com/microsoft/HoloLensForCV

Data Acquisition – Other Modalities – Depth (2)

- Microsoft HoloLens has a «Research Mode» which allows to access:
  - short-range depth
  - long-range depth;
  - IR reflectivity;

Microsoft HoloLens Research Mode



https://docs.microsoft.com/en-us/windows/mixed-reality/research-mode

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

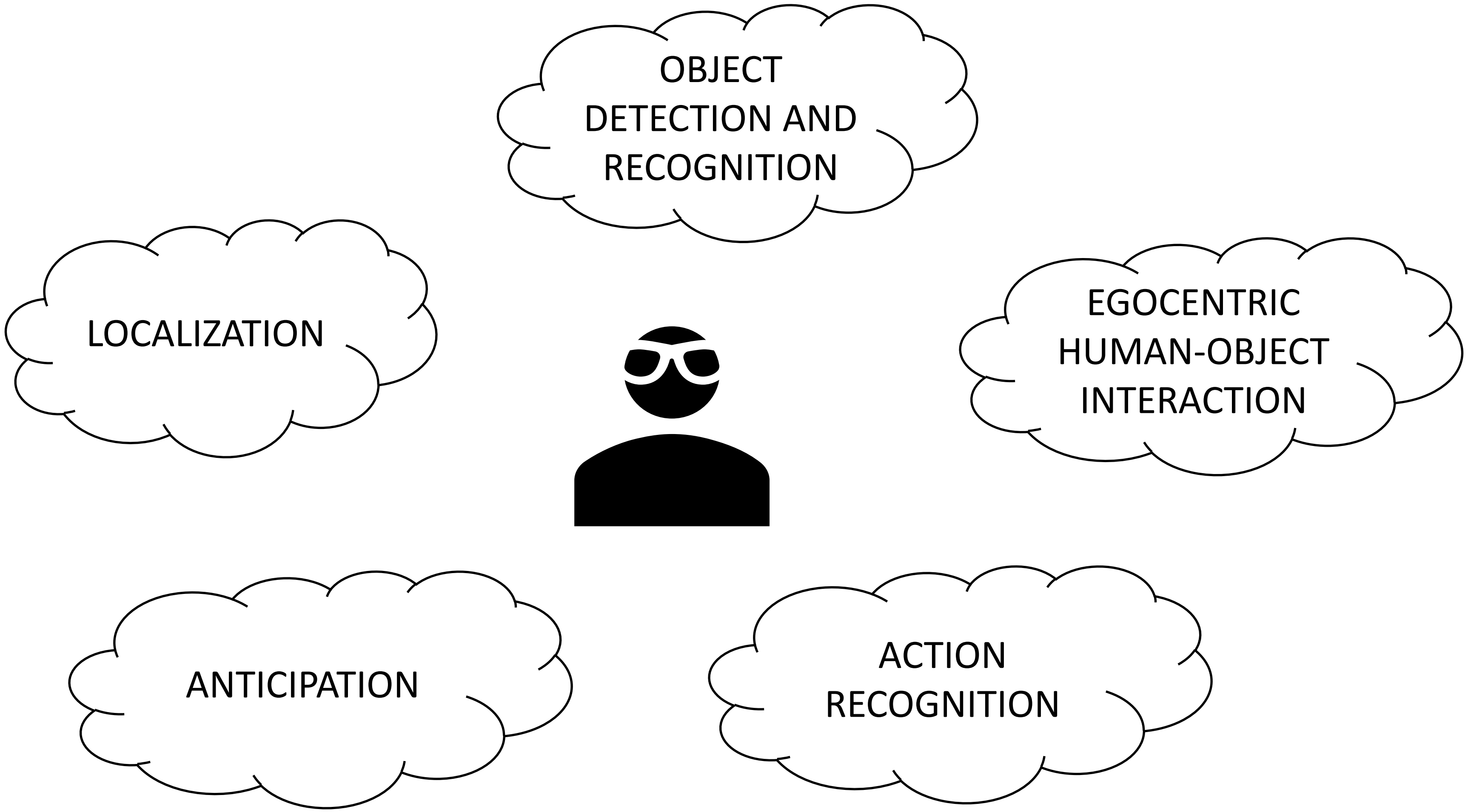
Datasets (non-exhaustive)

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

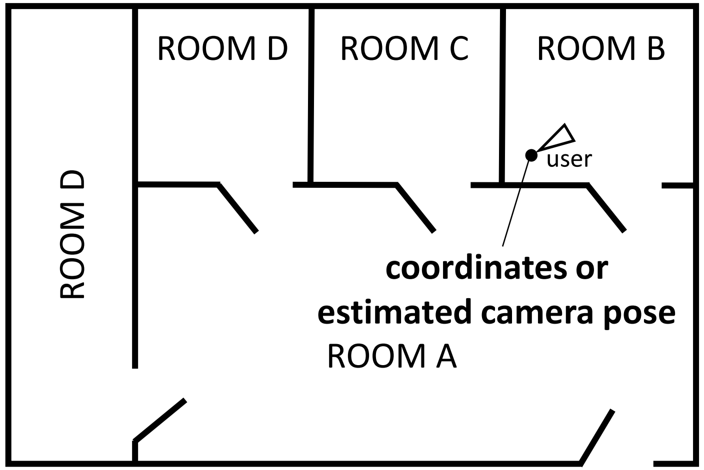
- Knowing the location of the user for a First Person Vision system is important to implement contextual awareness
  - Behave differently depending on the environment
    - Generate reminders when I get to a particular place
      - «remember to do the laundary when you get home»;
    - Turn notifications 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

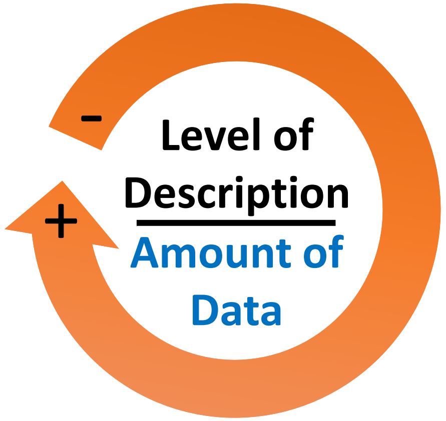
SCENE RECOGNITION

off-the-shelf detectors

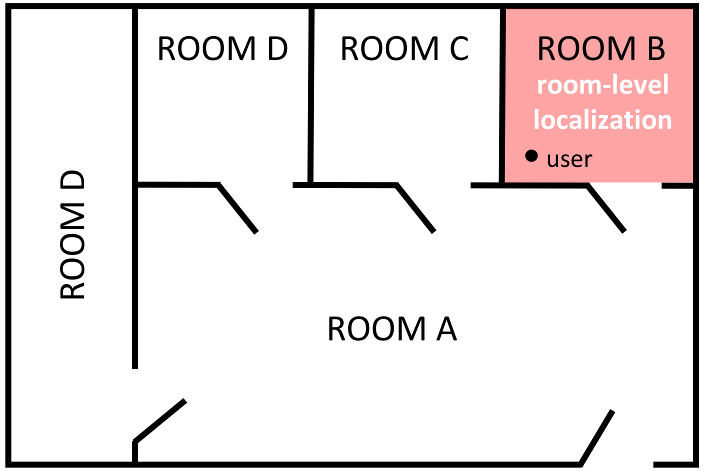
CAMERA POSE-ESTIMATION



3D reconstruction of the building



ROOM-LEVEL RECOGNITION



moderate amount of training data

Scene Recognition

?

Inside city

Street

Highway

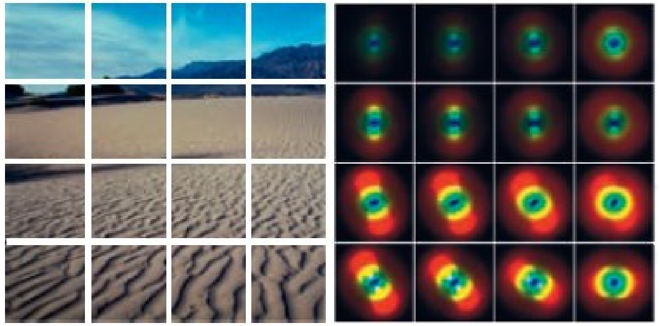
Coast

…

- 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.

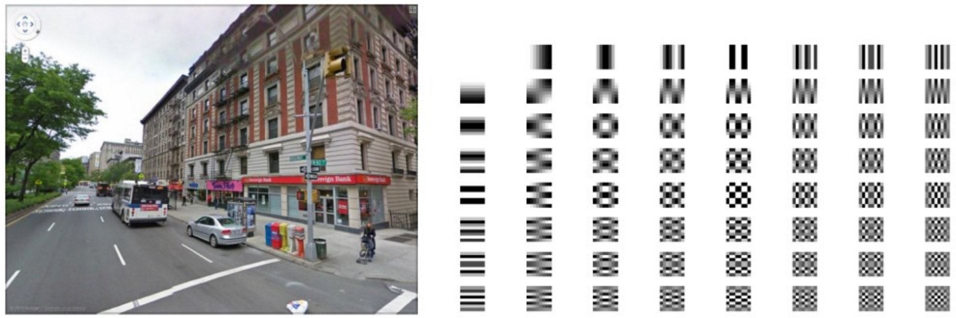
COMPUTATIONALLY INEXPENSIVE ALGORITHMS

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)

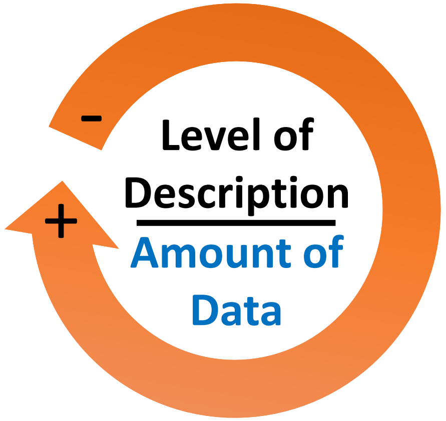
- Places is a large (10M images – 400+ classes) dataset for scene recognition;
- CNN models trained to recognize 365 scene classes available for download;
- Can be used off-the-shelf!

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*

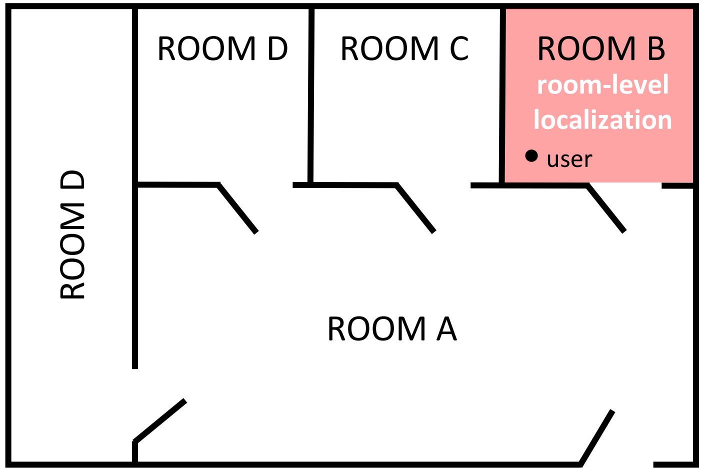
Localization – Levels of Granularity

SCENE RECOGNITION

off-the-shelf detectors

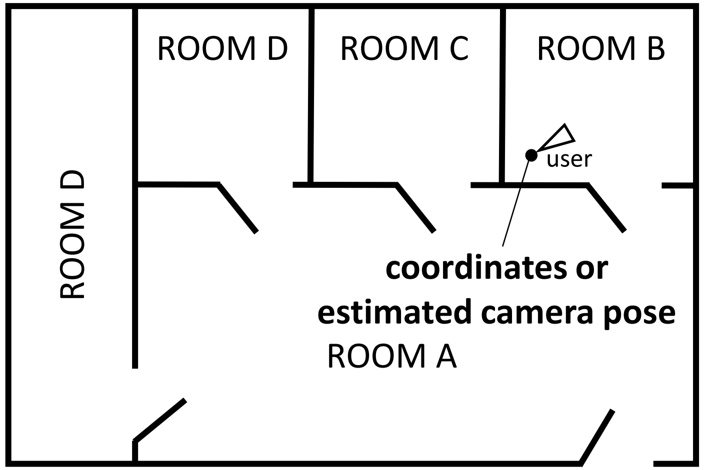


ROOM-LEVEL RECOGNITION



moderate amount of training data

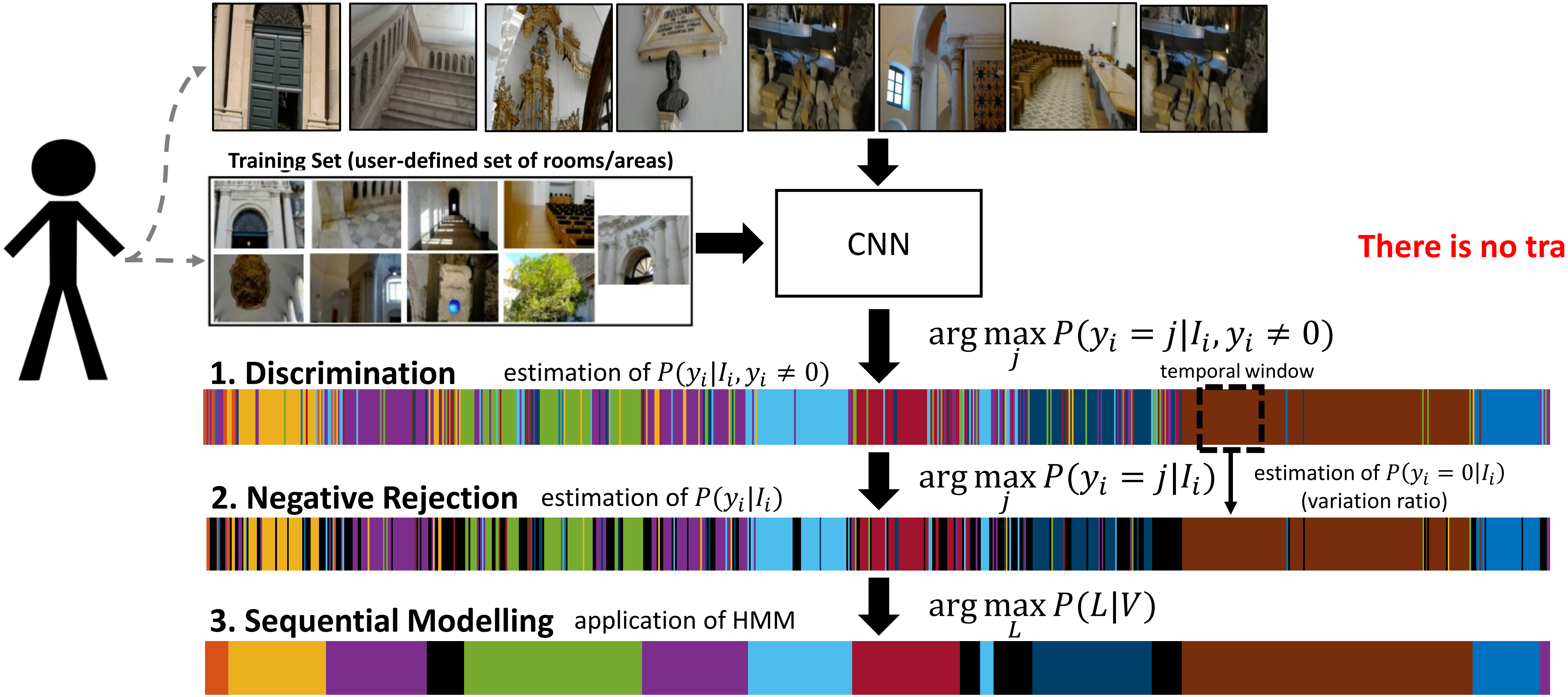
CAMERA POSE-ESTIMATION



3D reconstruction of the building

CODE HERE -> https://iplab.dmi.unict.it/VEDI/

Room-Level Localization – Full Model

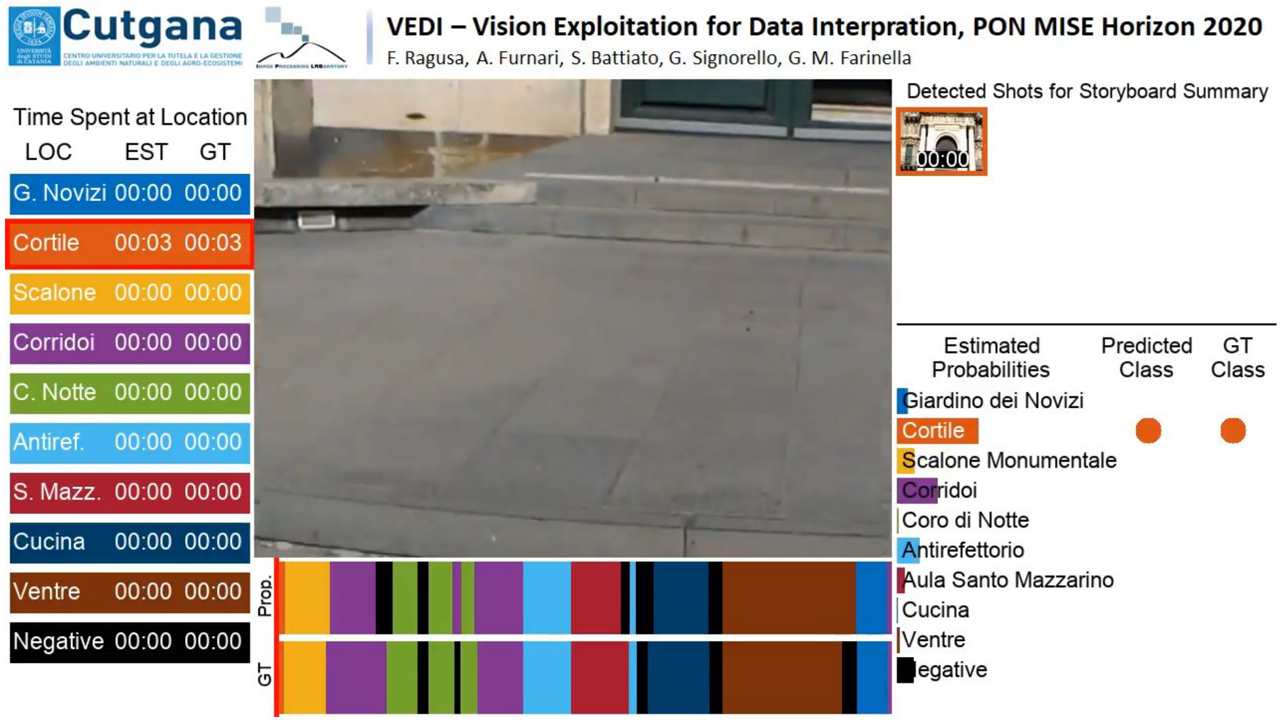


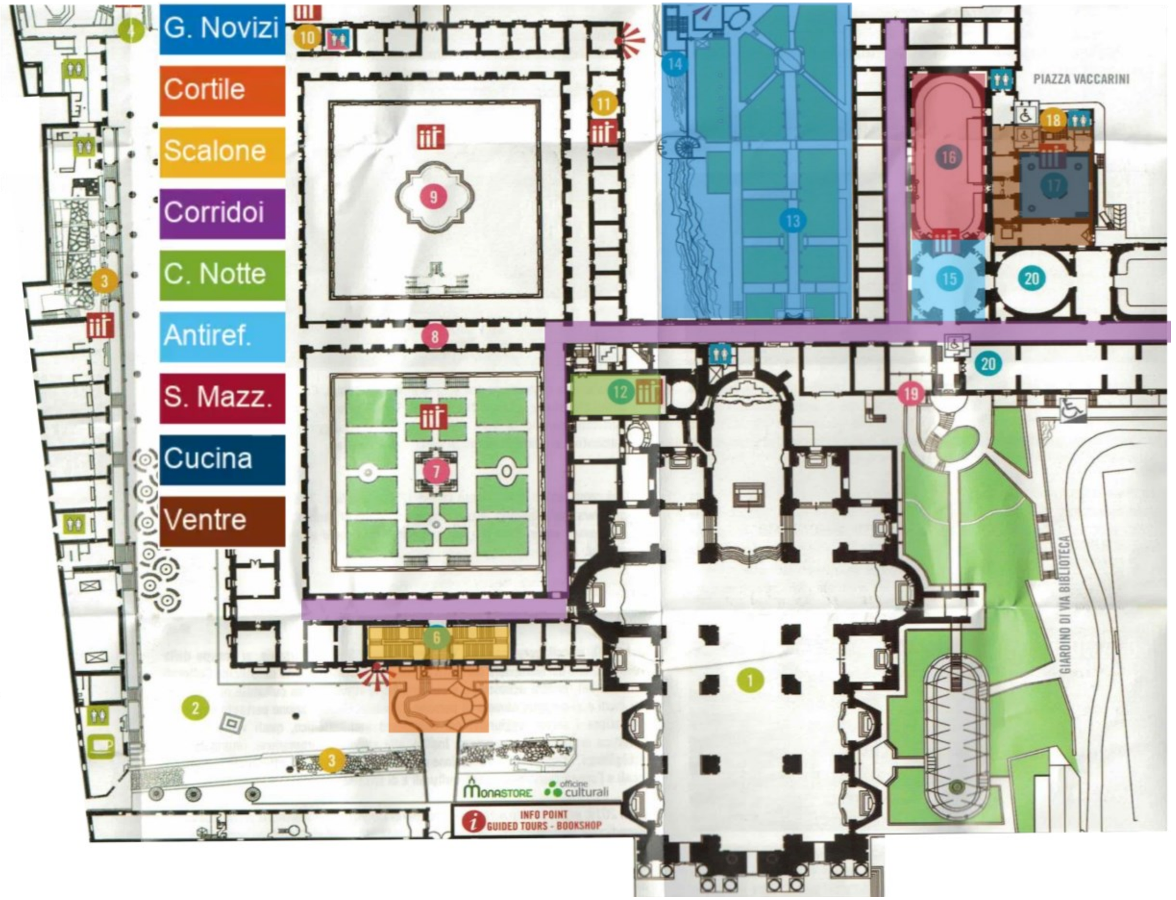
There is no training negatives!

F. Ragusa, A. Furnari, S. Battiato, G. Signorello, G. M. Farinella. Egocentric Visitors Localization in Cultural Sites. In Journal on Computing and Cultural Heritage (JOCCH), 2019.

Room-Level Localization

Localizing the user in a larger environment (e.g., a museum).



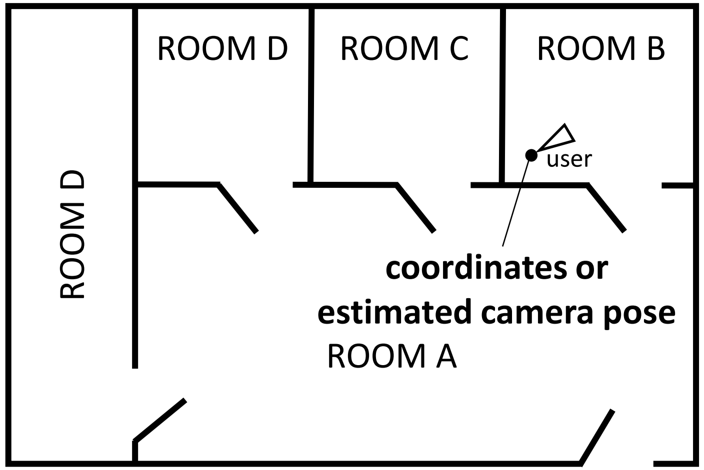


Localization – Levels of Granularity

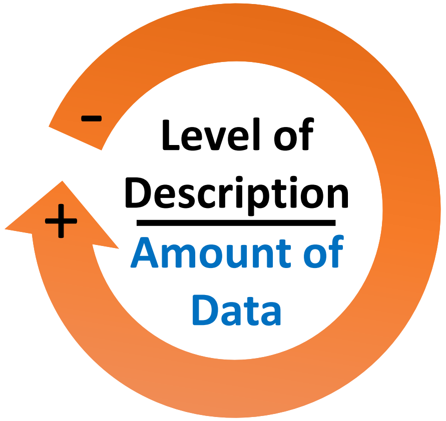
SCENE RECOGNITION

off-the-shelf detectors

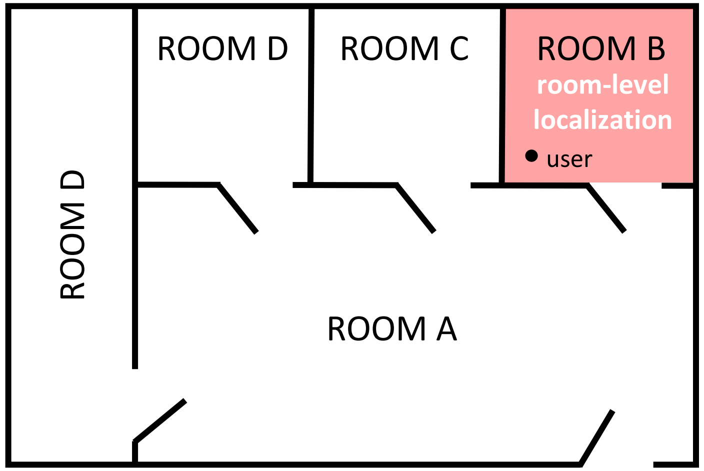
CAMERA POSE-ESTIMATION



3D reconstruction of the building



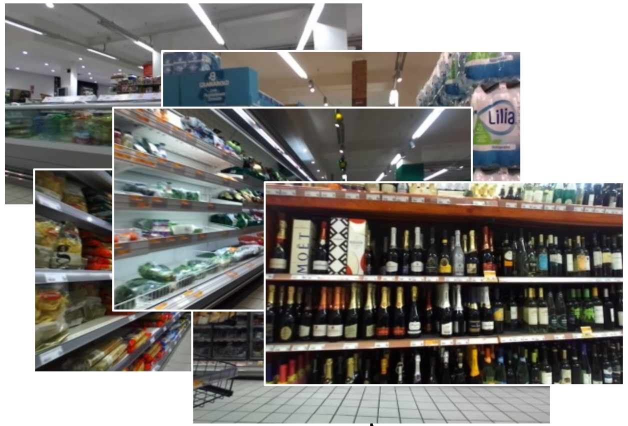
ROOM-LEVEL RECOGNITION



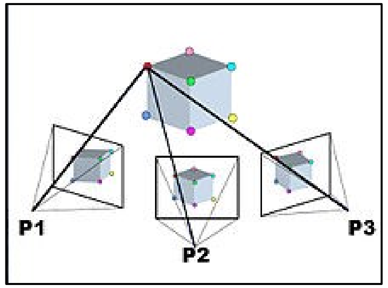
moderate amount of training data

Camera Pose Estimation – Dataset Creation

Images



Structure from Motion (SfM)

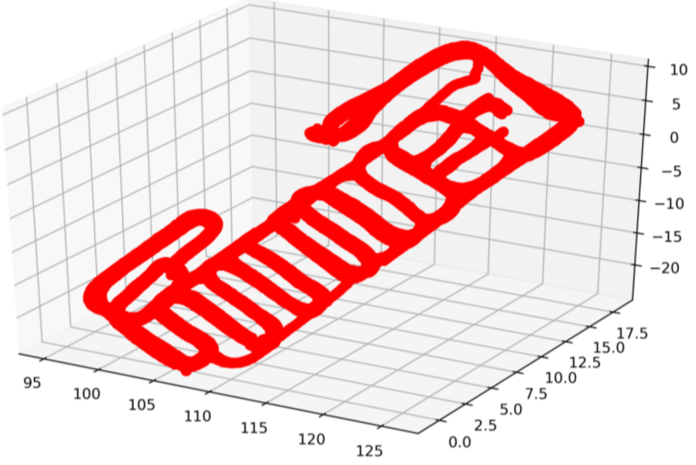


(P,Q)

Attach estimated 6DOF pose to each image

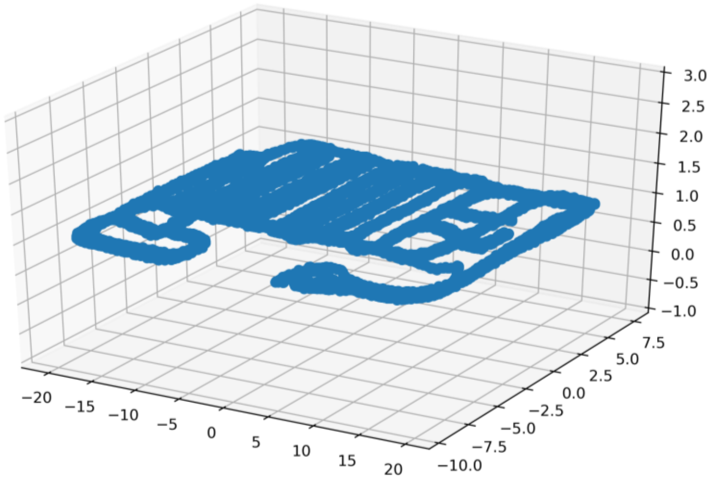
3D Model

Arbitrary Coordinate System (pose/scale)

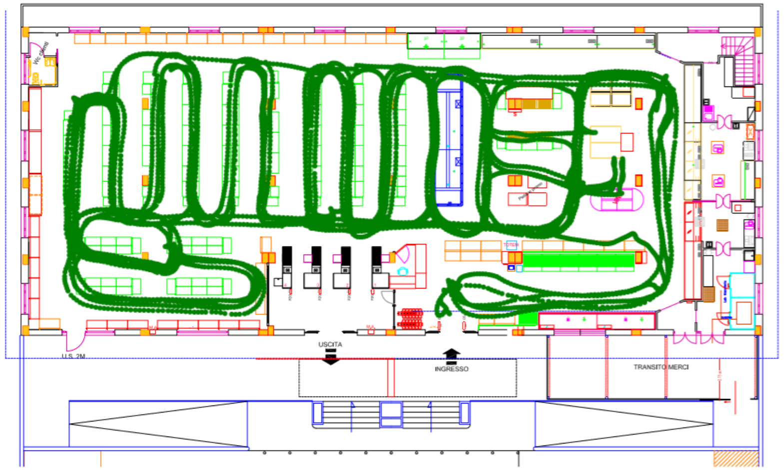


camera poses

PCA



rotated poses



scaled/aligned poses

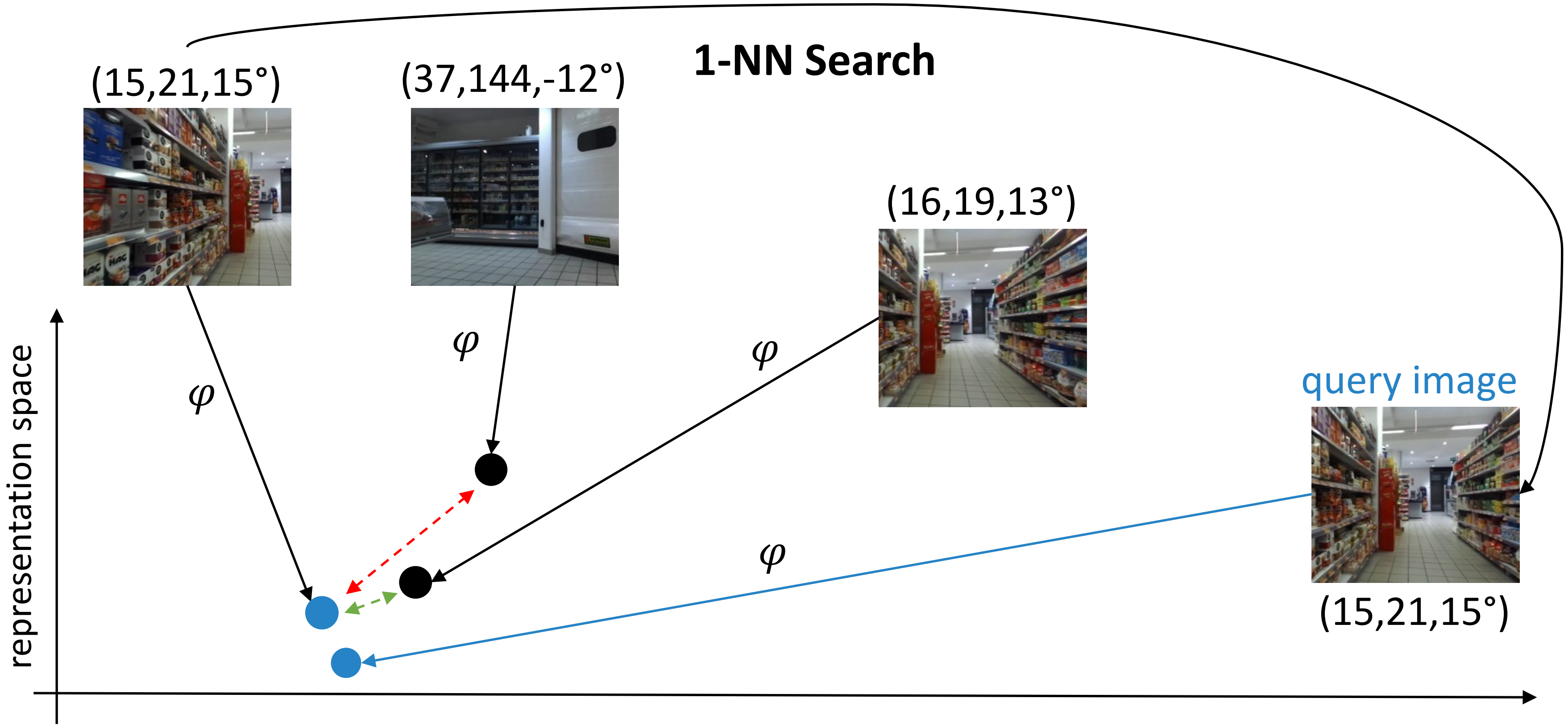
Structure from Motion (SfM) Softwares

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 learn a representation function φ which maps close to each other images of nearby locations



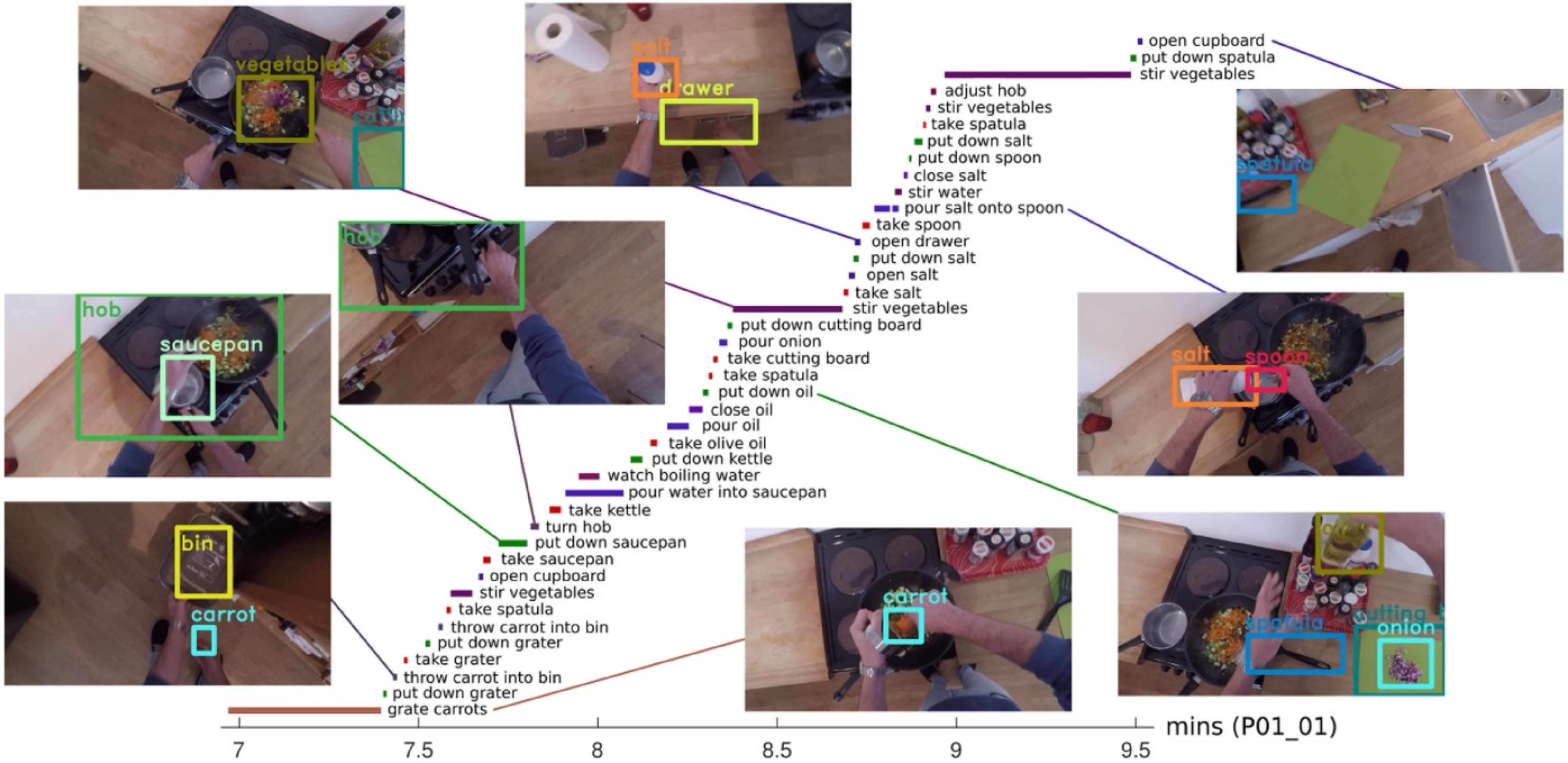
E. Spera, A. Furnari, S. Battiato, G. M. Farinella, Egocentric Shopping Cart Localization, International Conference on Pattern Recognition (ICPR), 2018
S. A. Orlando, A. Furnari, S. Battiato, G. M. Farinella. Image-Based Localization with Simulated Egocentric Navigations. VISAPP 2019

Objects and Actions are tight!

Useful to know what is in the scene

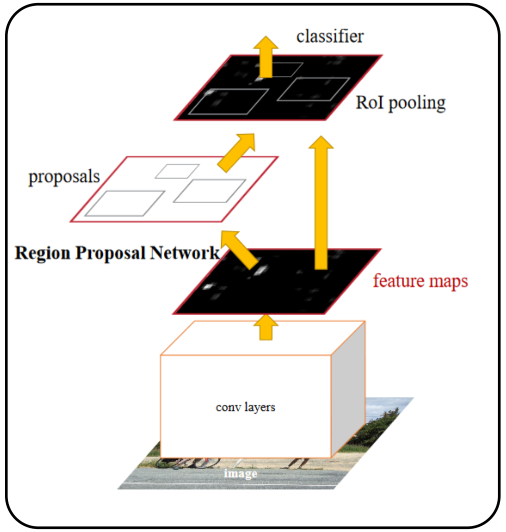
Useful to know what actions can be performed

Object Detection

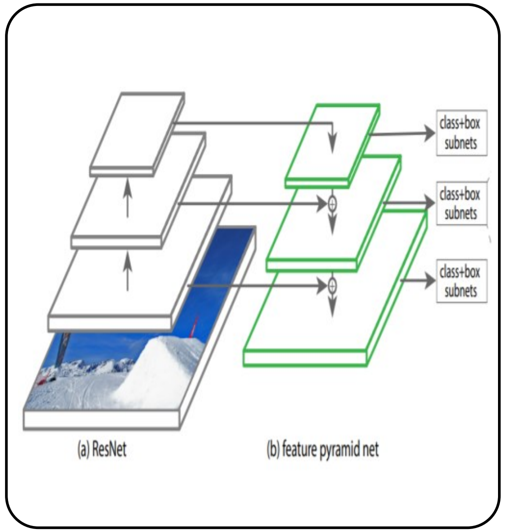


D. Damen, H. Doughty, G. M. Farinella, S. Fidler, A. Furnari, E. Kazakos, D. Moltisanti, J. Munro and T. Perrett, W. Price, M. Wray (2018). Scaling Egocentric Vision: The EPIC-KITCHENS Dataset. In European Conference on Computer Vision.

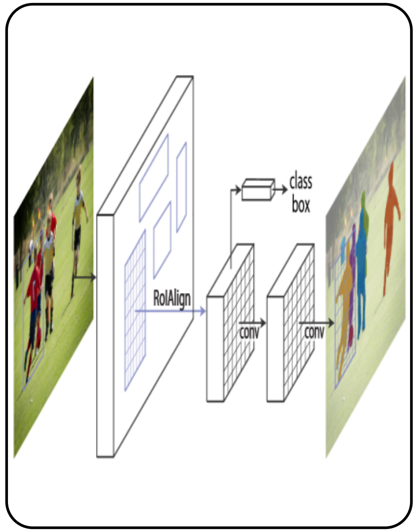
Off-the-shelf object detectors



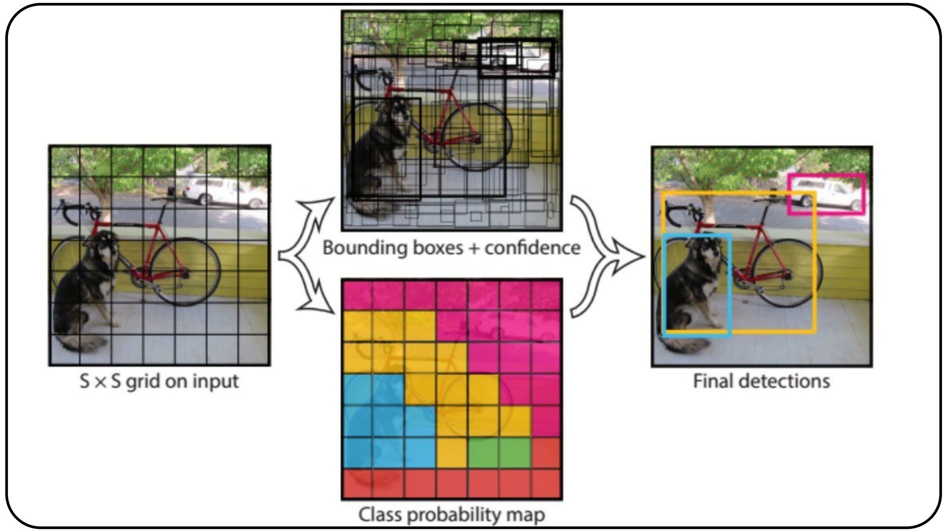
Faster-RCNN
(bounding boxes)



RetinaNet
(bounding boxes - faster)



Mask-RCNN
(boxes + segments)



YOLO
(much faster, but less accurate)

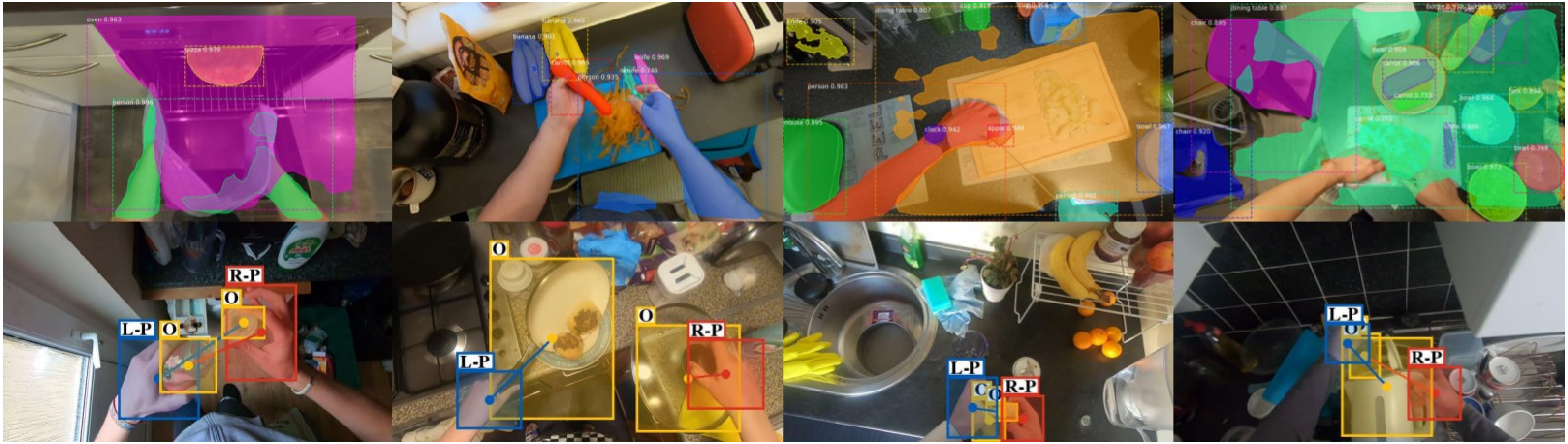
https://github.com/facebookresearch/detectron2

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/

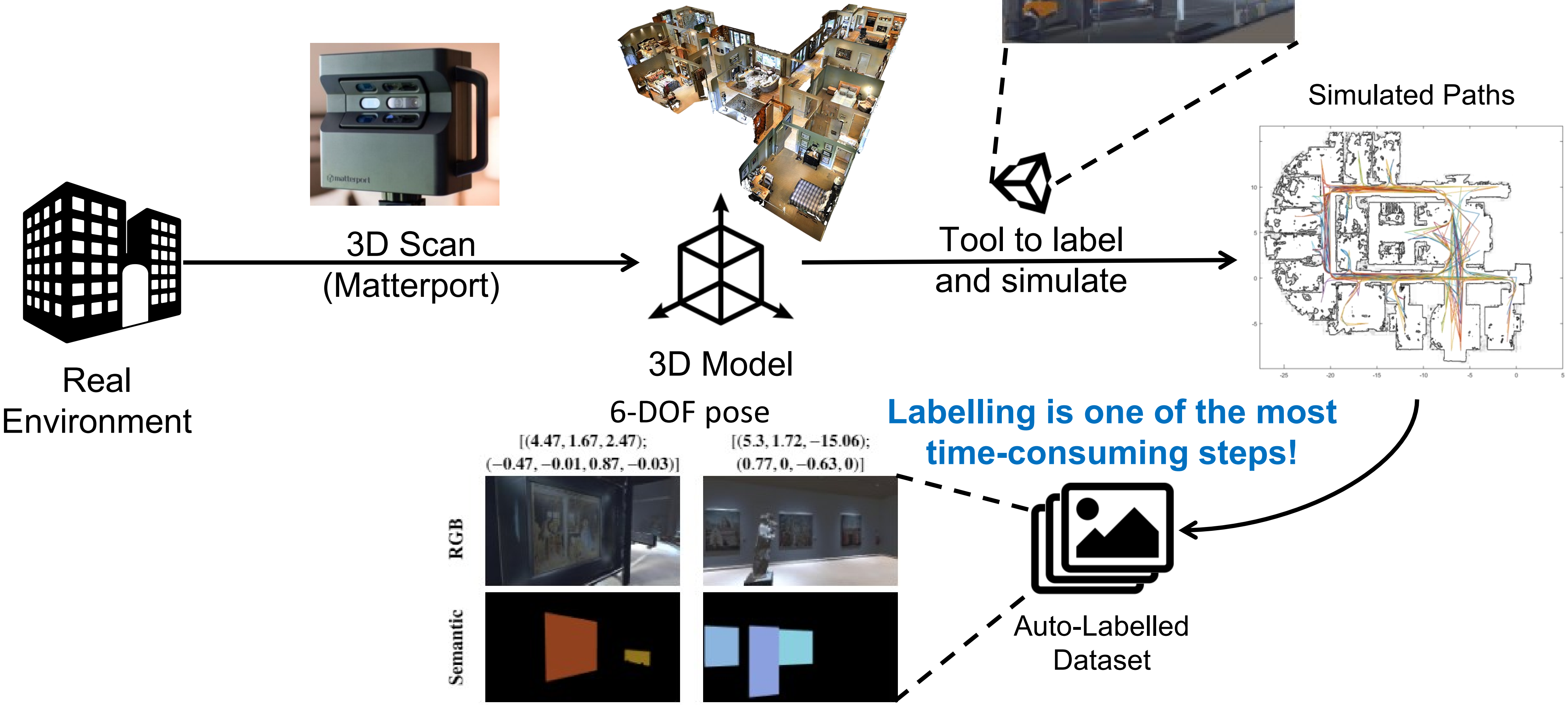
https://iplab.dmi.unict.it/EGO-CH/

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

- In some scenario, it could be necessary to fine-tune an object-detector with application-specific data.
- Main egocentric datasets providing bounding box annotations.
- EGO4D is multi-domain annotated with 295K 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.

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

Domain Adaptation for Semantic Object Segmentation Dataset

Synthetic Images

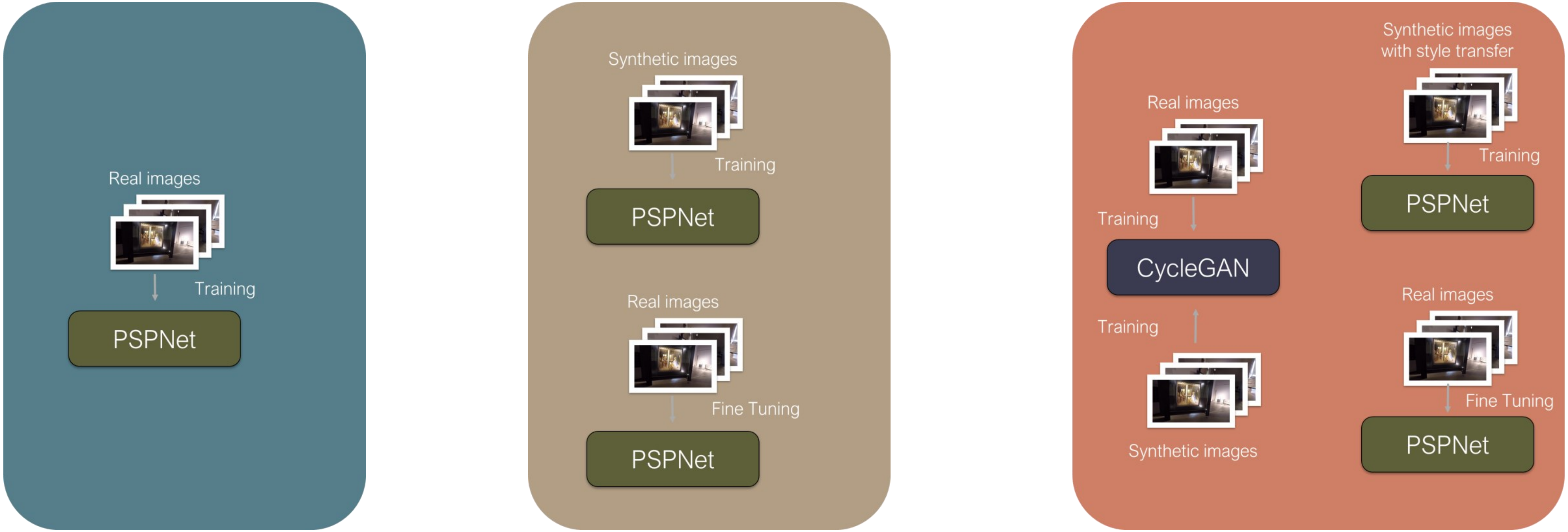
Real Images

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/

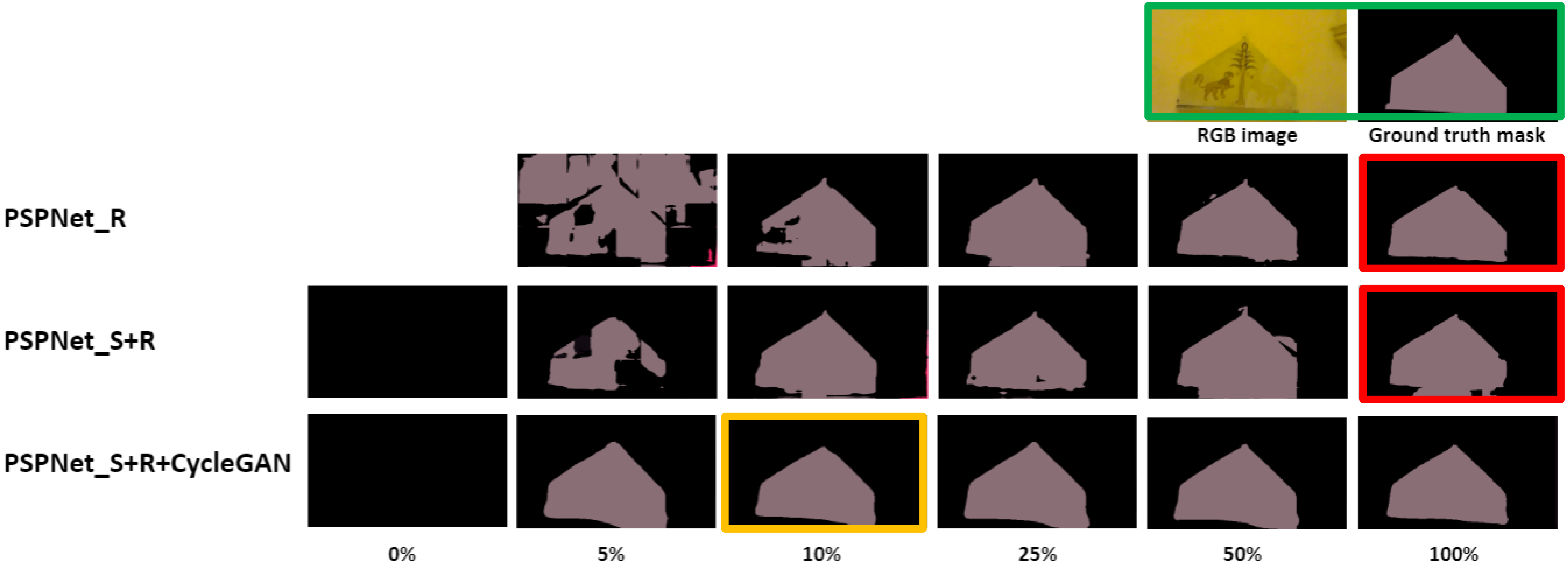
Domain Adaptation for Semantic Object Segmentation Dataset



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).

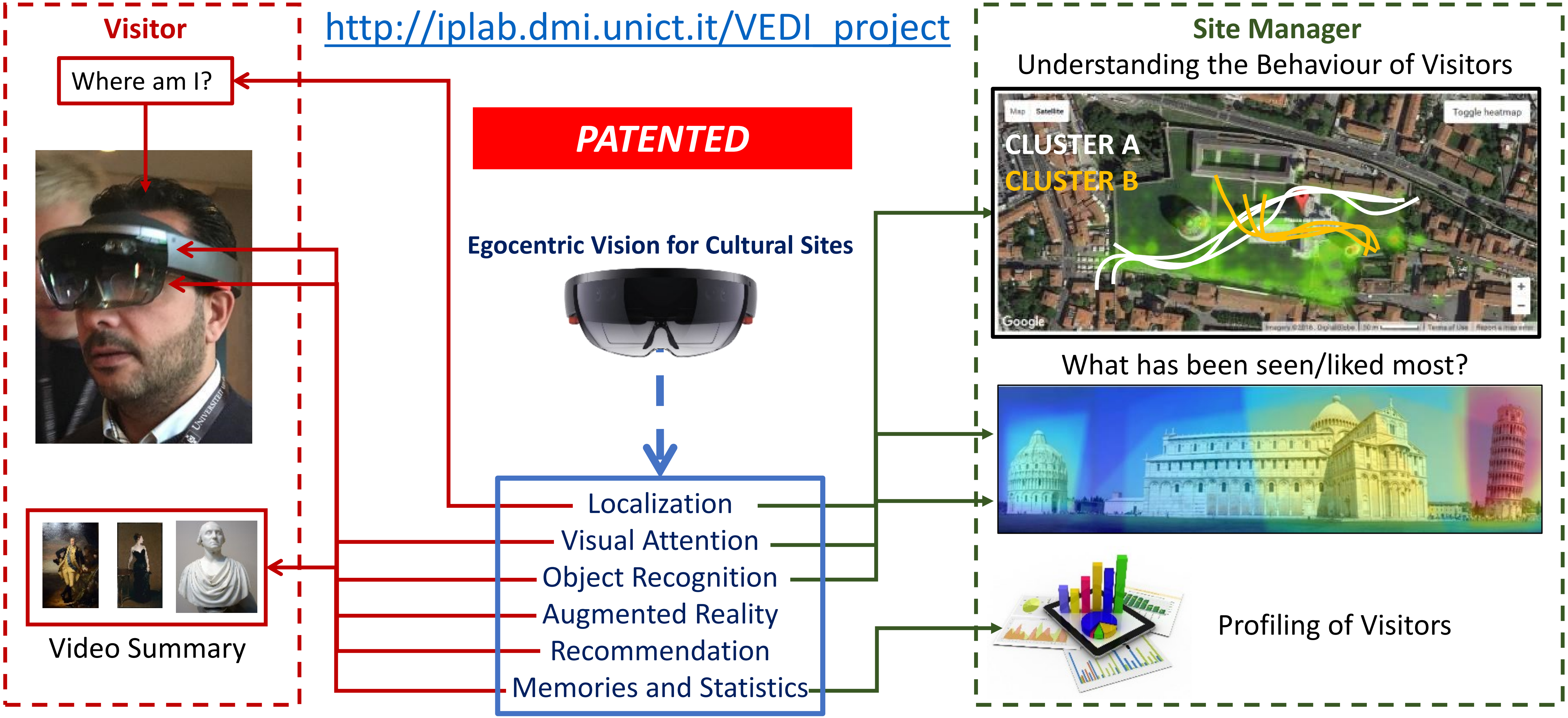
DATA HERE -> https://iplab.dmi.unict.it/EGO-CH-OBJ-SEG/

Domain Adaptation for Semantic Object Segmentation Dataset



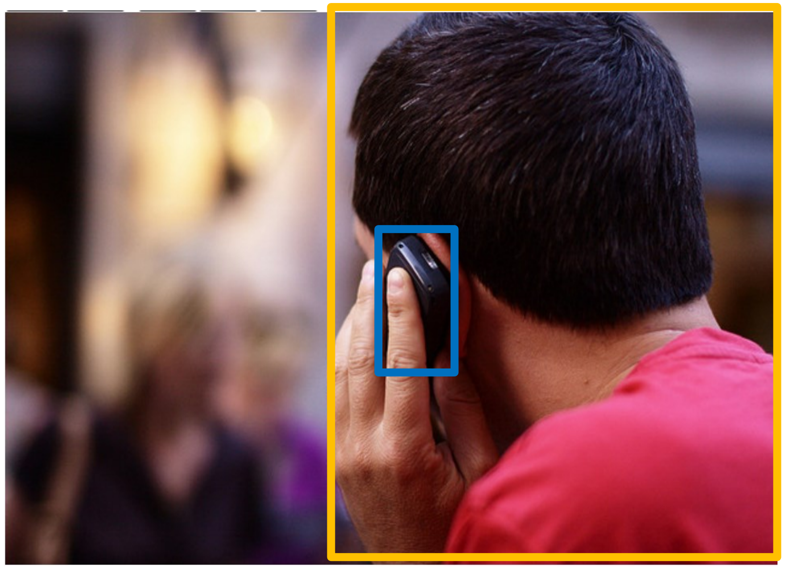
Francesco Ragusa, Daniele DiMauro, Alfio Palermo, Antonino Furnari, Giovanni Maria Farinella (2020). Semantic Object Segmentation in Cultural Sites using Real and Synthetic Data. International Conference on Pattern Recognition (ICPR).

Vision Exploitation for Data Interpretation (VEDI)

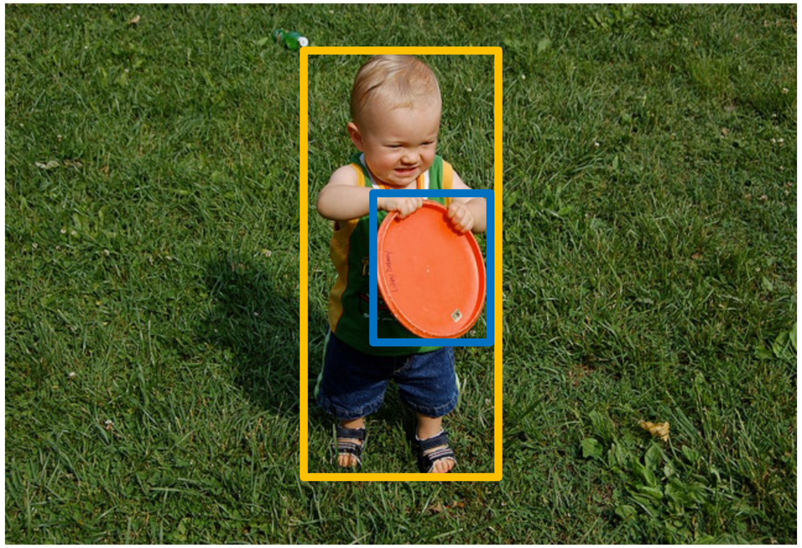


G. M. Farinella, G. Signorello, S. Battiato, A. Furnari, F. Ragusa, R. Leonardi, E. Ragusa, E. Scuderi, A. Lopes, L. Santo, M. Samarotto. VEDI: Vision Exploitation for Data Interpretation. In 20th International Conference on Image Analysis and Processing (ICIAP), 2019

Human-Object Interaction



<human, talks, cellphone>



<human, holds, freesbe>

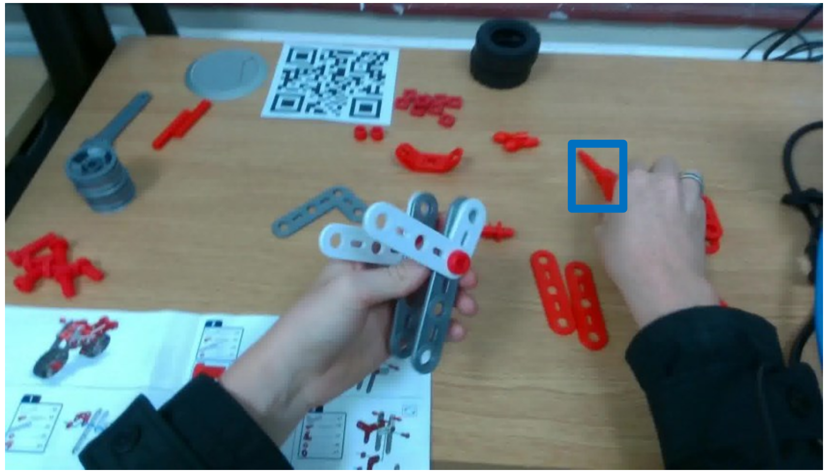
Georgia Gkioxari, Ross Girshick, Piotr Dollàr, Kaiming He. (2018). Detecting Human-Object Interactions. In IEEE Conference on Computer Vision and Pattern Recognition (CVPR).

Egocentric Human-Object Interaction

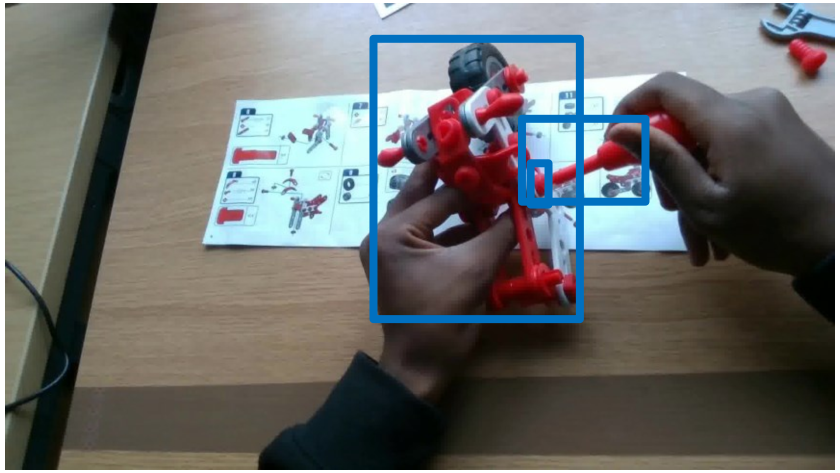
$O = \{o_1, o_2, \ldots, o_n\}$

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

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



<take, screwdriver>

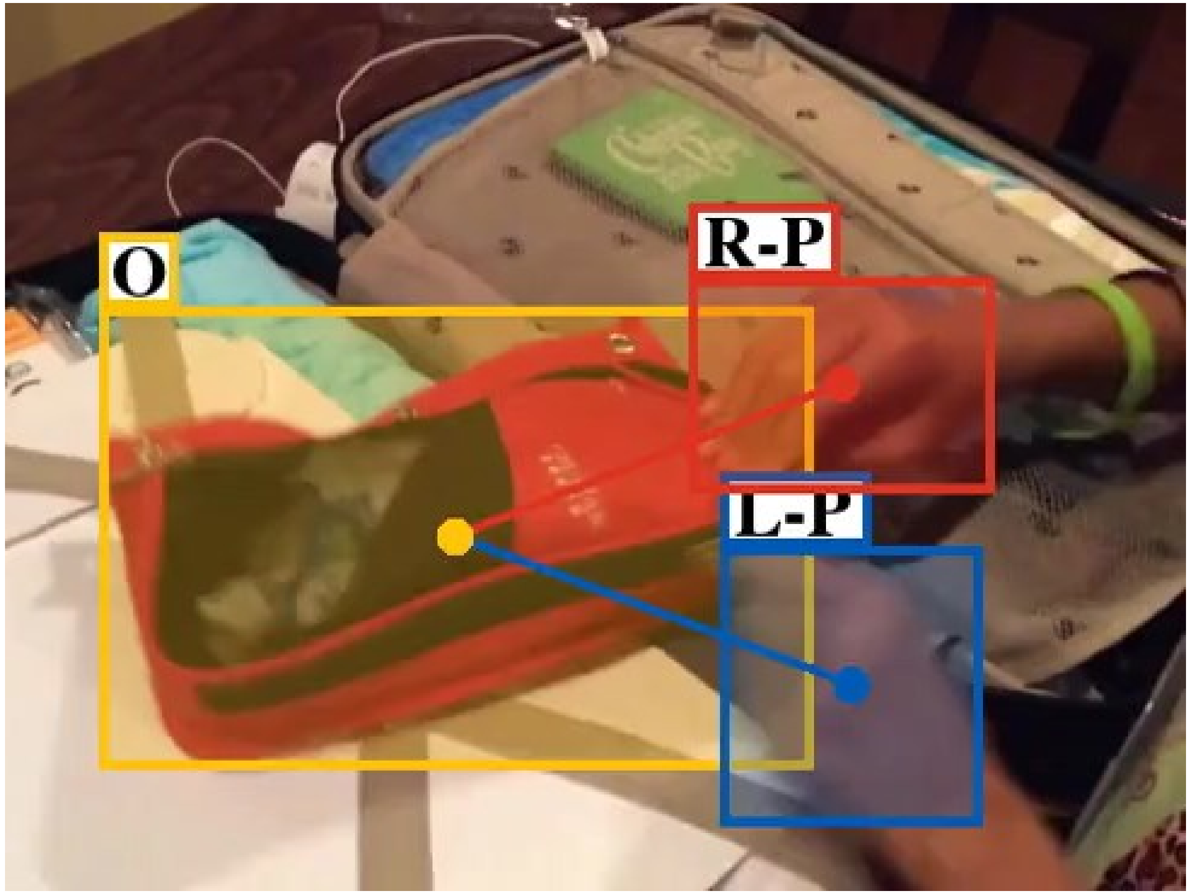


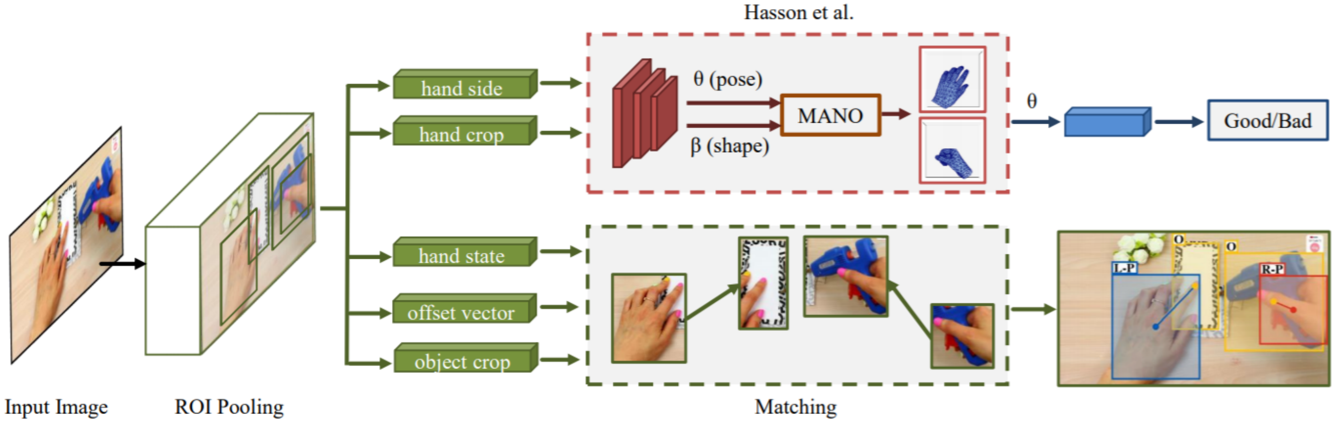
<screw, {screwdriver, screw, partial_model}>

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





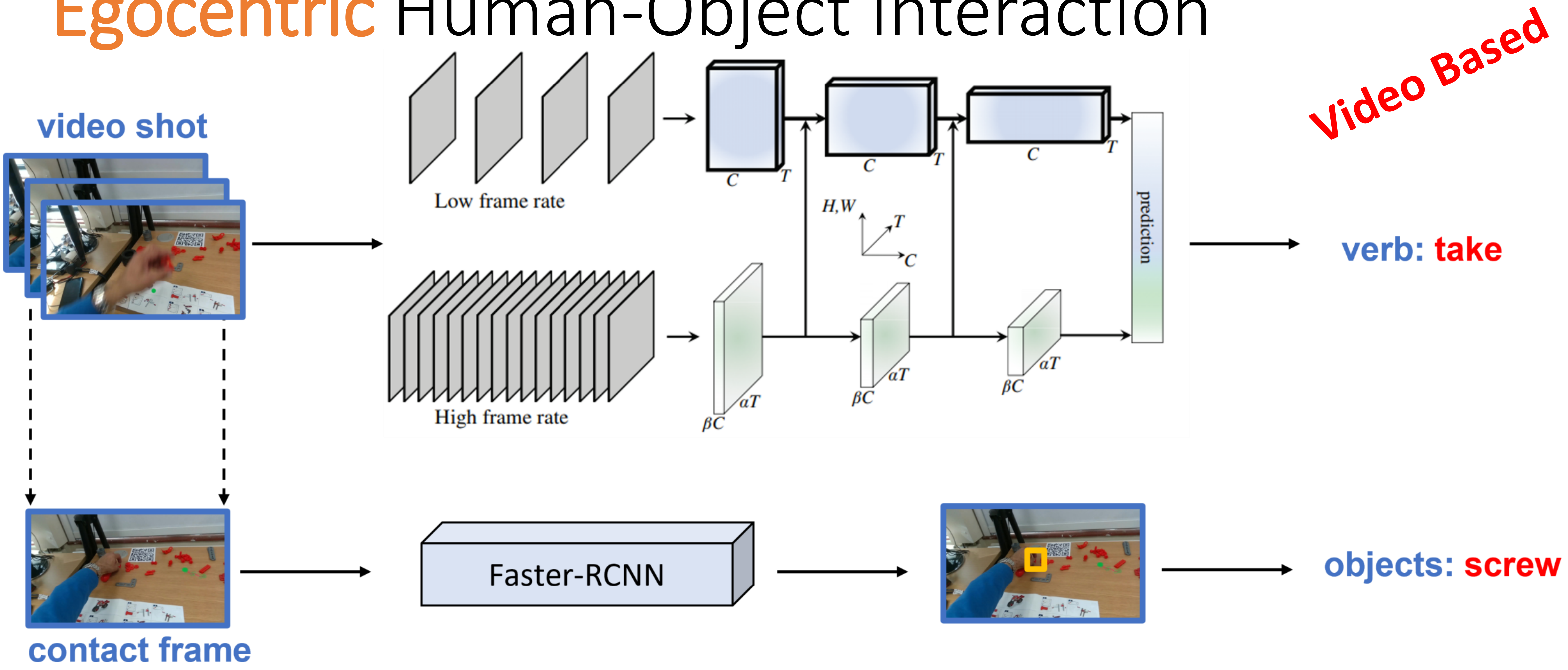
An «augmented» detector which recognizes:

- 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

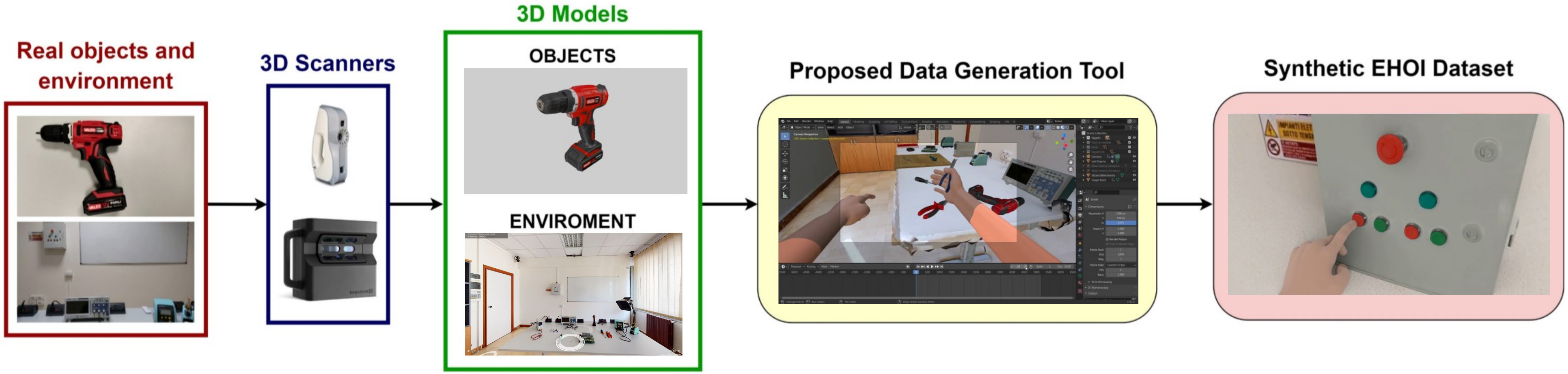
Egocentric Human-Object Interaction



F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding (CVIU), 2023.

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?

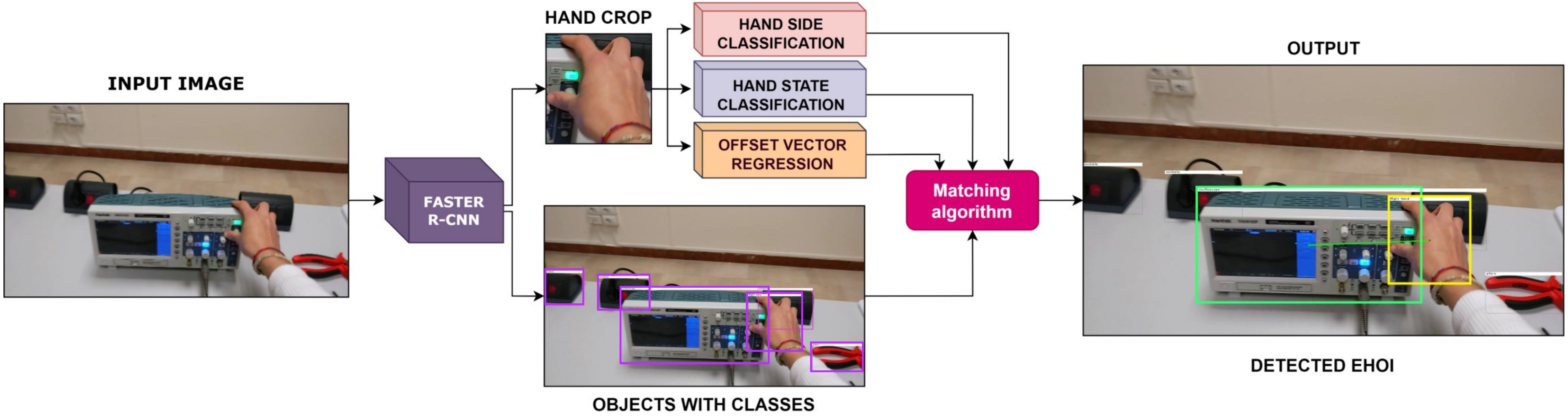
ENIGMA Laboratory

19 objects categories

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?

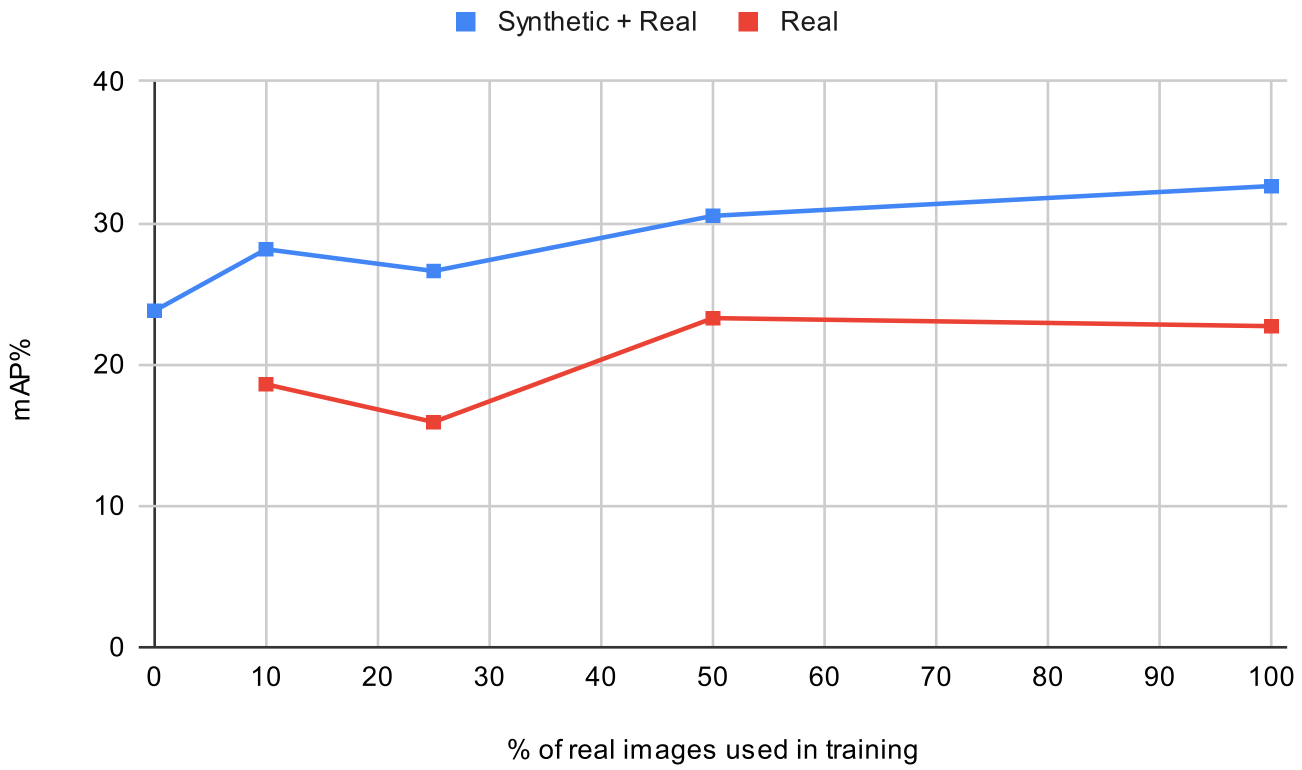


R. Leonardi, F. Ragusa, A. Furnari, G. M. Farinella (2022). Egocentric Human-Object Interaction Detection Exploiting Synthetic Data. In 21st International Conference on Image Analysis and Processing.

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

Can simulated data help?

Pretraining	Real Data%	mAP All
Synthetic	0	23.78
-	10	18.59
Synthetic	10	28.14
-	25	15.92
Synthetic	25	26.6
-	50	23.27
Synthetic	50	30.50
-	100	22.7
Synthetic	100	**32.61**

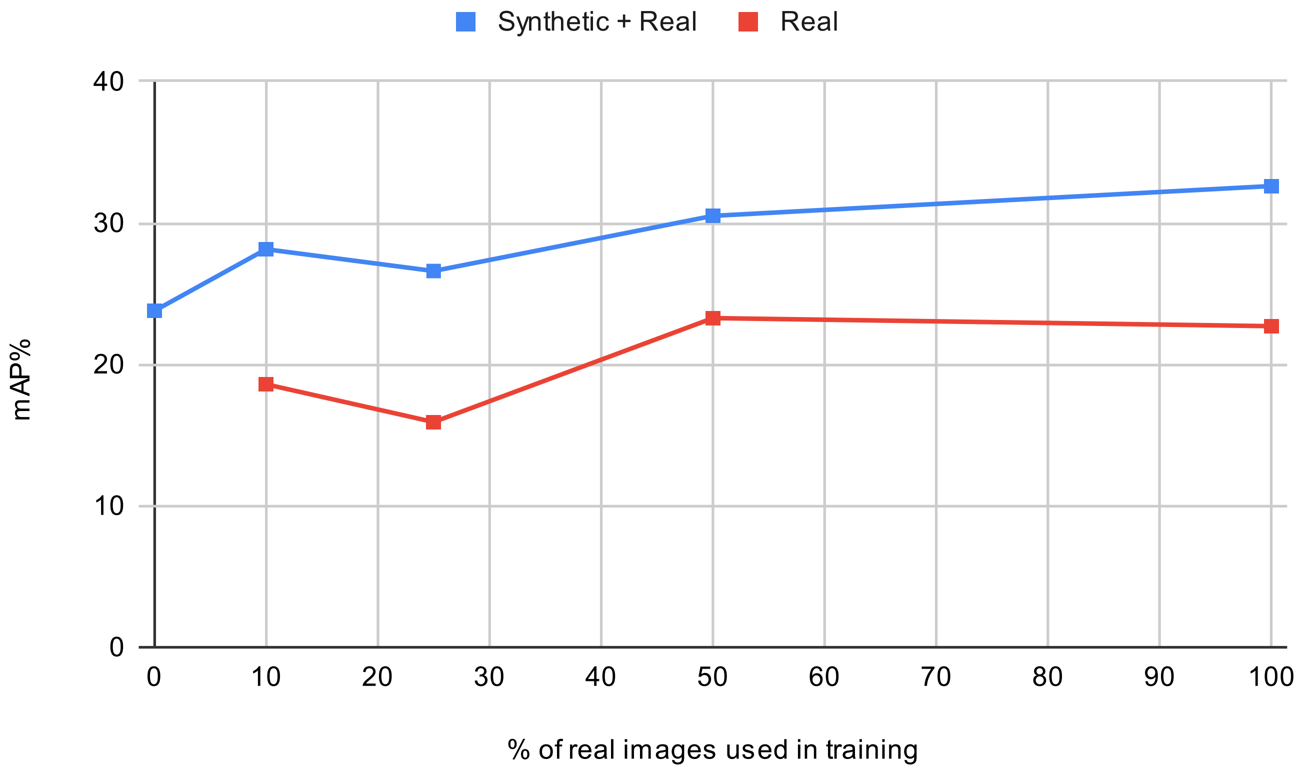


R. Leonardi, F. Ragusa, A. Furnari, G. M. Farinella (2022). Egocentric Human-Object Interaction Detection Exploiting Synthetic Data. In 21st International Conference on Image Analysis and Processing.

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

Can simulated data help?

Pretraining	Real Data%	mAP All
Synthetic	0	23.78
-	10	18.59
Synthetic	10	28.14
-	25	15.92
Synthetic	25	26.6
-	50	23.27
Synthetic	50	30.50
-	100	22.7
Synthetic	100	**32.61**

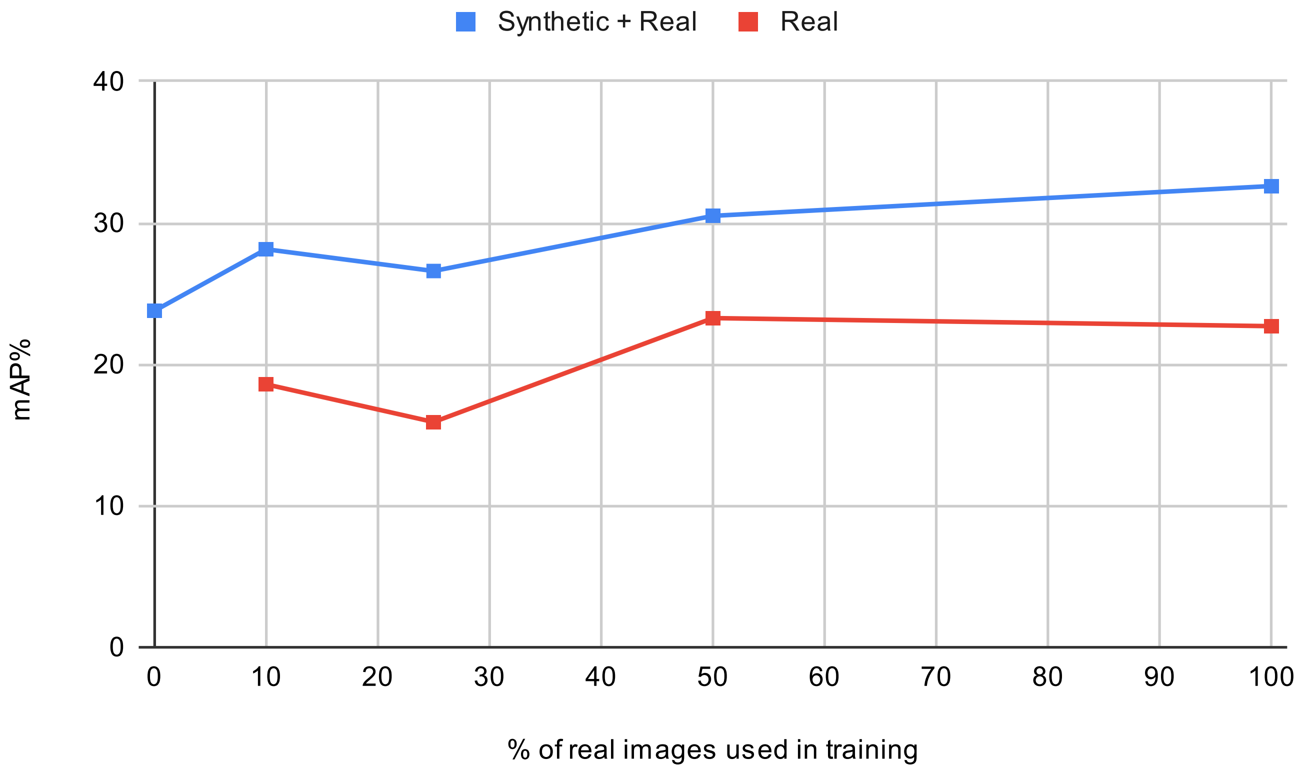


R. Leonardi, F. Ragusa, A. Furnari, G. M. Farinella (2022). Egocentric Human-Object Interaction Detection Exploiting Synthetic Data. In 21st International Conference on Image Analysis and Processing.

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

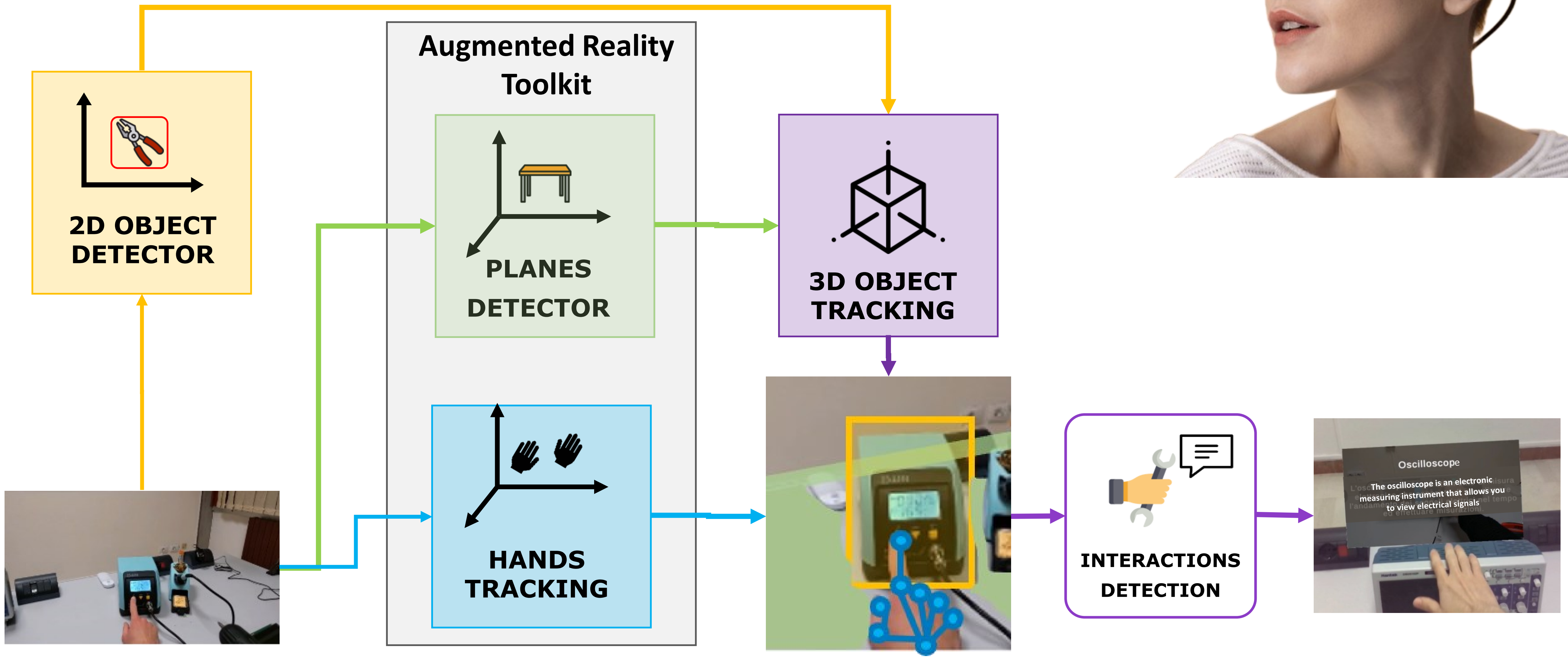
Can simulated data help?

Pretraining	Real Data%	mAP All
Synthetic	0	23.78
-	10	18.59
Synthetic	10	28.14
-	25	15.92
Synthetic	25	26.6
-	50	23.27
Synthetic	50	30.50
-	100	22.7
Synthetic	100	**32.61**



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.

Wearable Application



M. Mazzamuto, F. Ragusa, A. Resta, G. M. Farinella, Antonino Furnari (2023). A Wearable Device Application for Human-Object Interactions Detection. . In International Conference on Computer Vision Theory and Applications (VISAPP) .

Wearable Application

Understanding Actions

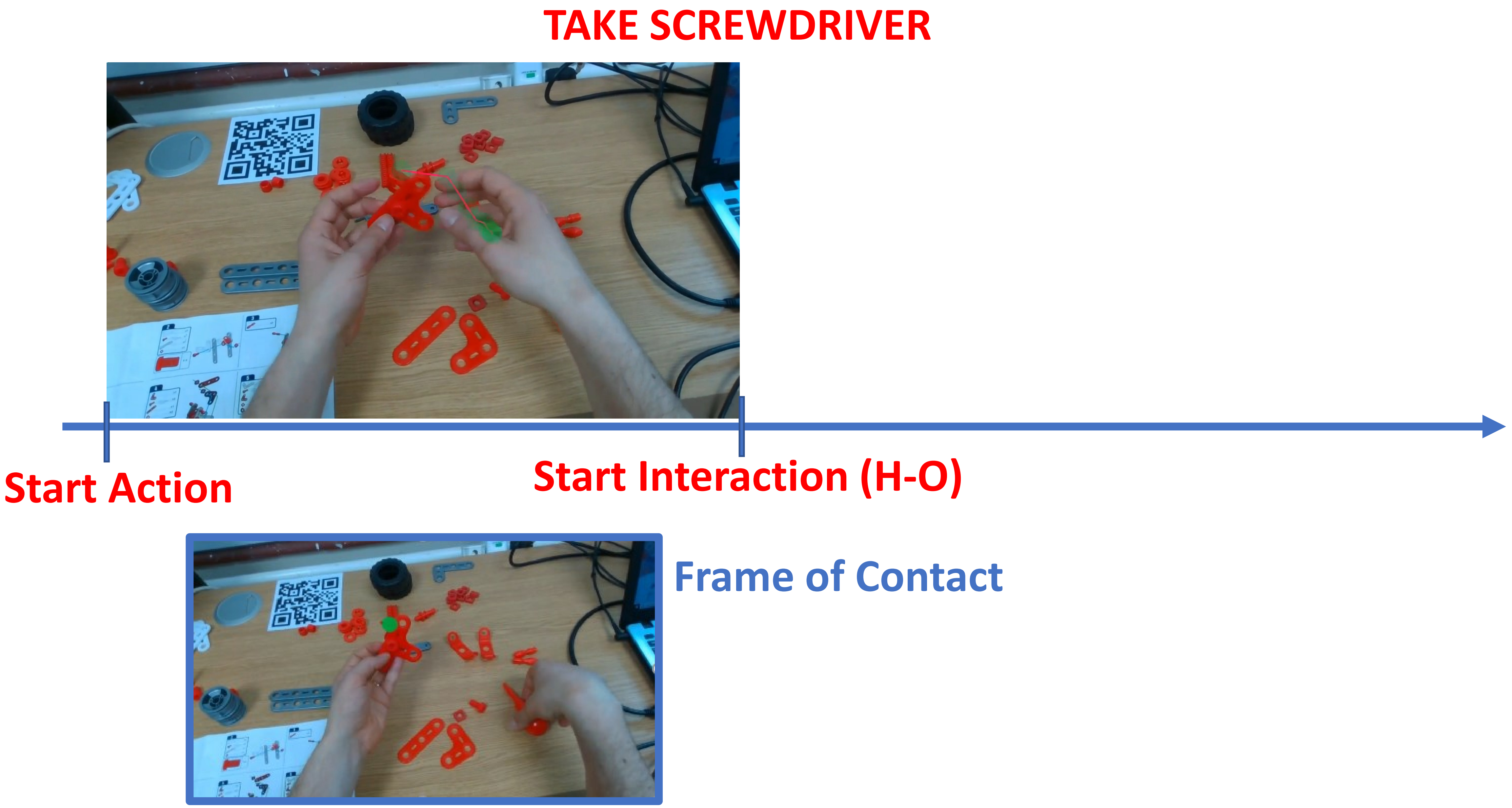
- Recognizing and detecting the actions performed by user allows to understand what happens in the video;
- This can be useful to:
  - Segment the video into coherent temporal units for:
    - Summarization;
    - Video understanding;
  - Understand the user’s goals to assist them;

Relation between Action and Interaction

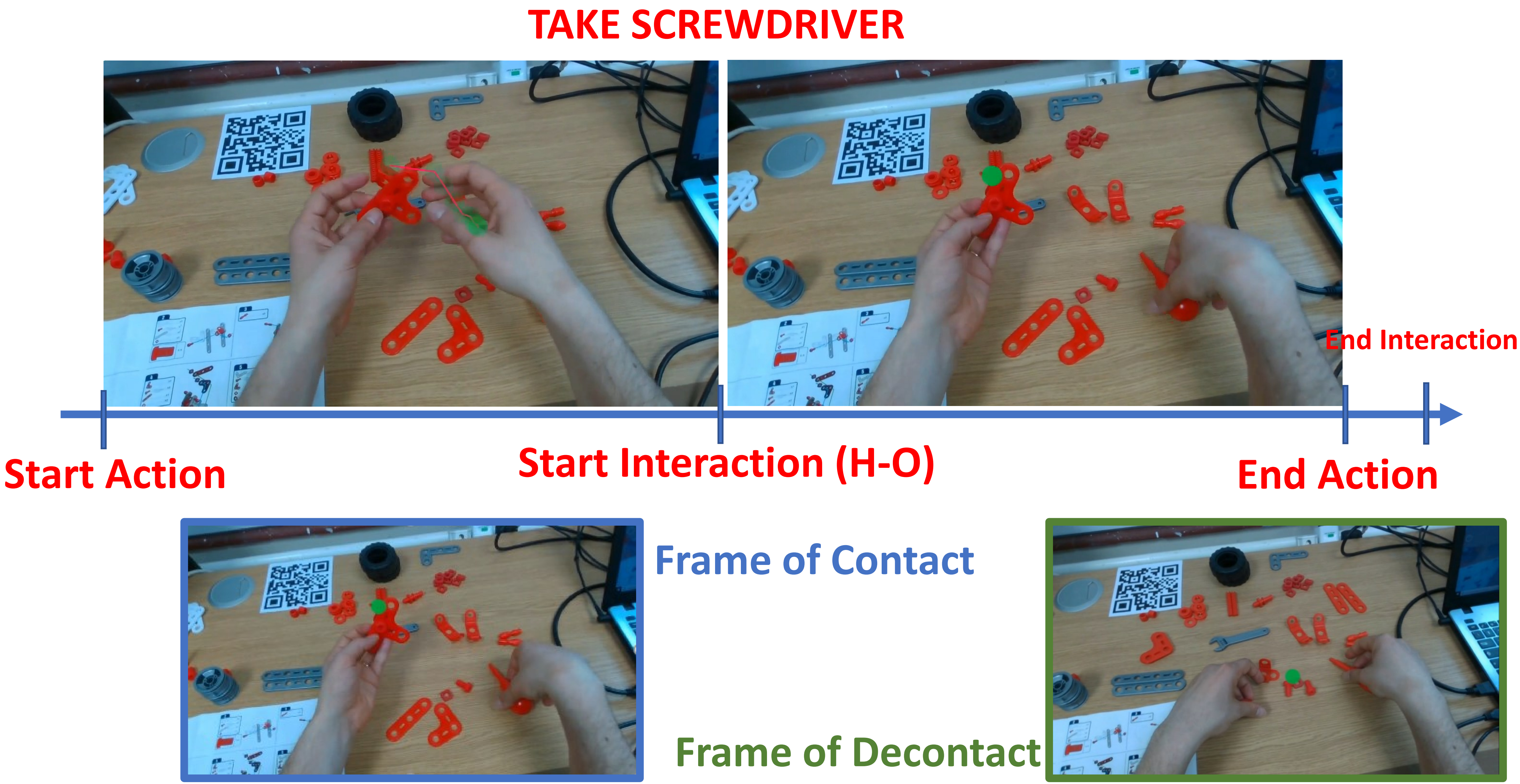
TAKE SCREWDRIVER

F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding (CVIU), 2023.

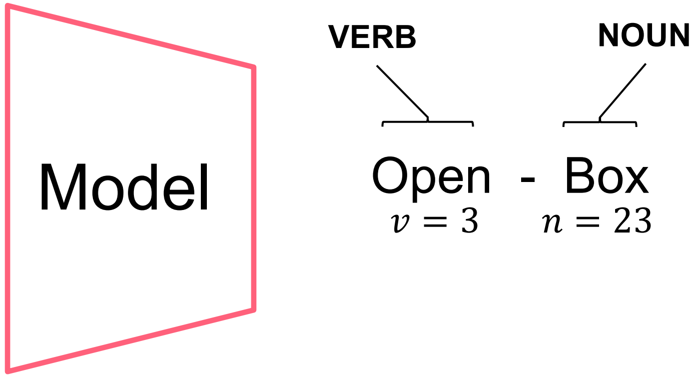
Relation between Action and Interaction

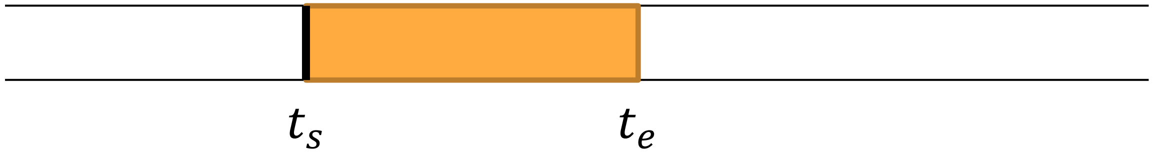


Relation between Action and Interaction



Action Recognition: Task



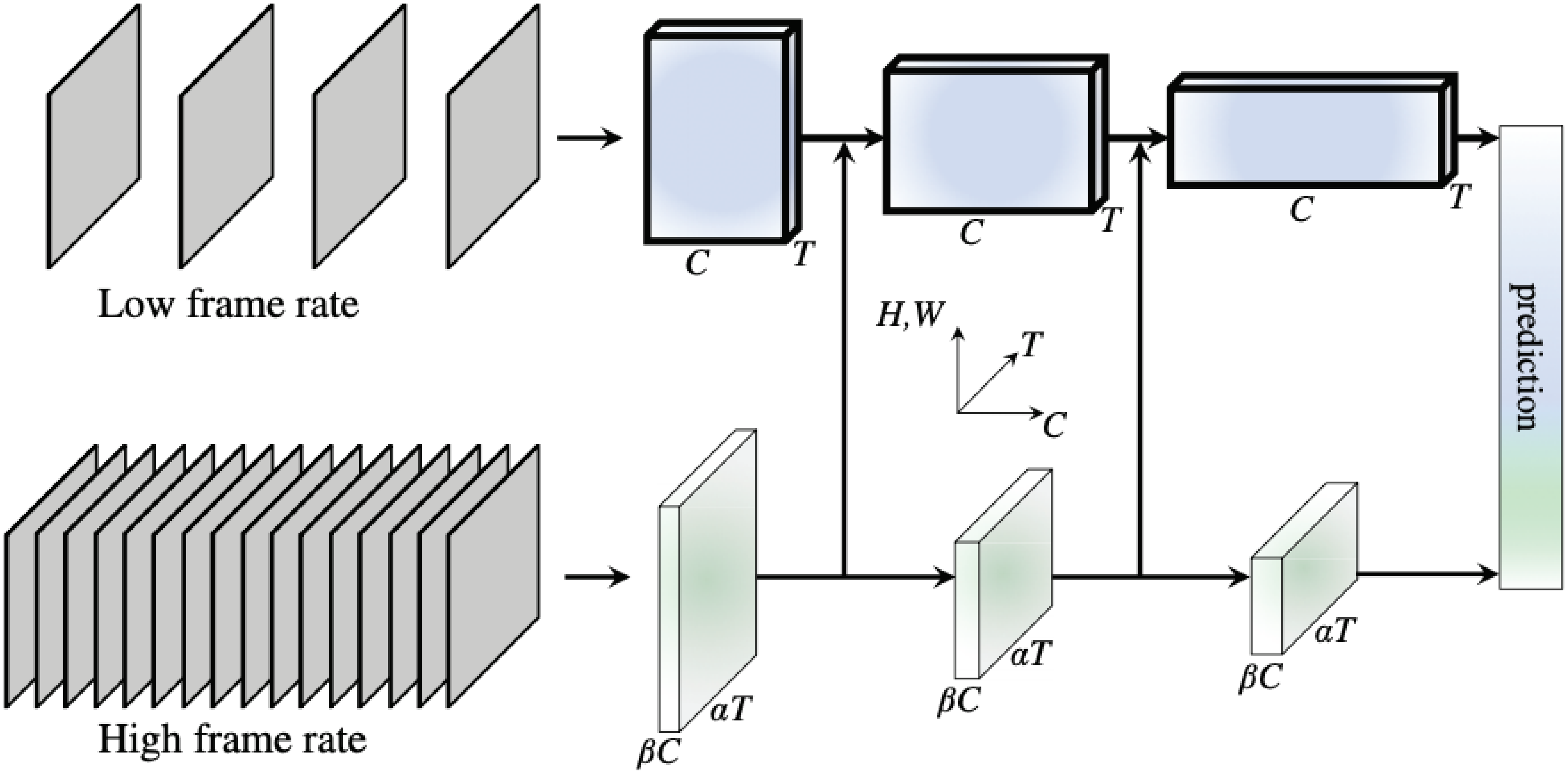


"observe a trimmed segment denoted by start and end time and classify the action present in the clip"

As defined in EPIC-KITCHENS-2020

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

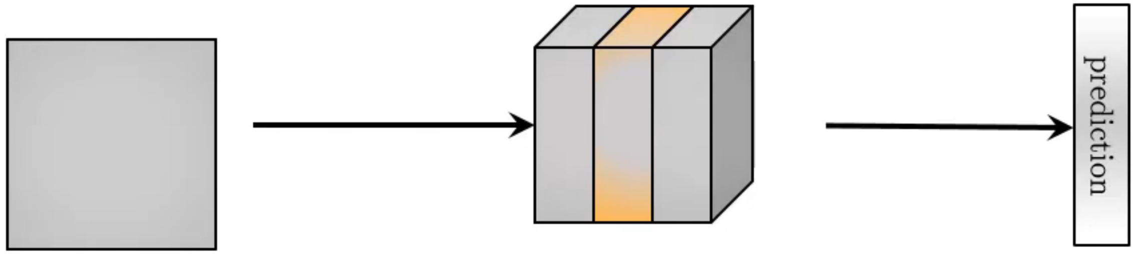
SlowFast Networks for Video Recognition



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).

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

X3D: Expanding Architectures for Efficient Video Recognition



- X-Fast
- X-Temporal
- X-Spatial
- X-Depth
- X-Width
- X-Bottleneck

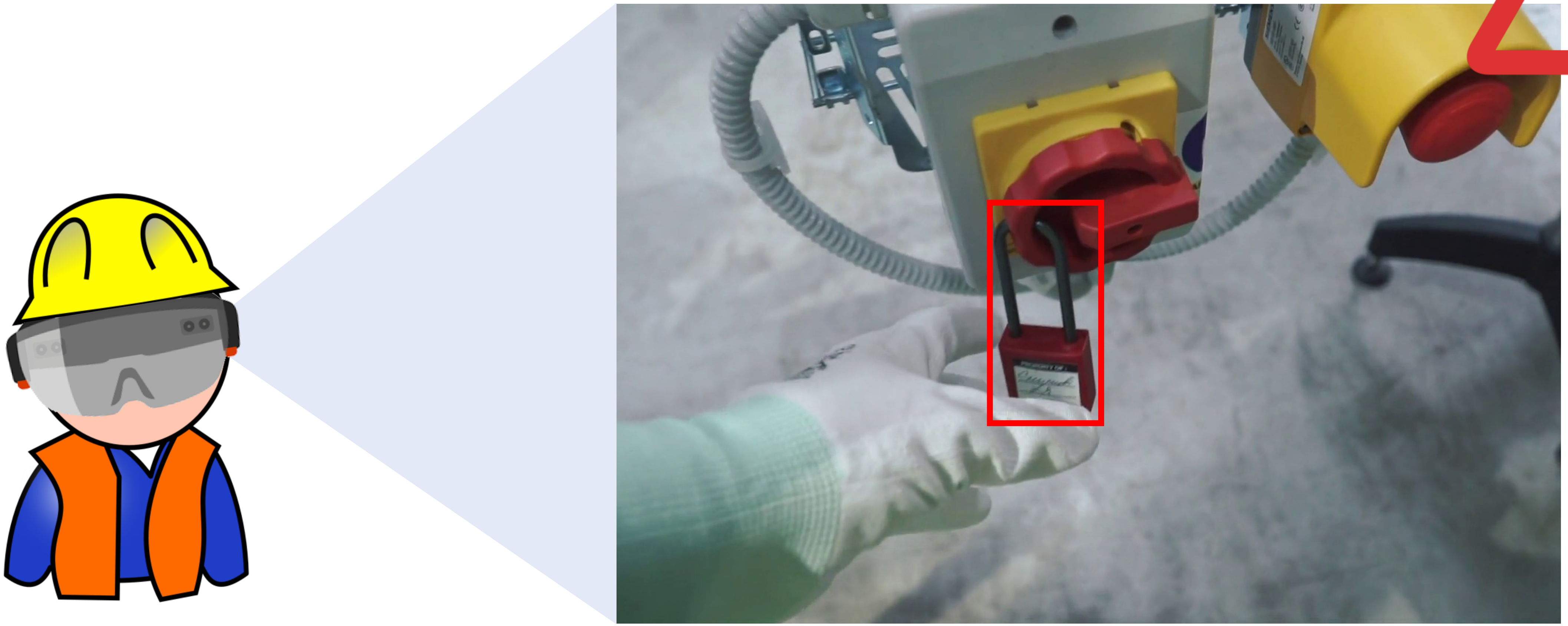
Feichtenhofer, C. (2020). X3D: Expanding Architectures for Efficient Video Recognition. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 200-210.

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**

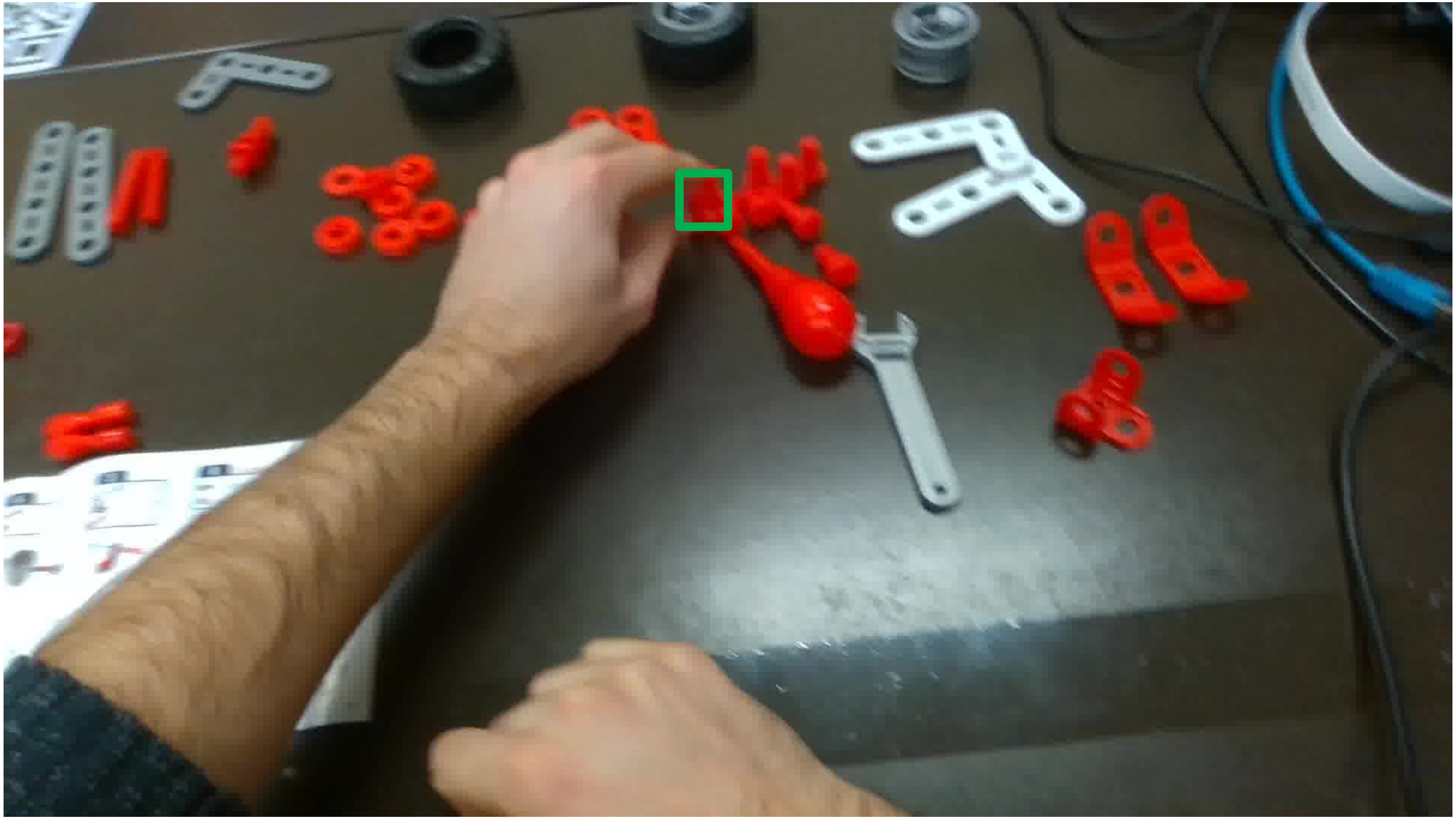


Next-Active Objects Detection

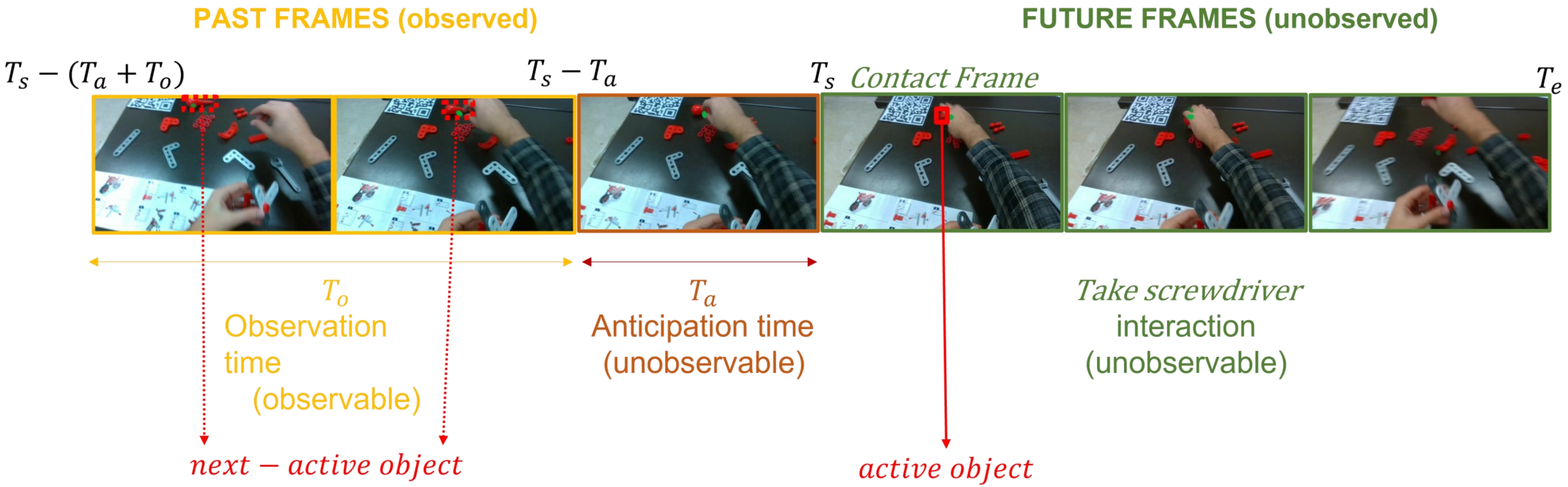
Next-Active Objects Detection

Next-Active Objects Detection

Next-Active Objects Detection



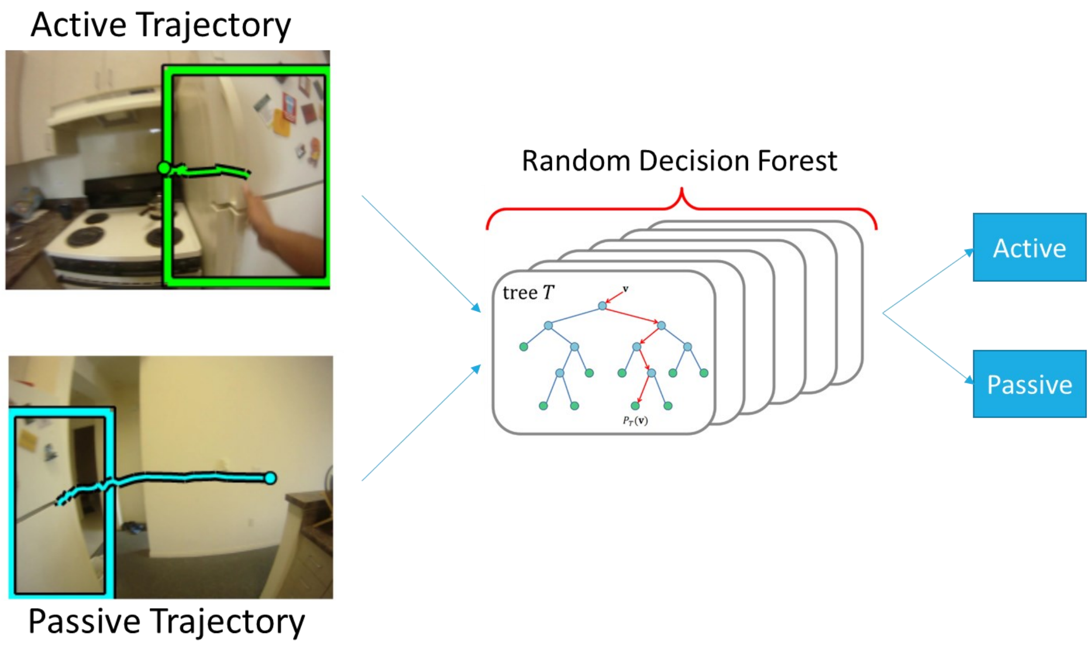
Next-Active Objects Detection

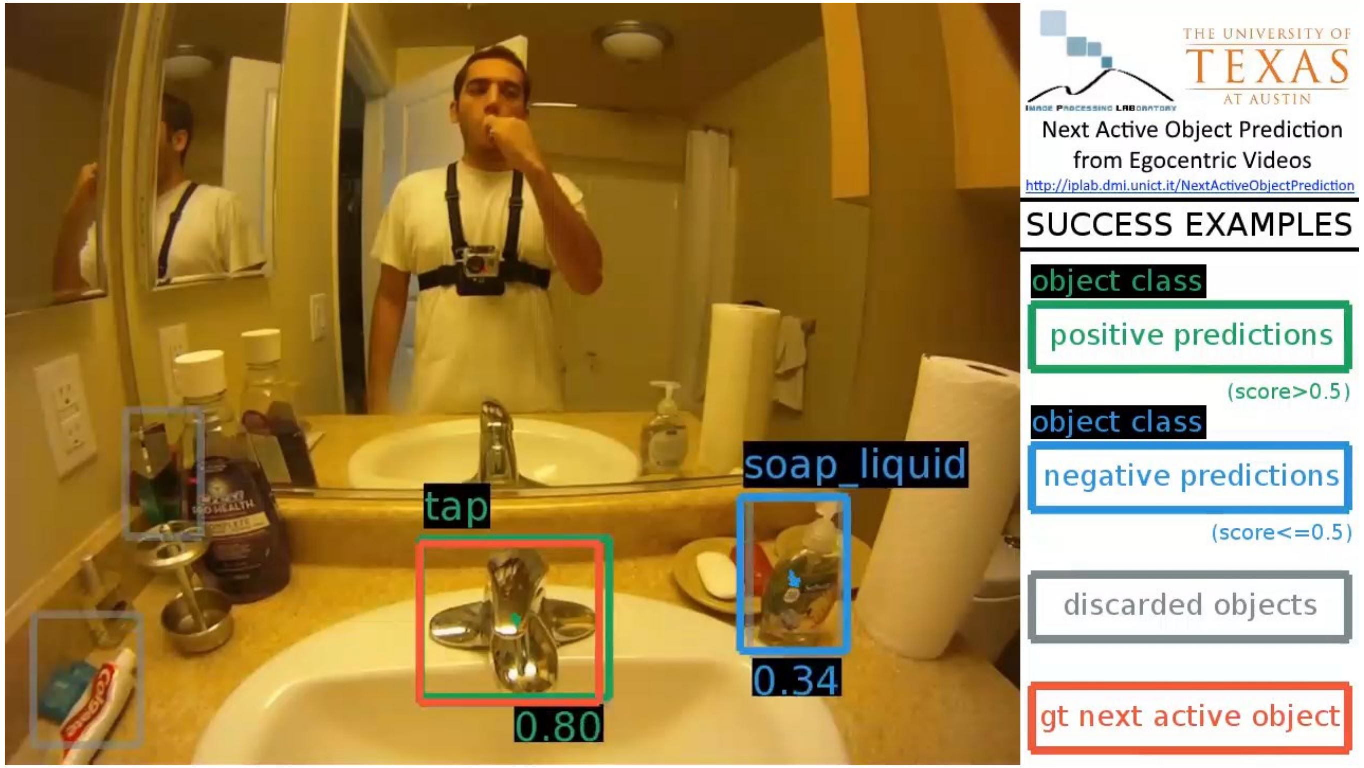


F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding (CVIU), 2023.

Anticipation – Next-Active-Objects

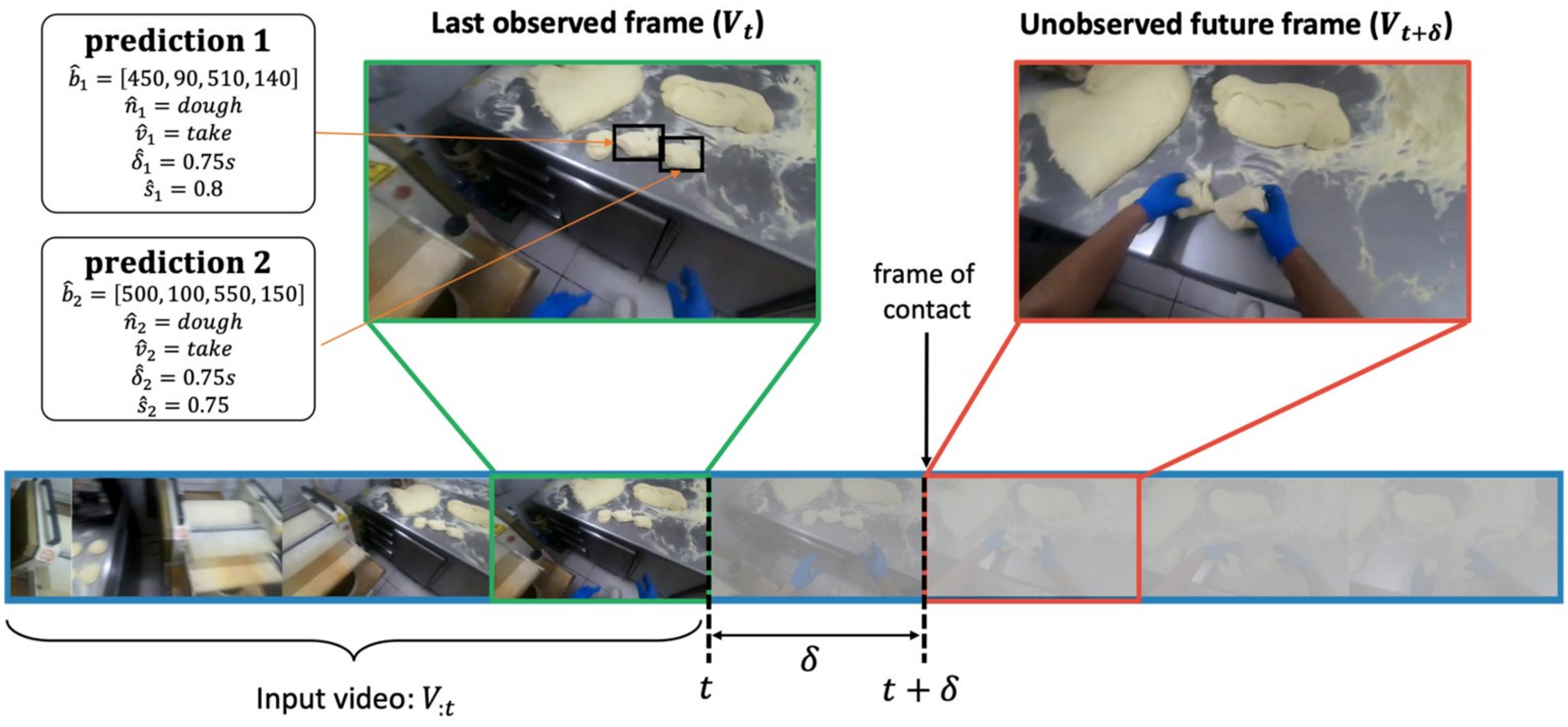
Use egocentric object trajectories to distinguish passive from next-active-objects (i.e., those which will be used soon by the user).





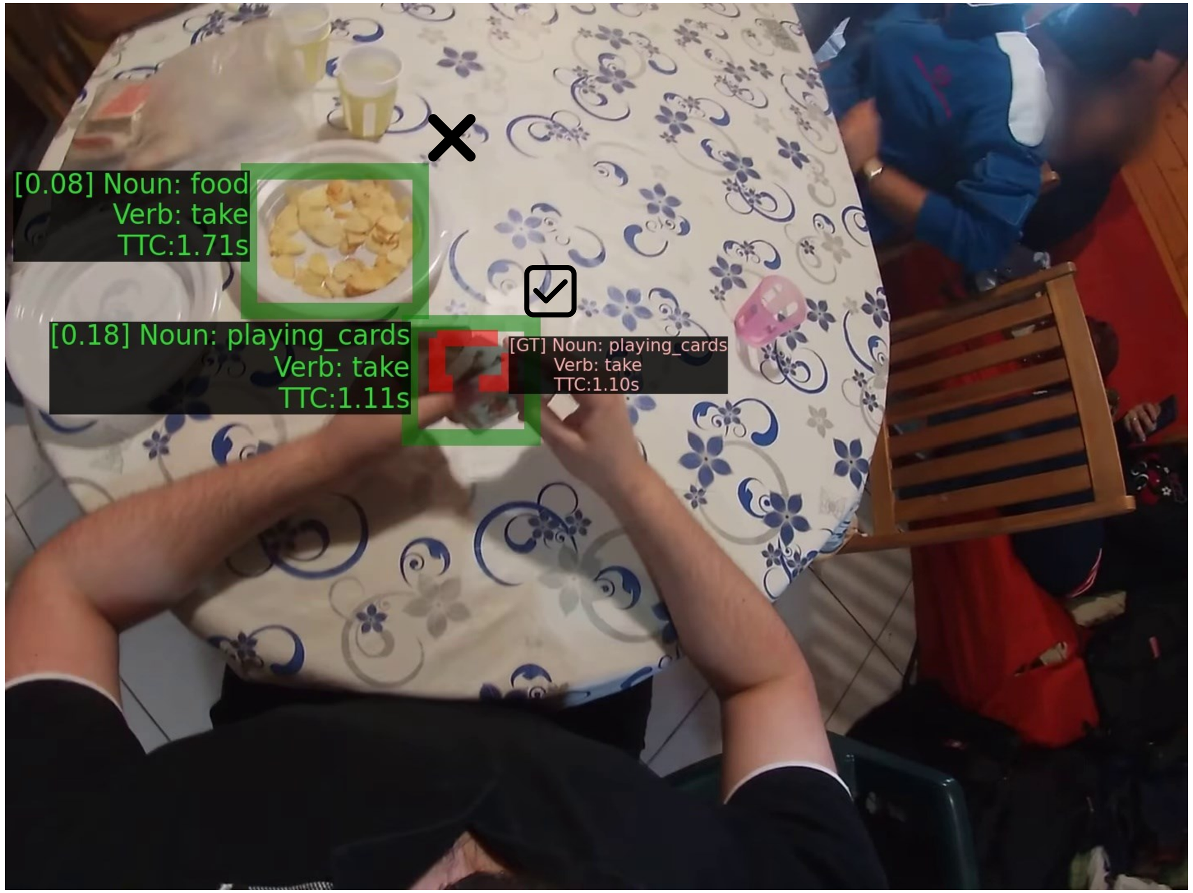
A. Furnari, S. Battiato, K. Grauman, G. M. Farinella, Next-Active-Object Prediction from Egocentric Videos, Journal of Visual Communication and Image Representation, 2017

Short Term Object Interaction Anticipation (STA)

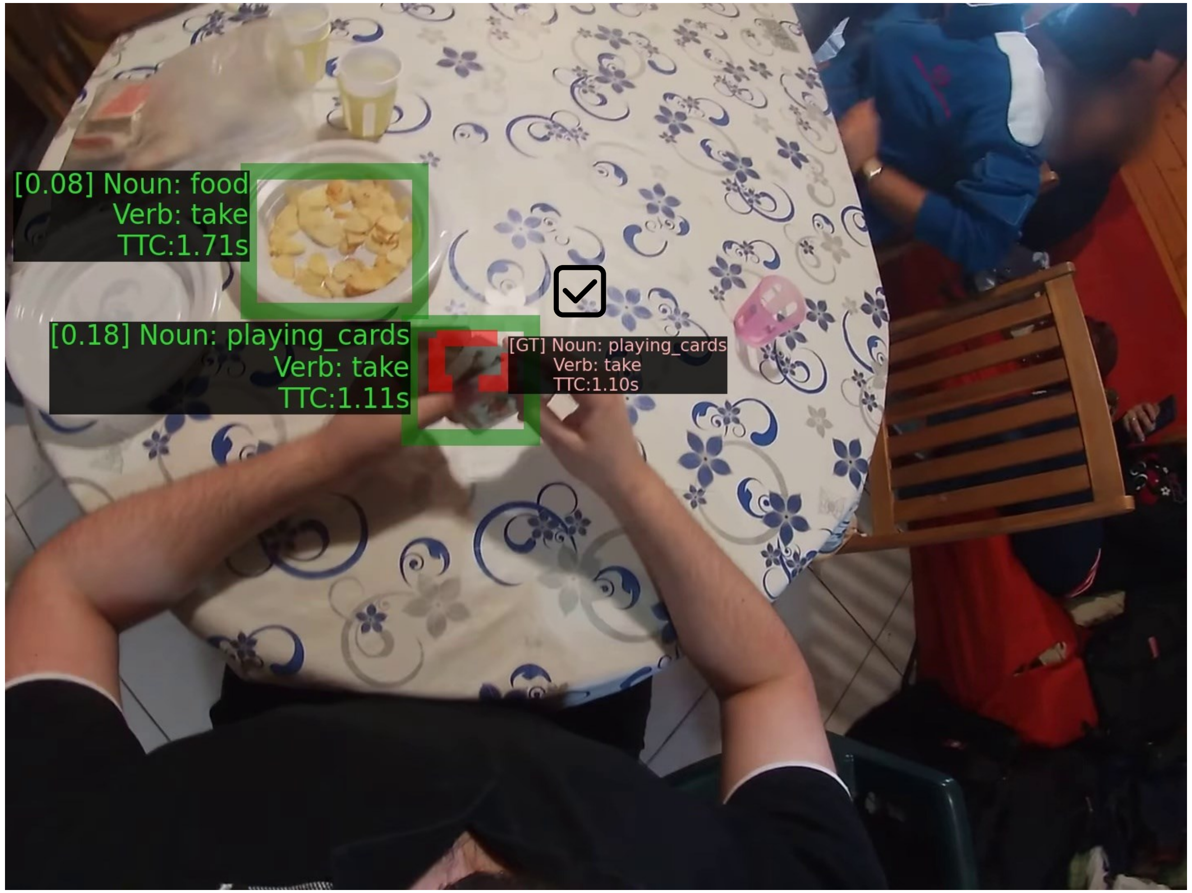


Short Term Object Interaction Anticipation (STA)

Top-5 mAP “discounts” up to 4 false positives per GT box

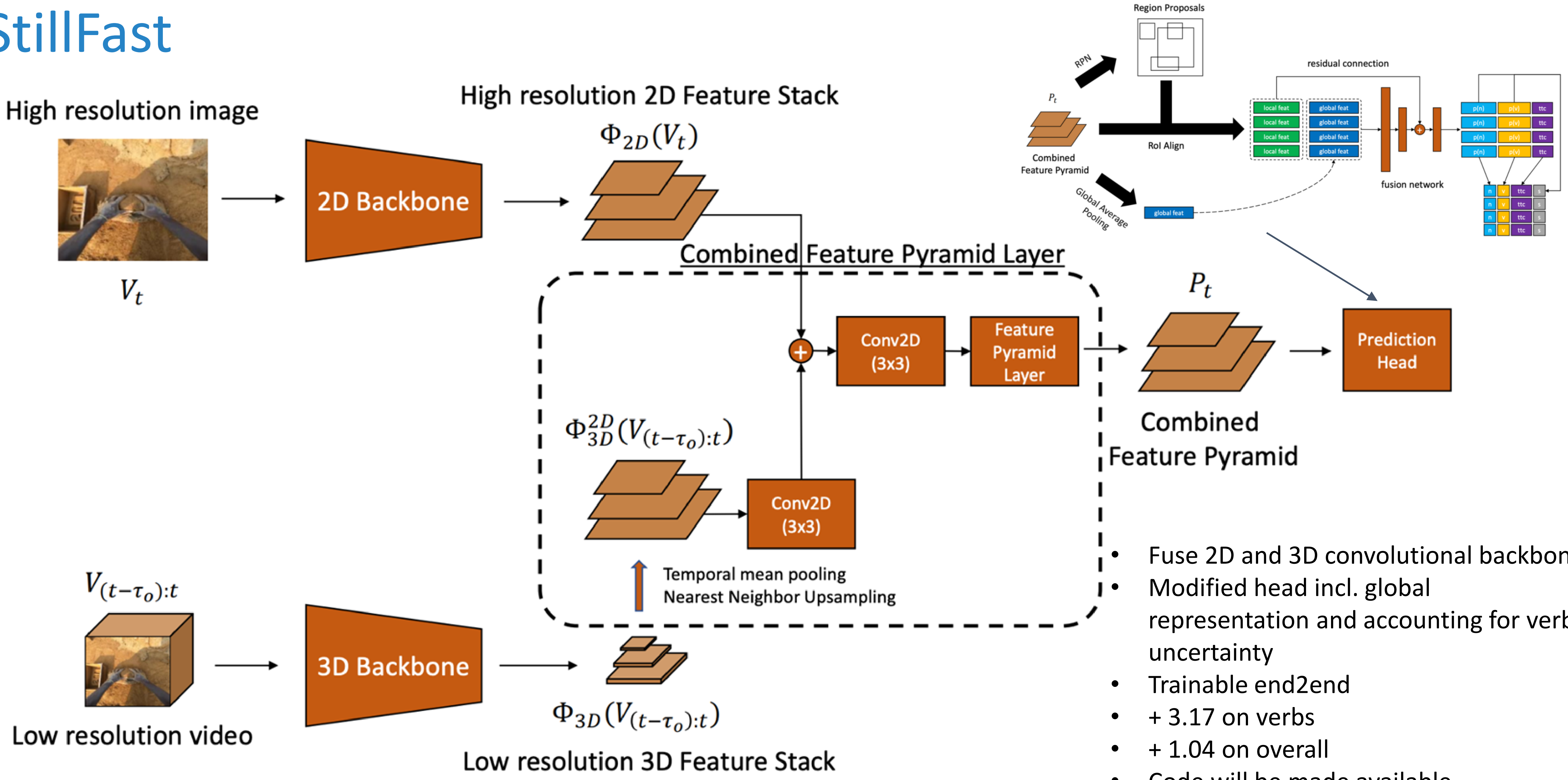


mAP: 1 True Positive + 1 False Positive



Top-5 mAP: 1 True Positive

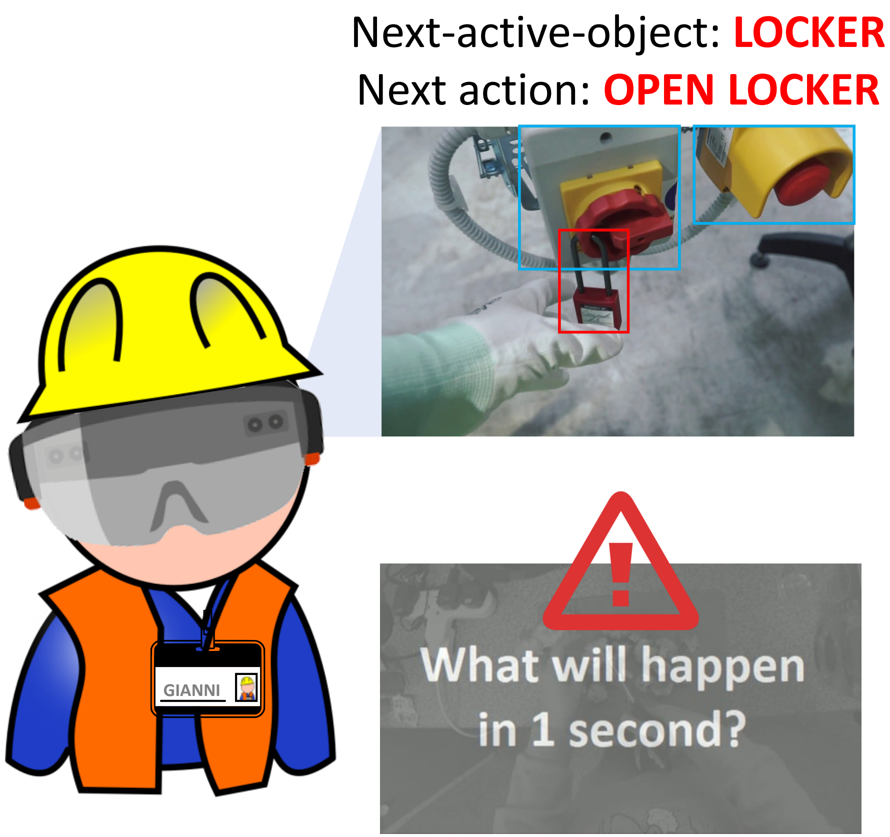
StillFast



- Fuse 2D and 3D convolutional backbone
- Modified head incl. global representation and accounting for verb uncertainty
- Trainable end2end
- + 3.17 on verbs
- + 1.04 on overall
- Code will be made available

F. Ragusa, G. M. Farinella, A. Furnari. StillFast: An End-to-End Approach for Short-Term Object Interaction Anticipation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, 2023.

Can we bring egocentric vision to industry?



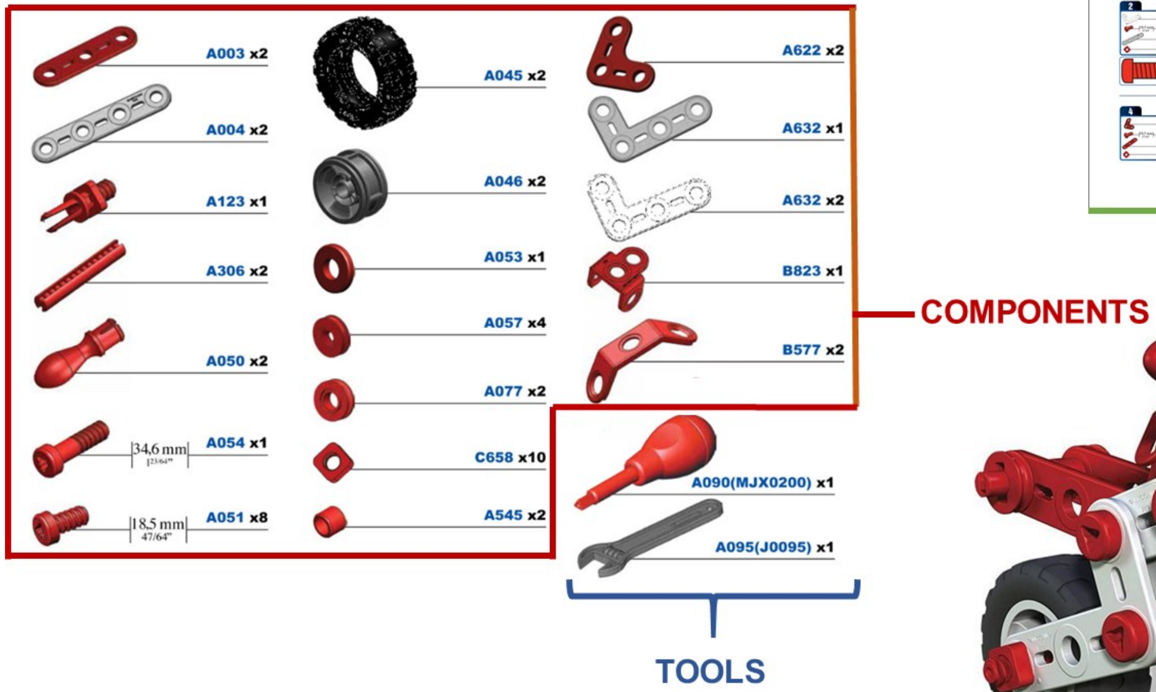
- The factory is a natural place for a wearable assistant;
- Closed-world assumption;
- Current research has considered different scenarios;
- No datasets in industrial-like scenarios;

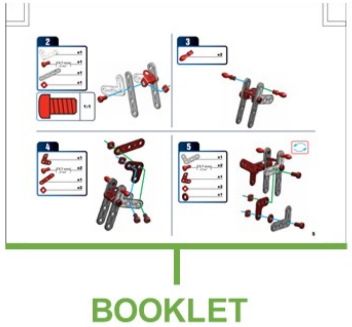
The MECCANO Dataset

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.



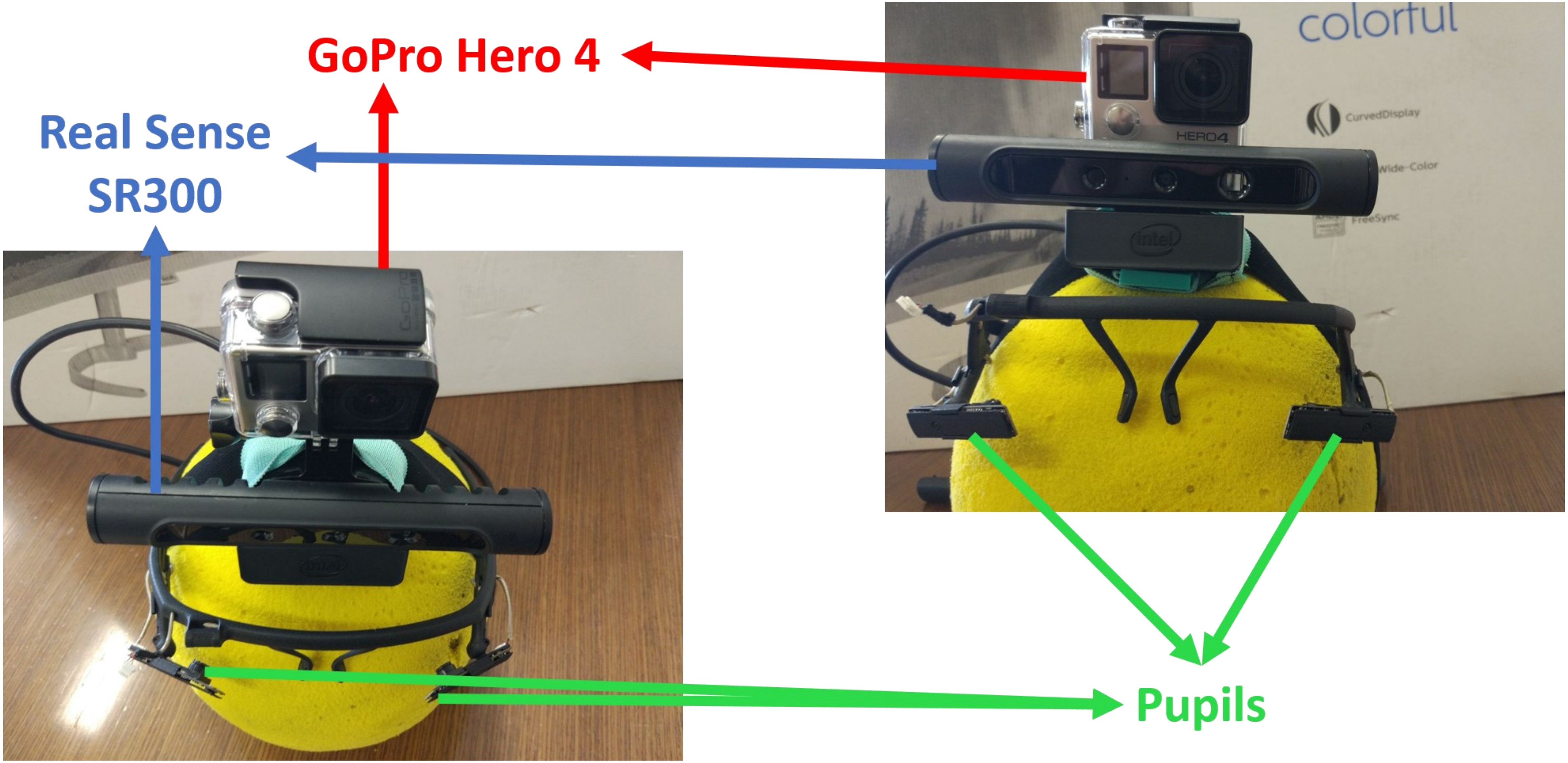


The scenario is industrial-like, with subjects undertaking interactions with tiny objects and tools in a sequential fashion to reach a goal.

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.

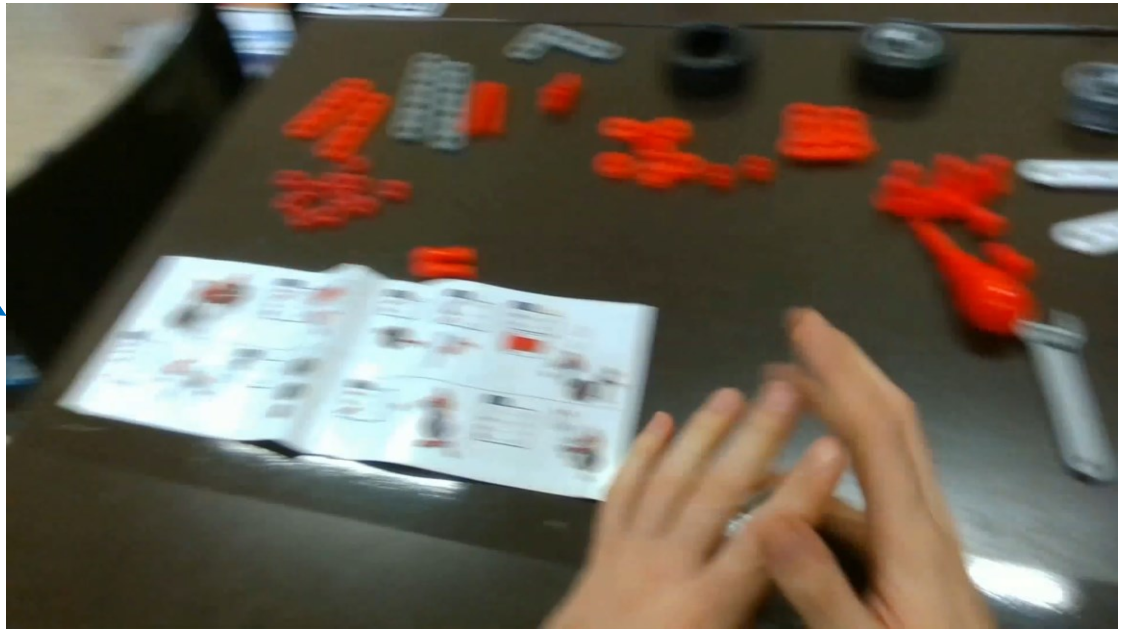
F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding (CVIU), 2023.

Data Collection



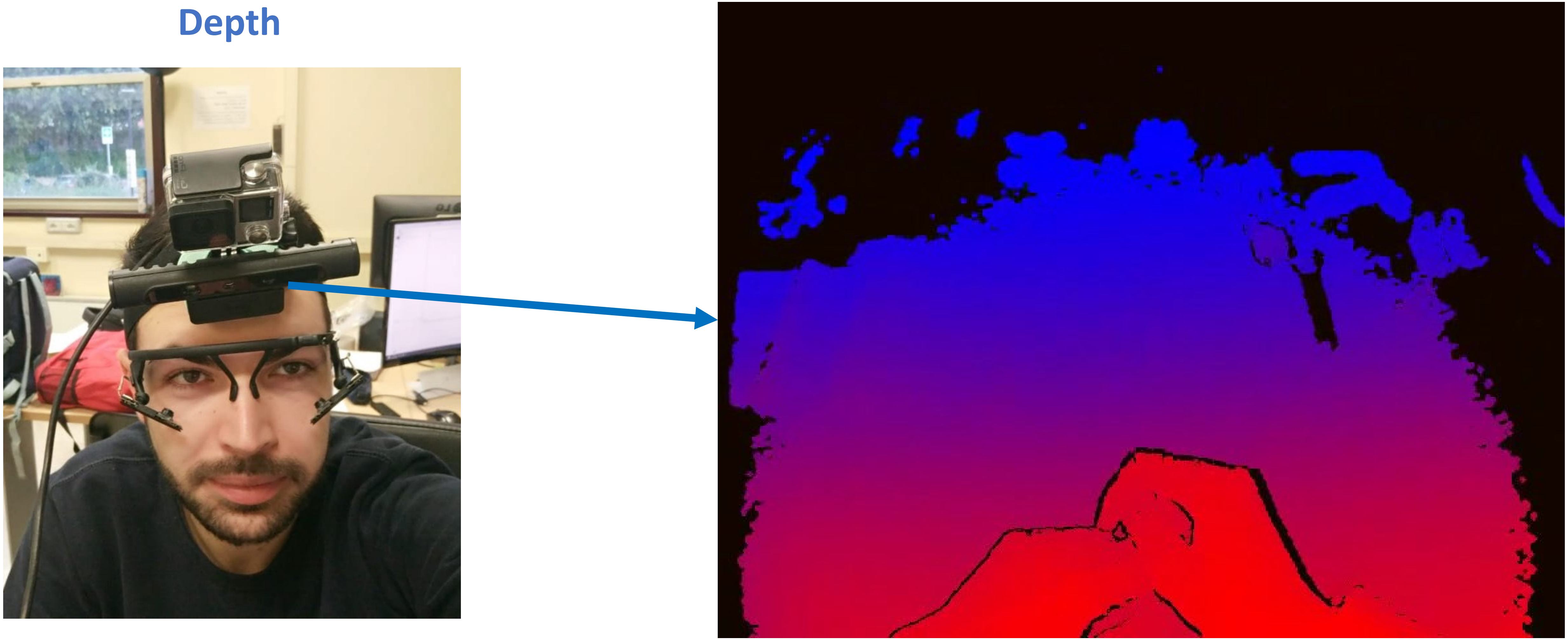
The MECCANO Dataset

RGB



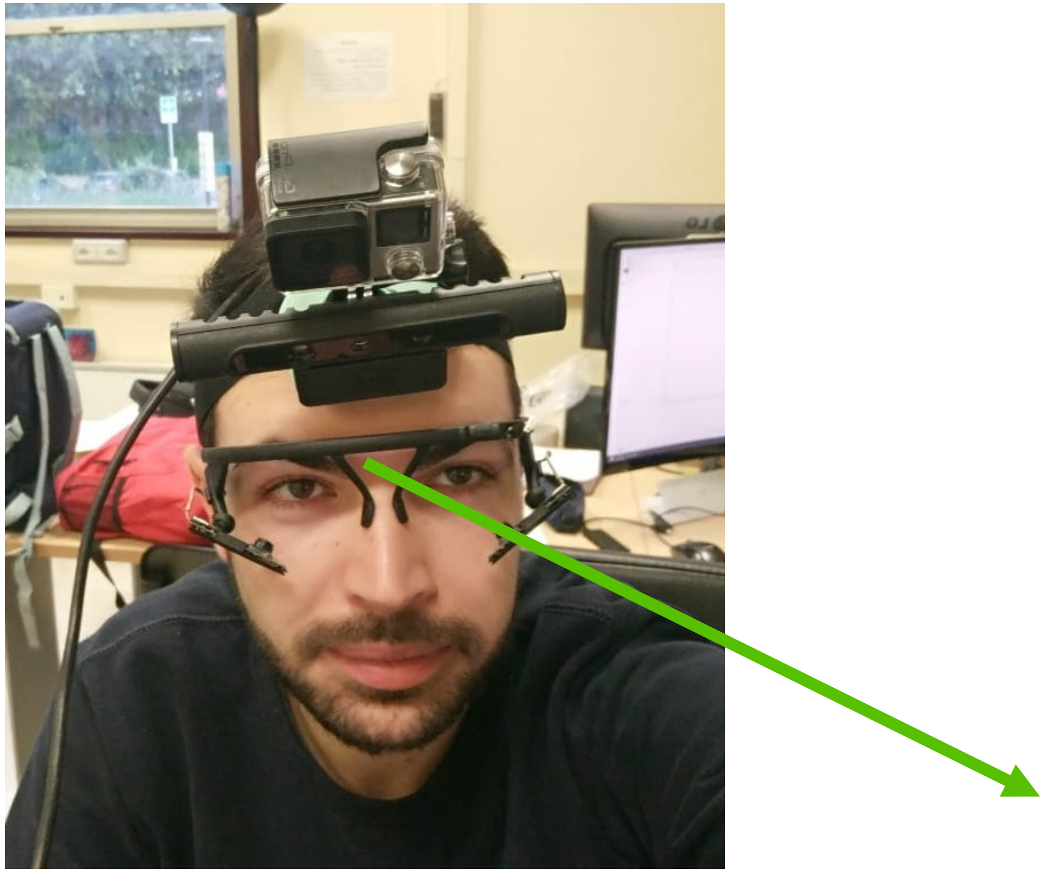
The MECCANO Dataset

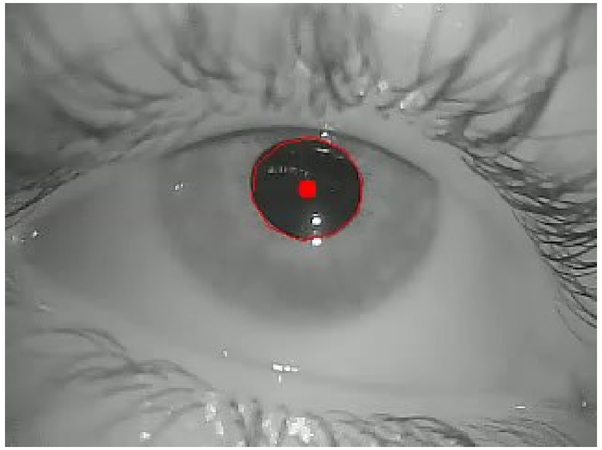
Depth

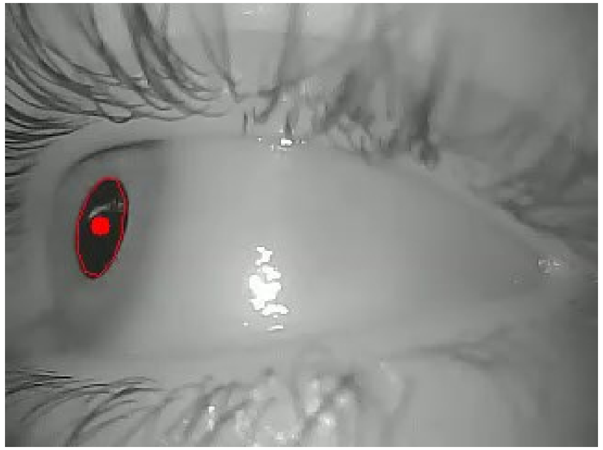


The MECCANO Dataset

Gaze







The MECCANO Dataset: Statistics

20 Subjects

3 Modalities

20 min. avg. Video length

5 Tasks

8858 Segments

64349 Boxes

20 Objects

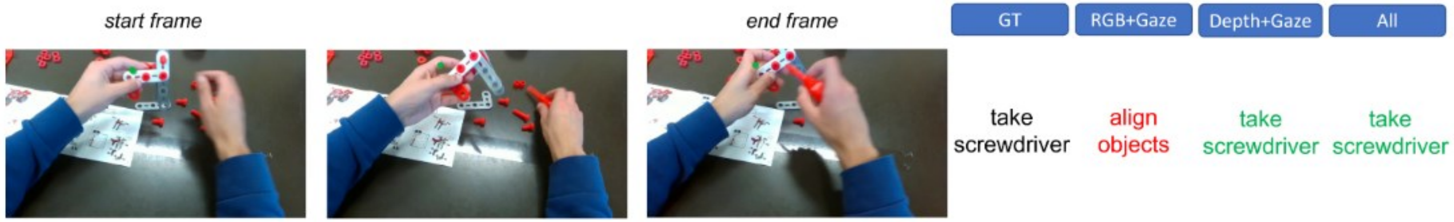
12 Verbs

61 Actions

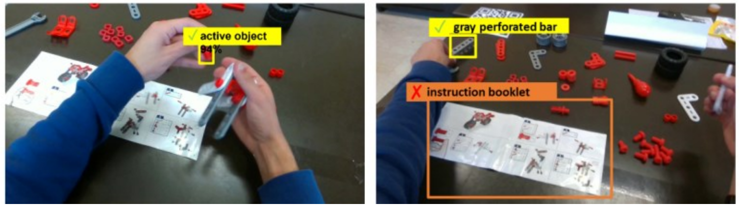
F. Ragusa, A. Furnari, S. Livatino, G. M. Farinella. The MECCANO Dataset: Understanding Human-Object Interactions from Egocentric Videos in an Industrial-like Domain. WACV, 2021 (https://arxiv.org/abs/2010.05654). ORAL.

The MECCANO Dataset: Tasks

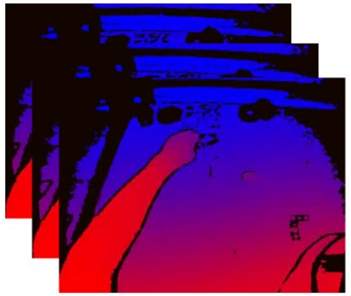
1) Action Recognition

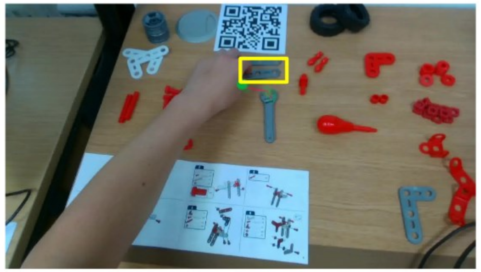


2) Active Object Detection and Recognition



3) EHOI Detection


<take>

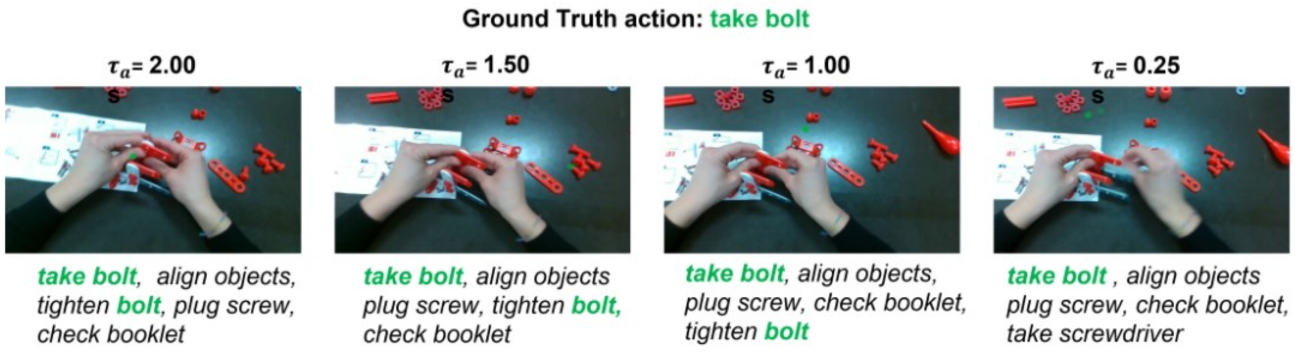

<gray perforated bar>

F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding (CVIU), 2023.

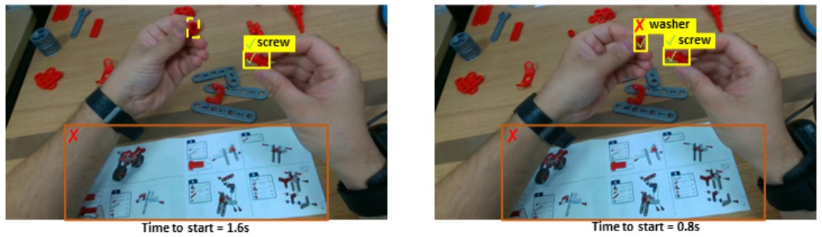
The MECCANO Dataset: Tasks

4) Egocentric Gaze Estimation

5) Action Anticipation



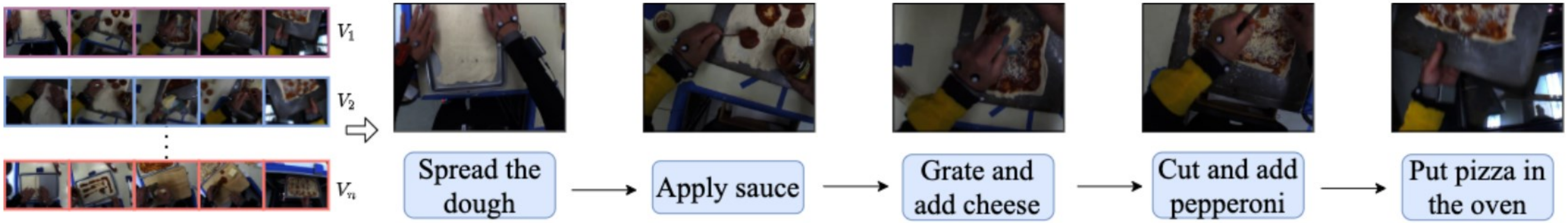
6) Next-Active Object (NAO) Detection



F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding (CVIU), 2023.

Procedural Learning

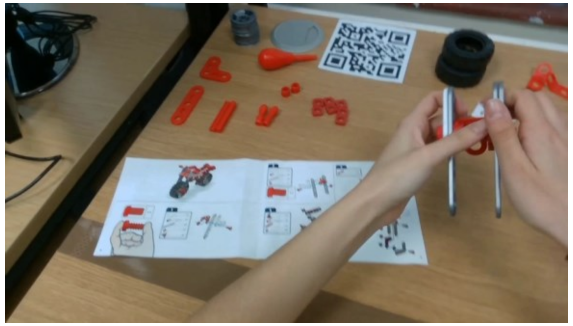
Given multiple videos of a task, the goal is to identify the key-steps and their order to perform the task.

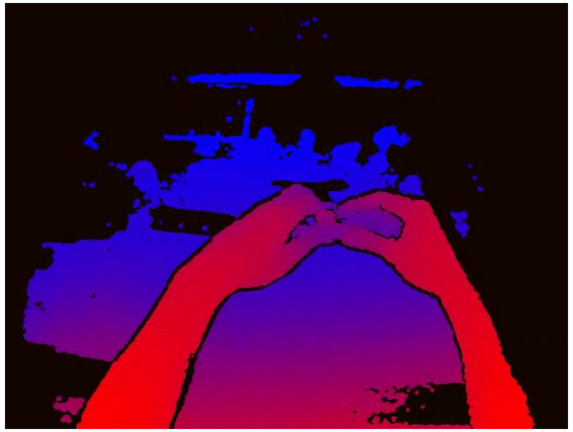


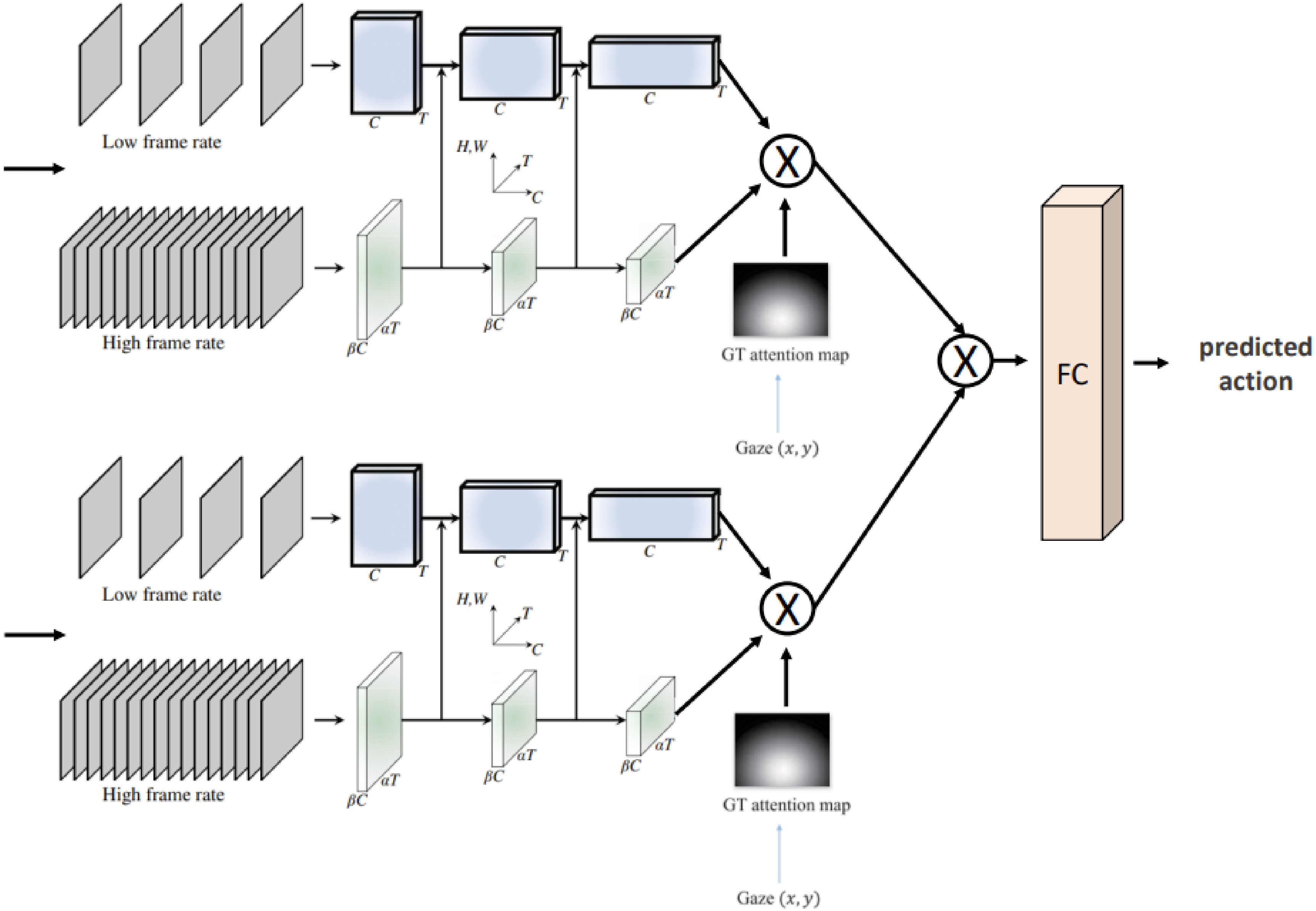
1) EgoProceL (proposed)
2) CMU-MMAC
3) EGTEA Gaze+
4) MECCANO
5) EPIC-Tent

B. Siddhant, A. Chetan, C. V. Jawahar, My View is the Best View: Procedure Learning from Egocentric Videos. In European Conference on Computer Vision (ECCV), 2022.

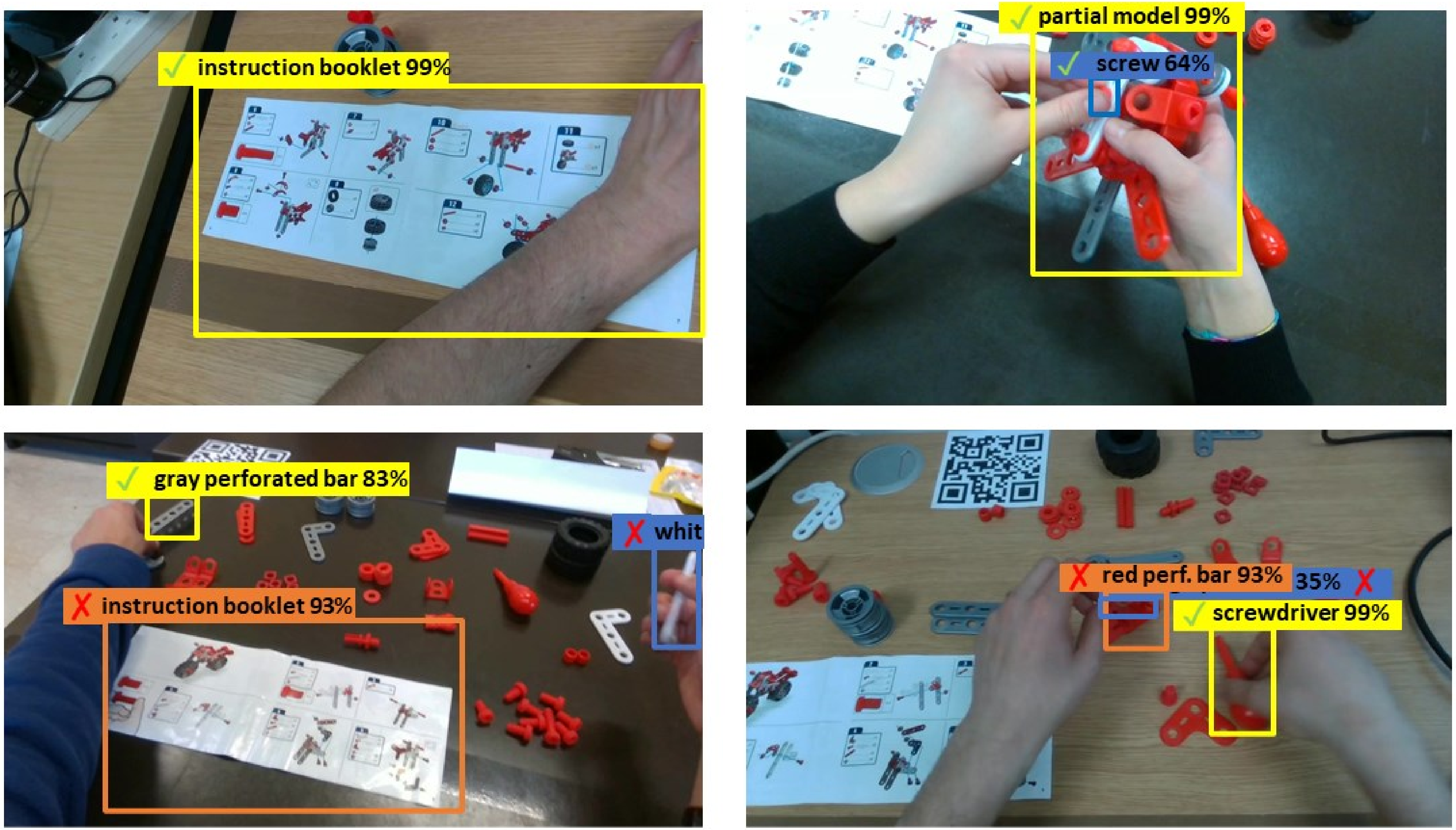
Action Recognition





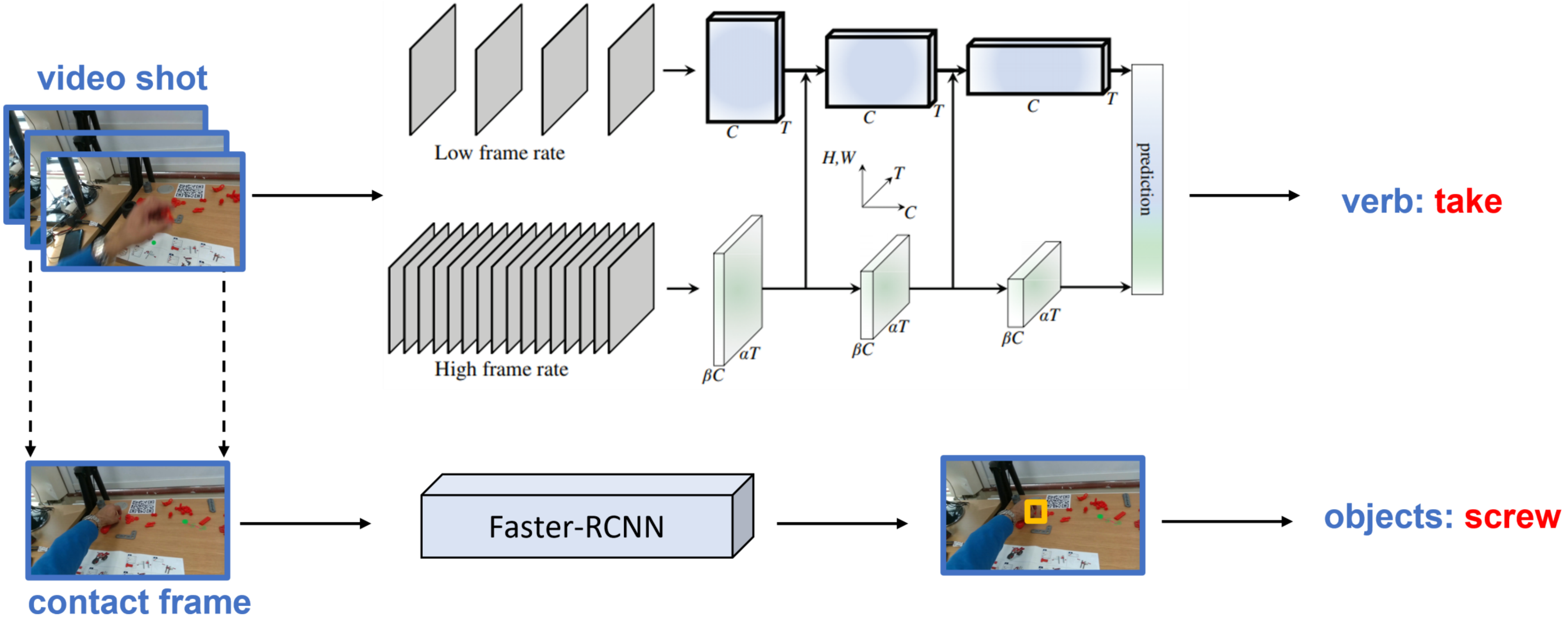


Active Object Detection and Recognition



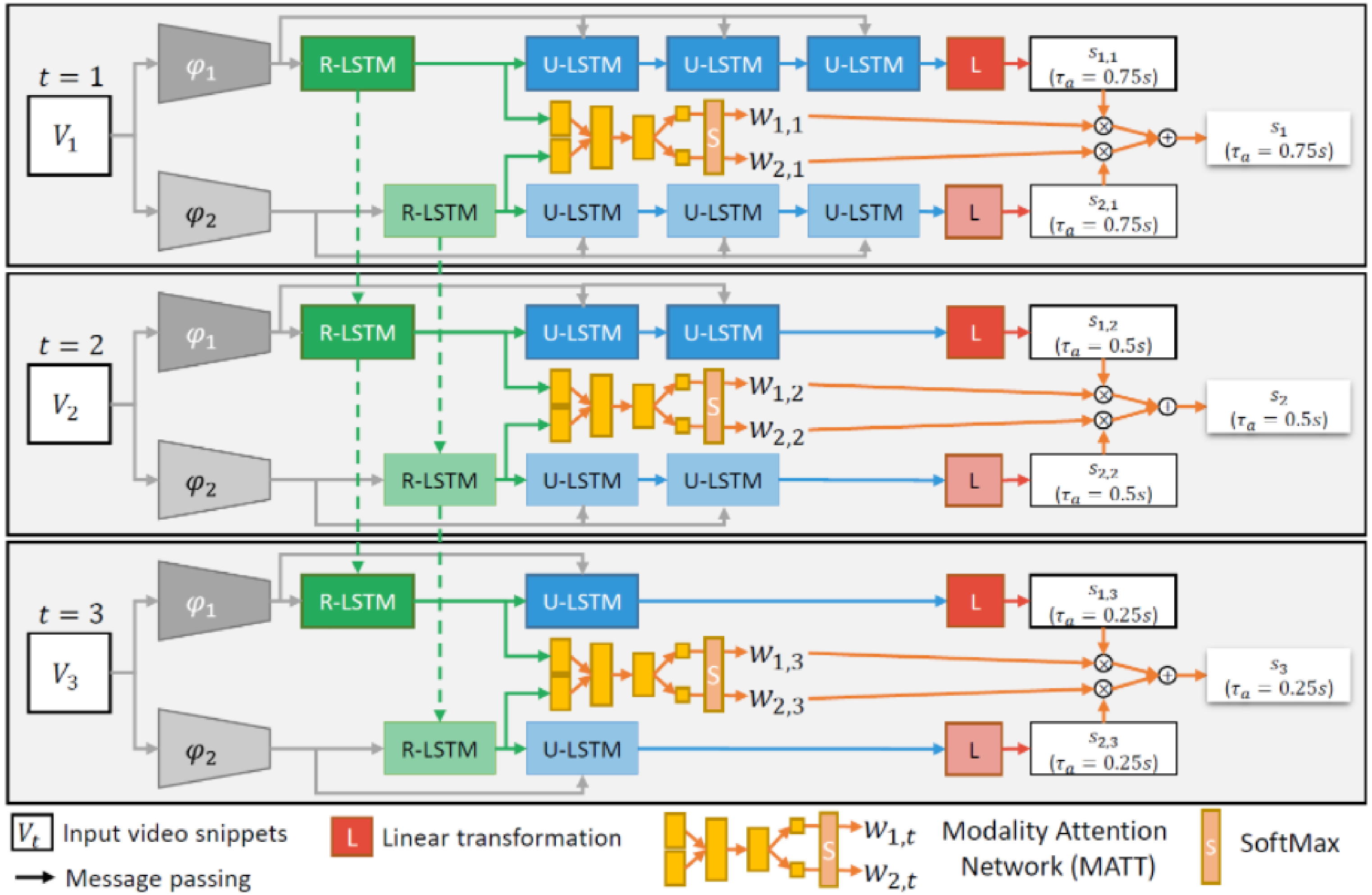
F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding (CVIU), 2023.

EHOI Detection



F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding (CVIU), 2023.

Action Anticipation



Modalities:

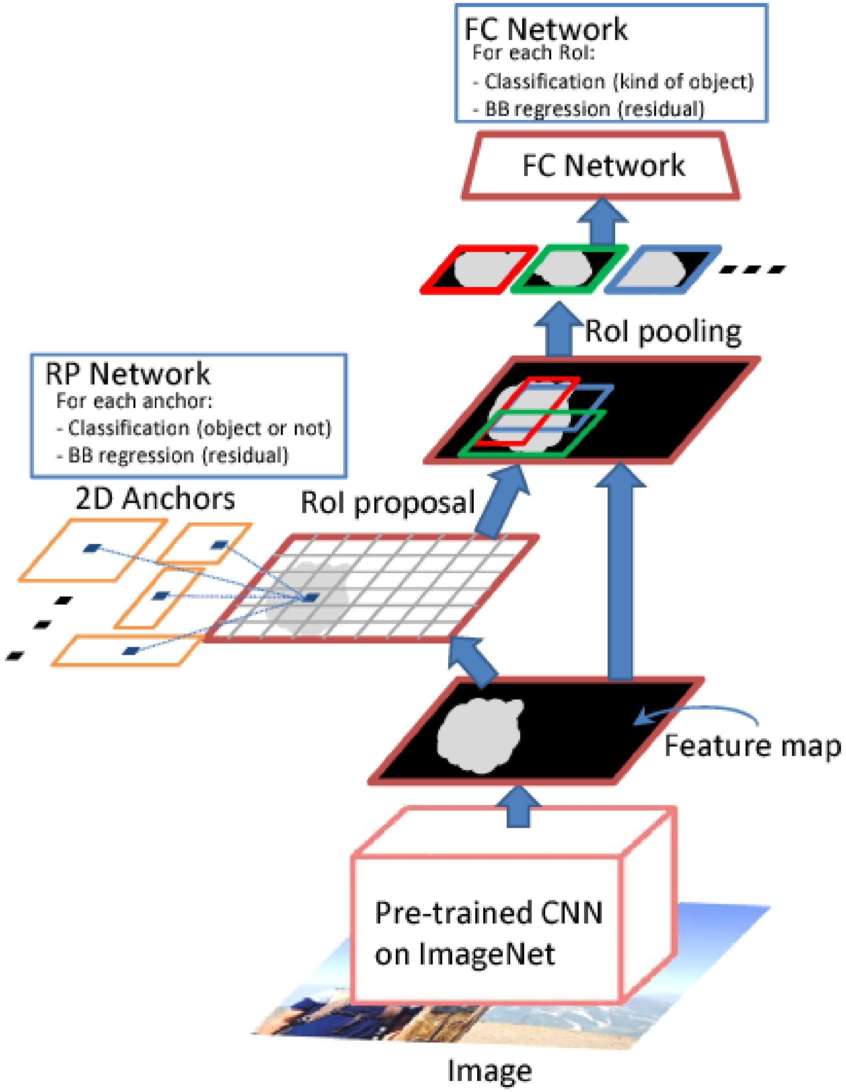
- RGB
- Optical Flow
- Objects

Our Modalities:

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

F. Ragusa, A. Furnari, G. M. Farinella. MECCANO: A Multimodal Egocentric Dataset for Humans Behavior Understanding in the Industrial-like Domain. Computer Vision and Image Understanding (CVIU), 2023.

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/

SICILIAE STVDIVM GENERALE · 1434 ·

IMAGE PROCESSING LABORATORY

Innovation

Microsoft HoloLens 2

NREAL LIGHT

Magic Leap 2

VUZIX BLADE

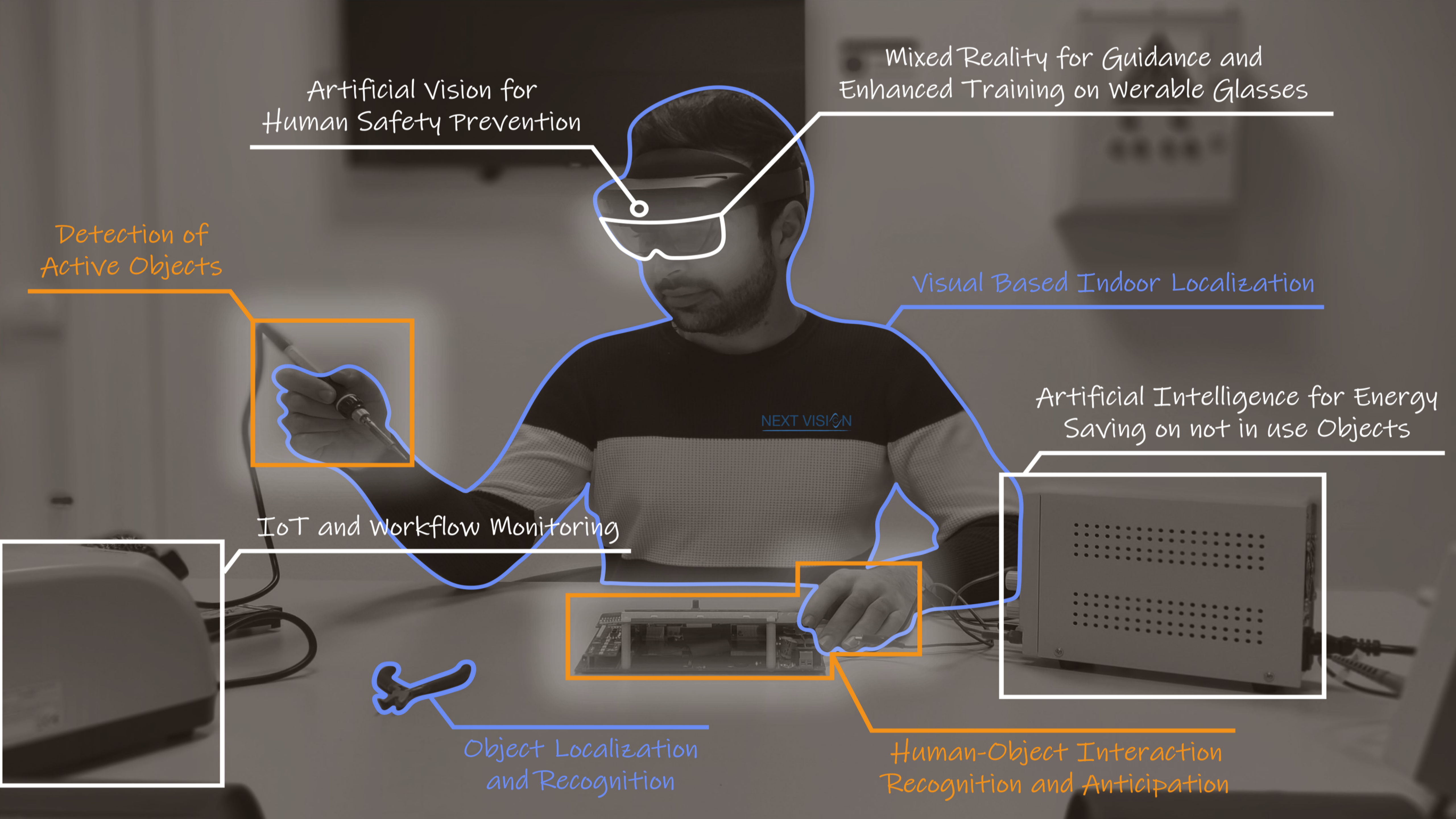
+ INTELLIGENCE

Smartphone Android

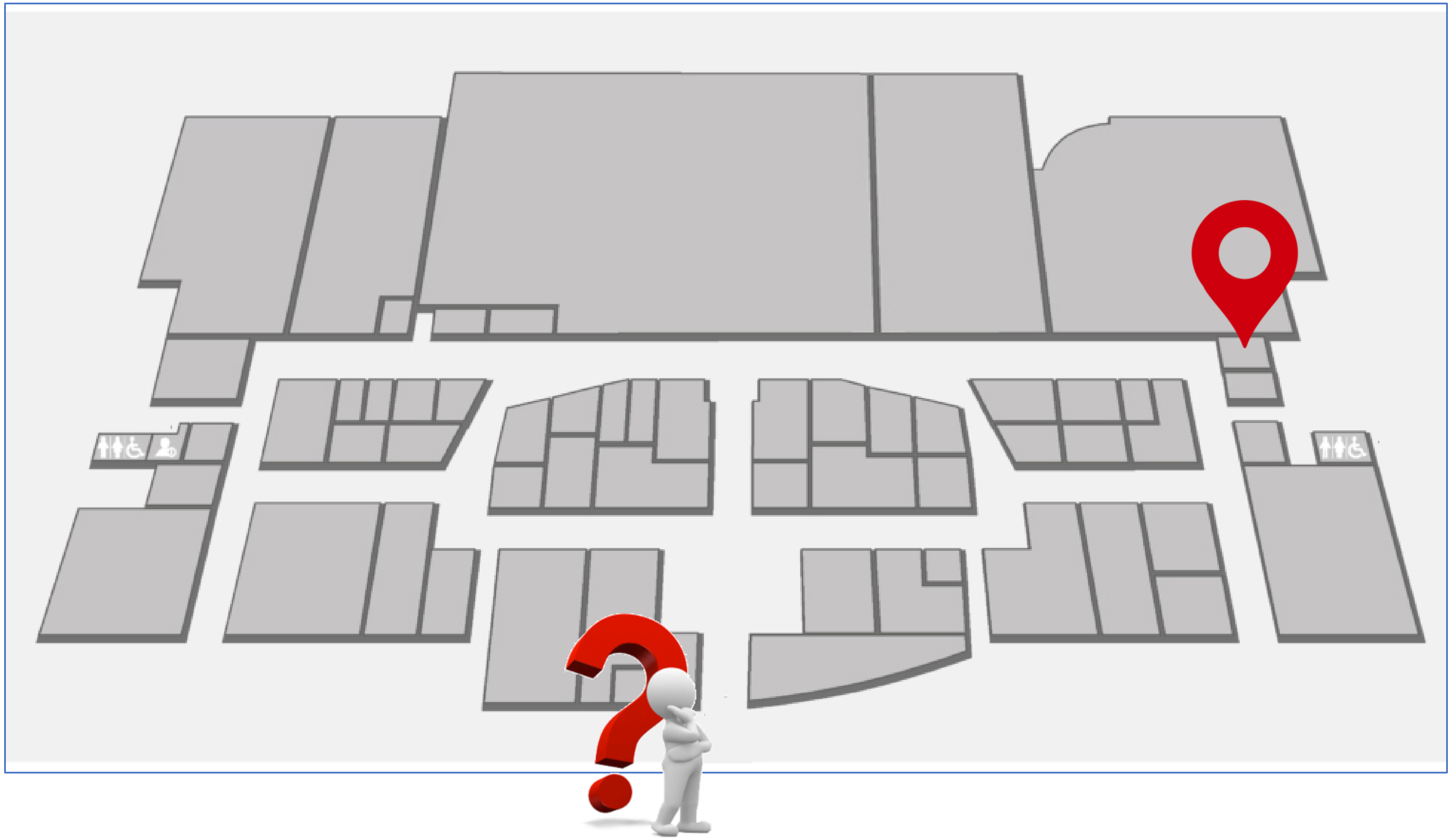
iOS

Tablet Android

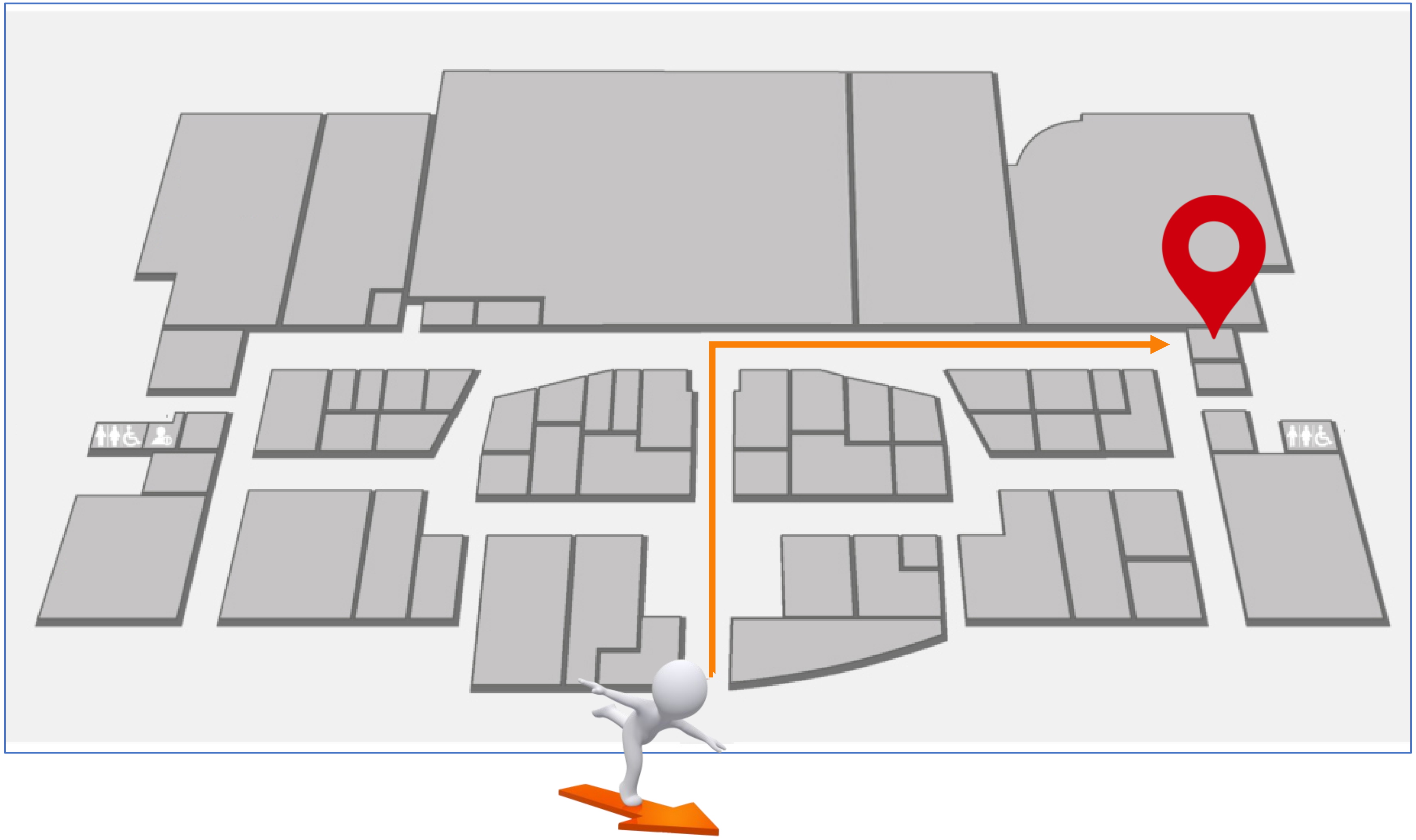
Ipad


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

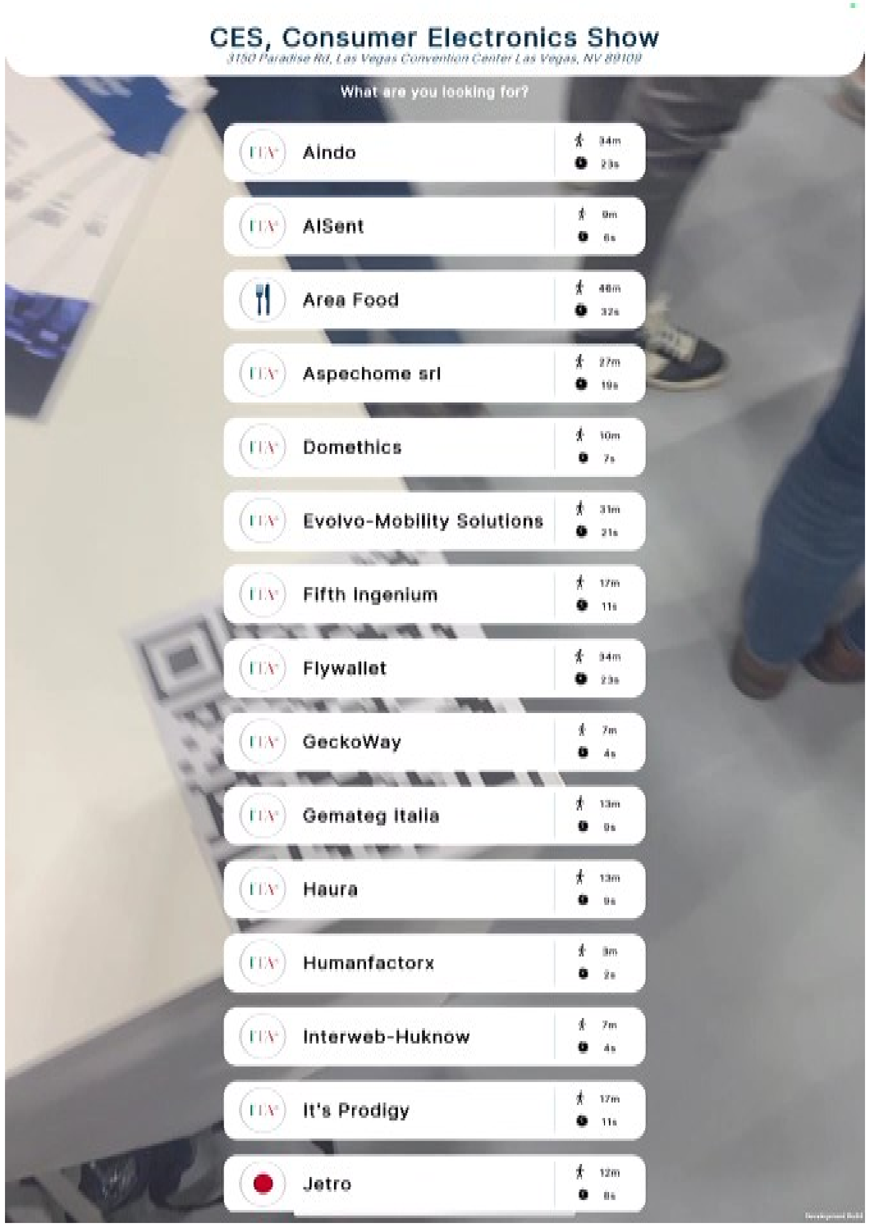
Navigation

Navigation



NAIROBI


CES, Consumer Electronics Show
3150 Paradise Rd, Las Vegas Convention Center Las Vegas, NV 89109
What are you looking for?
Aindo
AISent
Area Food
Aspechome srl
Domethics
Evolvo-Mobility Solutions
Fifth Ingenium
Flywallet
GeckoWay
Gemateg Italia
Haura
Humanfactorx
Interweb-Huknow
It's Prodigy
Jetro


Scheletro di Palaeoloxodon
Ingresso/Uscita
Mattonelle del refettorio settecentesco
Busto di Mario Rapisardi
Scheletro di Palaeoloxodon

NAOMI


NEXT VISION
Q 0025


electric screwdriver 98.0%
VALEX
Realtime recognition of
human-object interactions
NEXT VISION
Spin-off of the University of Catania


Trimble

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.

Thank you!

Antonino Furnari

Francesco Ragusa

Building Wearable Assistants with First Person (Egocentric Vision): History, Challenges, Opportunities 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

antonino.furnarni@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/