Hand-Object Interactions in Egocentric Vision

Rosario Leonardi

LIVE Group @ UNICT - https://iplab.dmi.unict.it/live/

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

Department of Mathematics and Computer Science - University of Catania

rosario.leonardi@unict.it

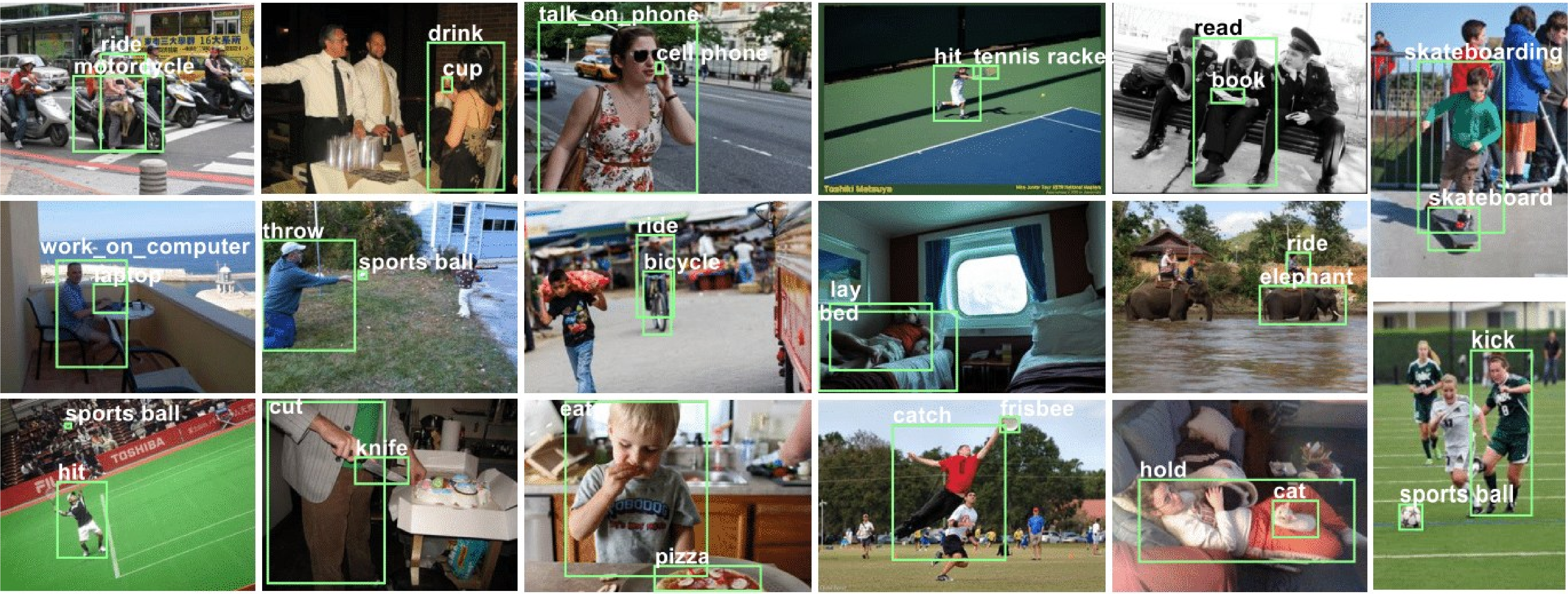
Agenda

1. **Introduction to Hand-Object Interactions Detection**
   - Human-Object Interaction vs Hand-Object Interaction
   - Definition and importance of Hand-Object Interactions
   - Applications in AR/VR, robotics, industrial monitoring, and assistive systems
2. **Datasets and Benchmarks for Hand-Object Interactions in Egocentric Vision**
   - Overview of popular datasets
   - Challenges in dataset collection and annotation
   - Synthetic datasets
3. **Models and Architectures for Hand-Object Interactions Detection**
   - Evaluation Protocols
   - Understanding Human Hands in Contact at Internet Scale
   - Exploiting multimodal synthetic data for egocentric human-object interaction detection in an industrial scenario
   - Are synthetic data useful for egocentric hand-object interaction detection?
4. **Open Challenges**
   - Standardization of evaluation protocols and task definitions
   - Updating and improving pre-existing architectures
   - Handling occlusions and complex interactions
   - Integration of multimodal data

1. Introduction to Hand-Object Interactions Detection

Human-Object Interactions

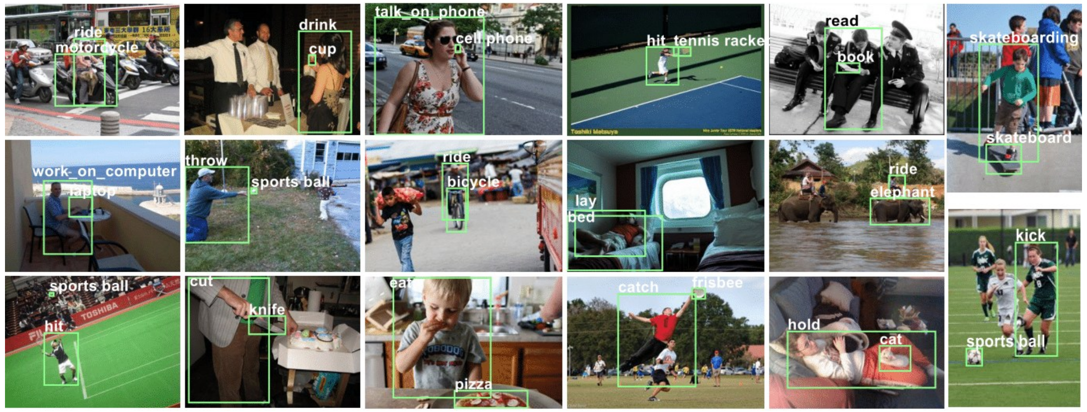
Human-Object Interactions



The set of actions or relationships that occur when a human interacts with an object, typically described as a ⟨**human, action, object**⟩ triplet.

- *Gupta Saurabh and Jitendra Malik. "Visual semantic role labeling" arXiv preprint arXiv:1505.04474 (2015)*

Human-Object Interactions



The set of actions or relationships that occur when a human interacts with an object, typically described as a ⟨**human, action, object**⟩ triplet.

- *Gupta Saurabh and Jitendra Malik. "Visual semantic role labeling" arXiv preprint arXiv:1505.04474 (2015)*
- *Chao Yu-Wei, et al. "Hico: A benchmark for recognizing human-object interactions in images." Proceedings of the IEEE international conference on computer vision. 2015.*
- *Chao, Yu-Wei, et al. "Learning to detect human-object interactions." 2018 ieee winter conference on applications of computer vision (wacv). IEEE, 2018.*
- *Gkioxari, Georgia, et al. "Detecting and recognizing human-object interactions." Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.*
- *Gao, Chen, Yuliang Zou, and Jia-Bin Huang. "ican: Instance-centric attention network for human-object interaction detection." arXiv preprint arXiv:1808.10437 (2018).*
- *Kim, Bumsoo, et al. "Hotr: End-to-end human-object interaction detection with transformers." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021.*

Humans Manipulate The World With Their Hands

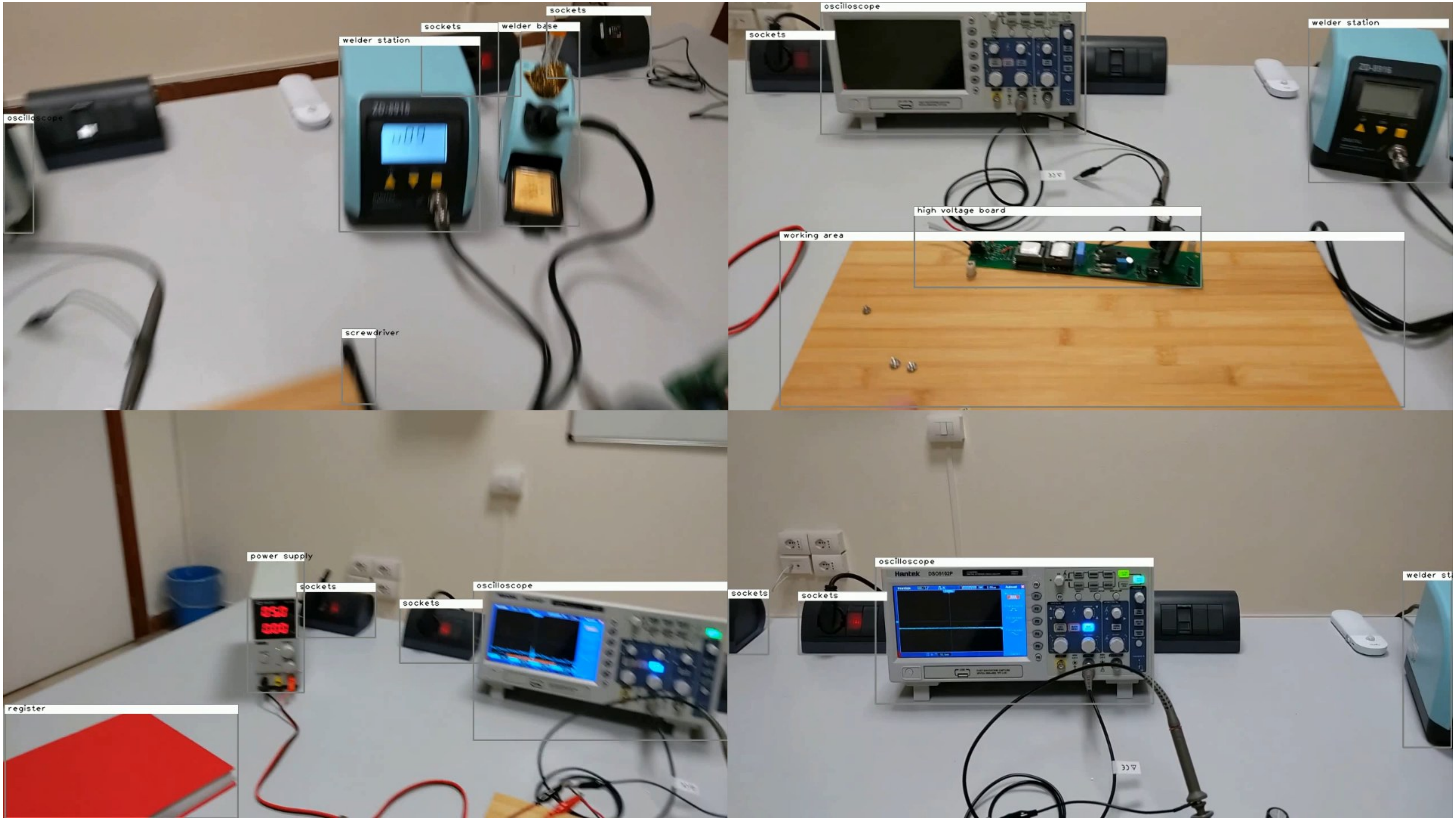
Shan Dandan, et al. "Understanding human hands in contact at internet scale" (CVPR 2020)

Hand-Object Interactions

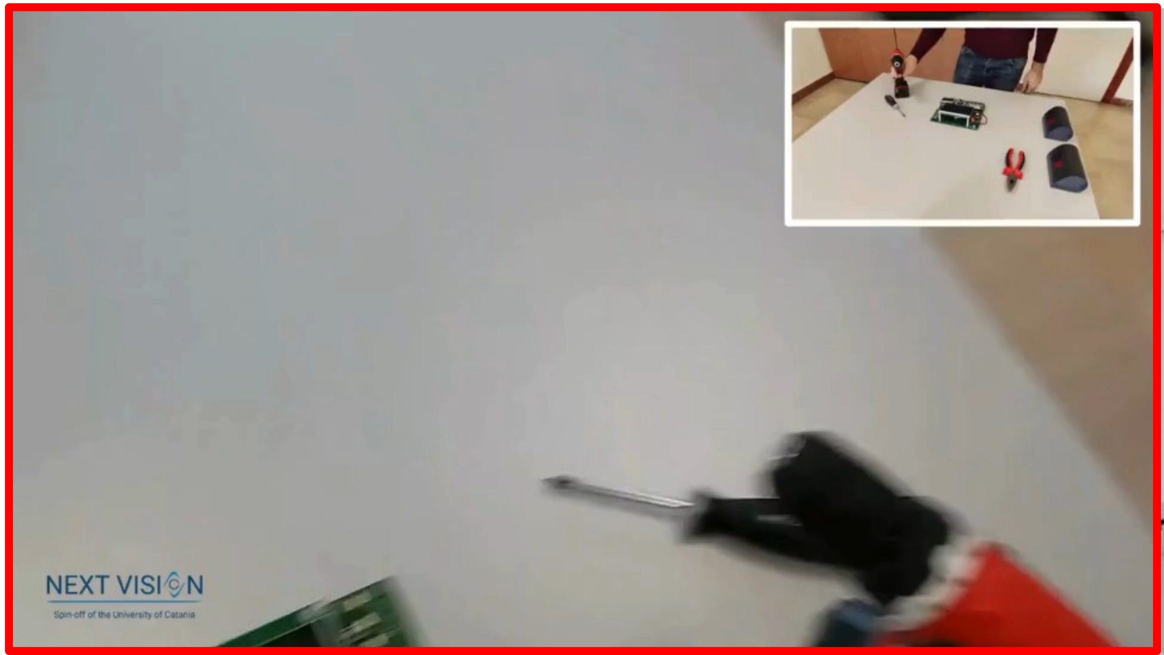
Hand-Object Interaction (HOI): ⟨**hand, object, contact state**⟩

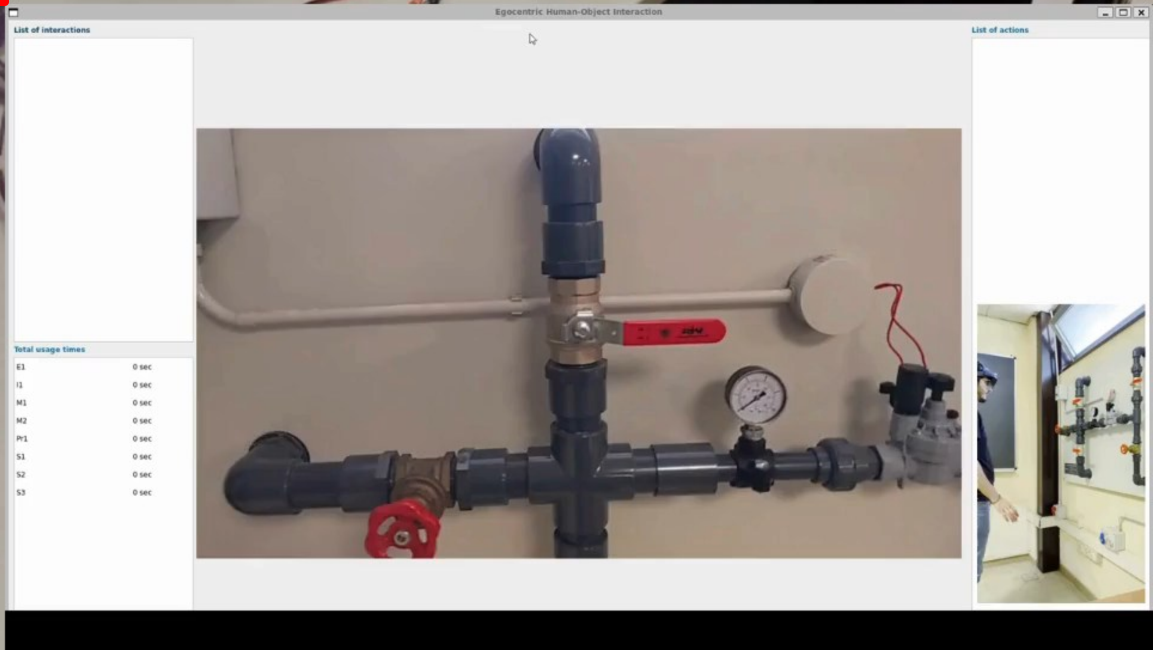
Shan Dandan, et al. "Understanding human hands in contact at internet scale" (CVPR 2020)

Applications



Applications





2. Datasets and Benchmarks for Hand-Object Interactions in Egocentric Vision

Hand-Object Interactions Datasets

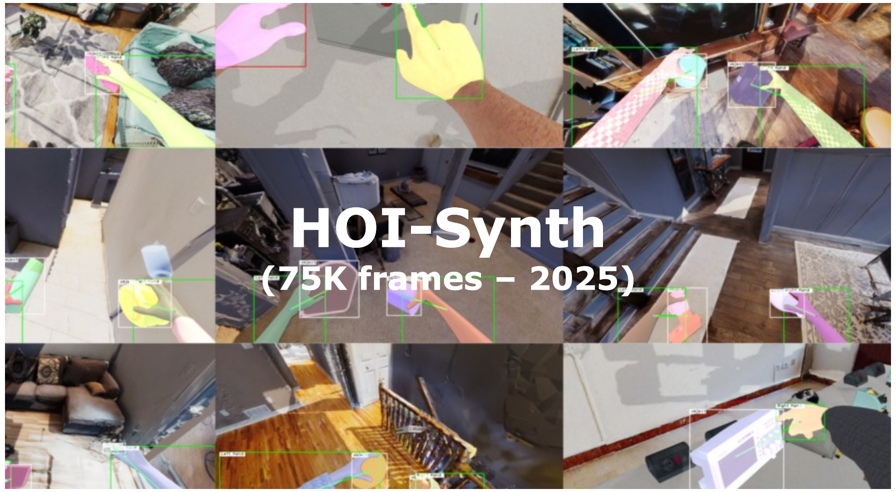
https://fouheylab.eecs.umich.edu/~dandans/

https://epic-kitchens.github.io/VISOR/

https://github.com/owenzlz/EgoHOS

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

https://iplab.dmi.unict.it/ENIGMA-51/



https://fpv-iplab.github.io/HOI-Synth/

100 Days of Hands

Shan Dandan, et al. "Understanding human hands in contact at internet scale" (CVPR 2020)

100 Days of Hands

Category	Amount	Storage	Brief Description
Boardgame	2,654	179G	playing board games, e.g. checkers, chess, monopoly, etc.
DIY	2,902	198G	diy clothes, gifts, food, cards or experiments etc.
Drinks	2,739	155G	making drinks, e.g. coffee, tea, smoothie, hydromel, etc.
Food	2,737	203G	making food, e.g. pasta, yogurt, seafood, chocolate, etc.
Furniture	2,813	145G	assembling furniture, e.g. bookcases, bedstands, tables, etc.
Gardening	955	79G	making gardens, e.g. growing trees, flowers, vegetable etc.
Housework	2,809	323G	doing housework or cleaning, decorating rooms, etc.
Packing	2,809	234G	packing or unpacking boxes, clothes, bags, gifts, etc.
Puzzle	2,825	176G	solving puzzles or magic cubes, building legos, etc.
Repair	2,764	177G	repairing cars, trucks, engines, computers, smartphones, etc.
Study	1,299	106G	studying and explaining videos for finals, midterms, engineering, etc.
Vlog	---	---	VLOG Dataset: videos documenting lives
Total	**27306**	**2.0T**	

131 Days

12 Categories

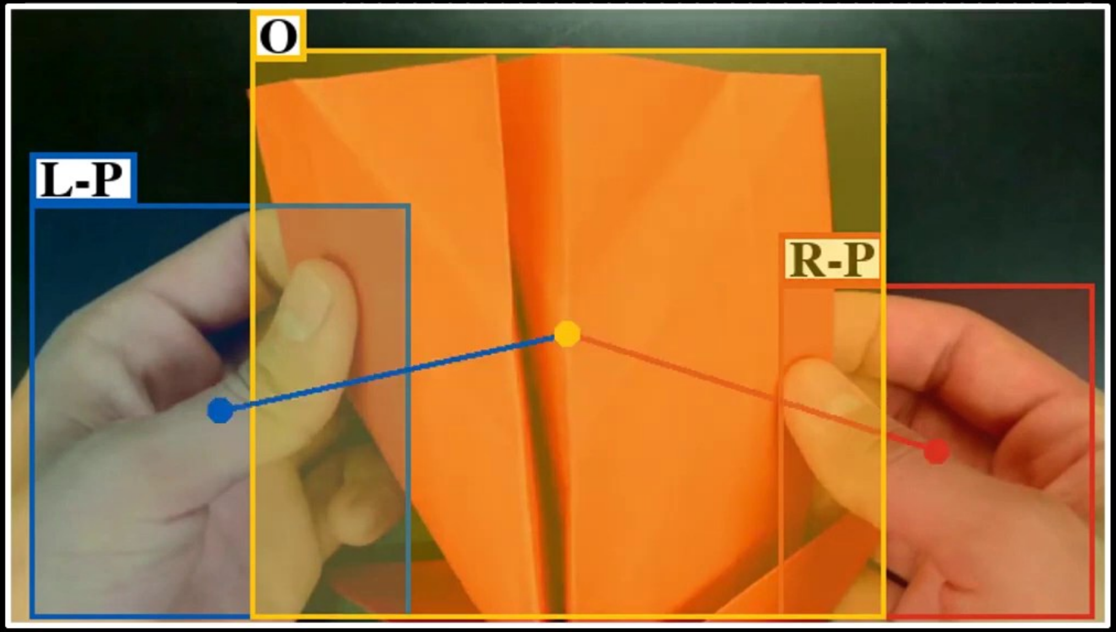
27.3K Videos

19.2K Uploaders

Shan Dandan, et al. "Understanding human hands in contact at internet scale" (CVPR 2020)

100 Days of Hands: 100K Frames

100K Frame-level Annotations

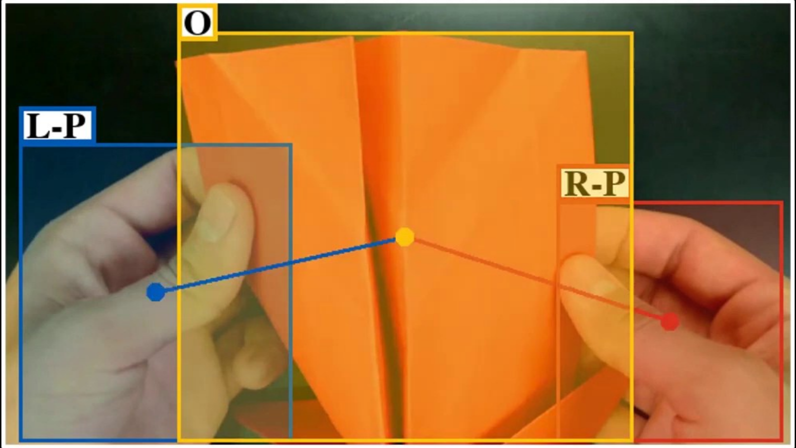


- Box around hand
- Side (left / right)
- Contact (no / self / other / portable / furniture)
- Box around object in contact
- Association

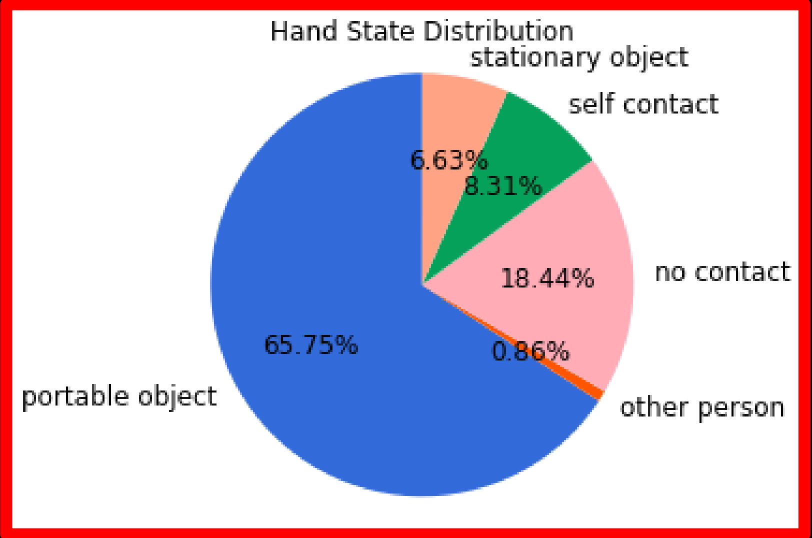
Shan Dandan, et al. "Understanding human hands in contact at internet scale" (CVPR 2020)

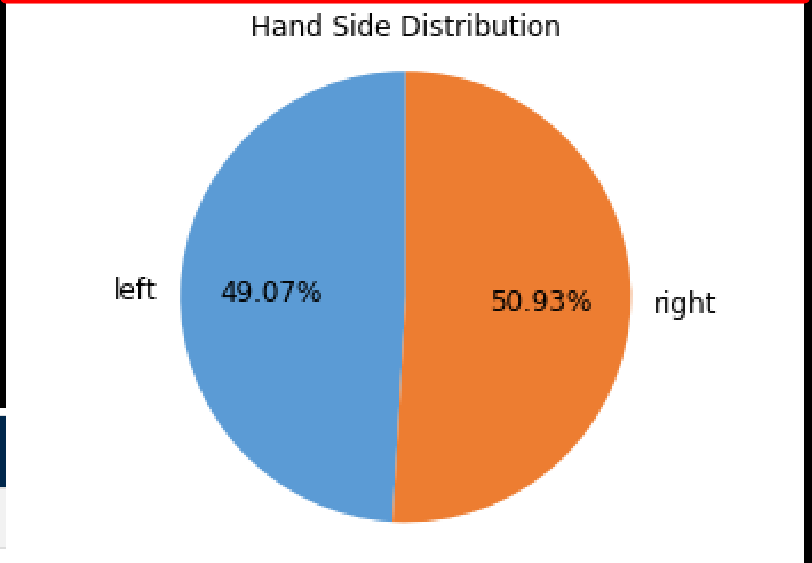
100 Days of Hands: 100K Frames

100K Frame-level Annotations



- Box around hand
- Side (left / right)
- Contact (no / self / other / portable / furniture)
- Box around object in contact
- Association





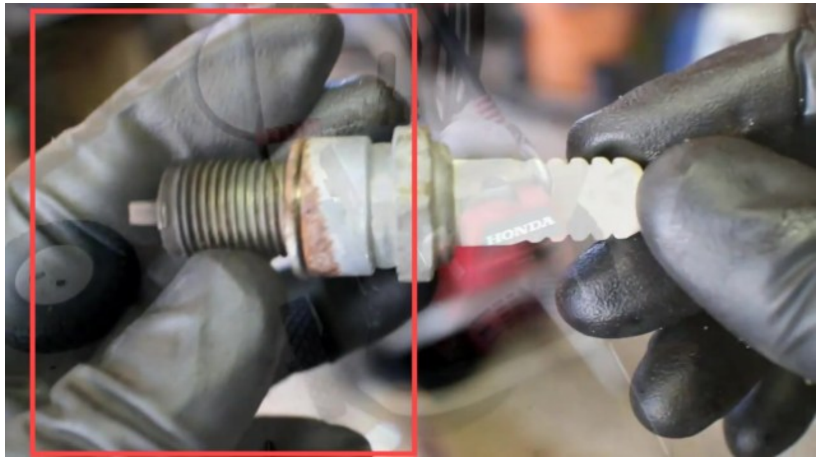
	#Frames	#Hand Box	#Hand Side	#Contact State	#Object Box
Total	99,899	189,426	189,426	189,426	140,431

Shan Dandan, et al. "Understanding human hands in contact at internet scale" (CVPR 2020)

100 Days of Hands: Hand Size

Small

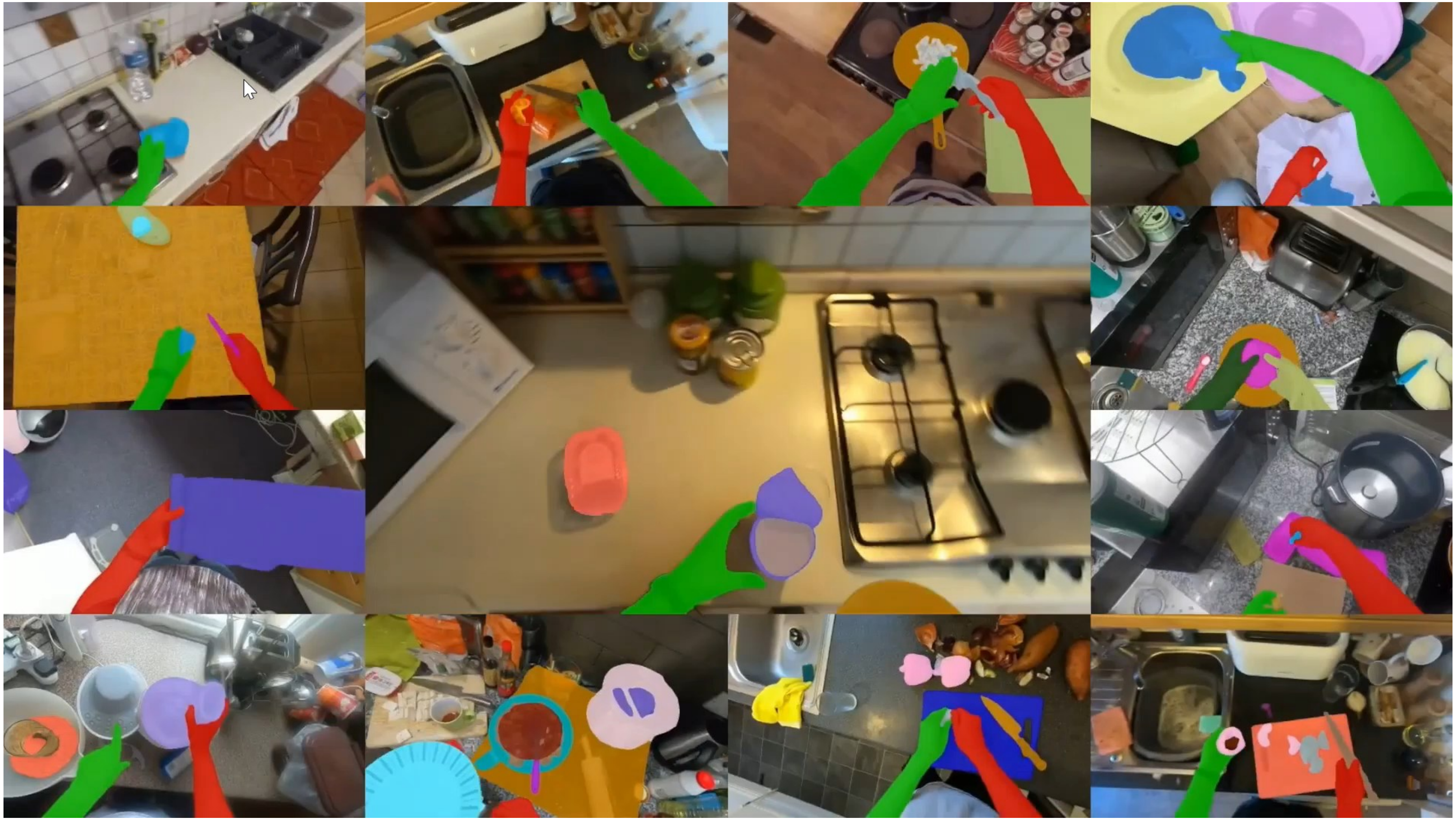
Medium



Large

Shan Dandan, et al. "Understanding human hands in contact at internet scale" (CVPR 2020)

EPIC-Kitchens VISOR



Darkhalil Ahmad, et al. "Epic-kitchens visor benchmark: Video segmentations and object relations" (Advances in Neural Information Processing Systems 2022)

EPIC-Kitchens VISOR: Annotations

Sparse Annotations

271K masks covering 36 hours of untrimmed video

Dense Annotations

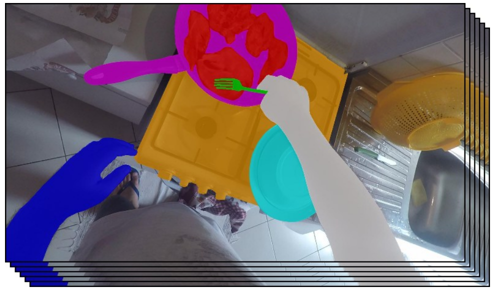
14.9M high quality automatic interpolations

Darkhalil Ahmad, et al. "Epic-kitchens visor benchmark: Video segmentations and object relations" (Advances in Neural Information Processing Systems 2022)

EPIC-Kitchens VISOR: Annotations

Sparse Annotations

Dense Annotations

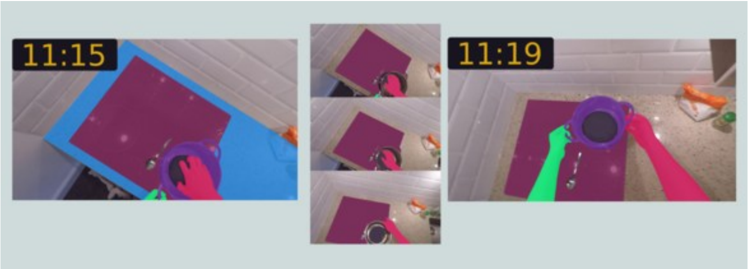


Darkhalil Ahmad, et al. "Epic-kitchens visor benchmark: Video segmentations and object relations" (Advances in Neural Information Processing Systems 2022)

EPIC-Kitchens VISOR: Challenges

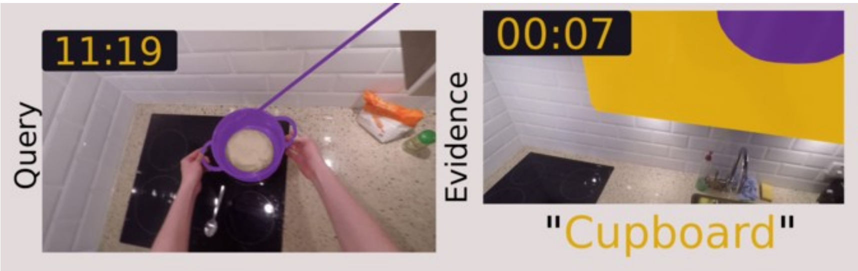
Video Object Segmentation

Goal: Track segments through video and occlusion



Where Did This Come From?

Goal: Name and point to where things came from



Hand Object Segmentation

Goal: Identify contact with 67K in-hand object masks

Darkhalil Ahmad, et al. "Epic-kitchens visor benchmark: Video segmentations and object relations" (Advances in Neural Information Processing Systems 2022)

EPIC-Kitchens VISOR: Annotation Stats

Hand Object Segmentation

Goal: Identify contact with 67K in-hand object masks

	Train+Val	Test
In Contact	52,685	14,233
Not-in-contact	4,144	1,341
Inconclusive	4,943	1,104

Darkhalil Ahmad, et al. "Epic-kitchens visor benchmark: Video segmentations and object relations" (Advances in Neural Information Processing Systems 2022)

EPIC-Kitchens VISOR: Annotation Stats

Hand Object Segmentation

Goal: Identify contact with 67K in-hand object masks

	Train+Val	Test
In Contact	52,685	14,233
Not-in-contact	4,144	1,341
Inconclusive	4,943	1,104

Darkhalil Ahmad, et al. "Epic-kitchens visor benchmark: Video segmentations and object relations" (Advances in Neural Information Processing Systems 2022)

EgoHOS

Zhang Lingzhi, et al. "Fine-grained egocentric hand-object segmentation: Dataset, model, and applications" (ECCV 2022)

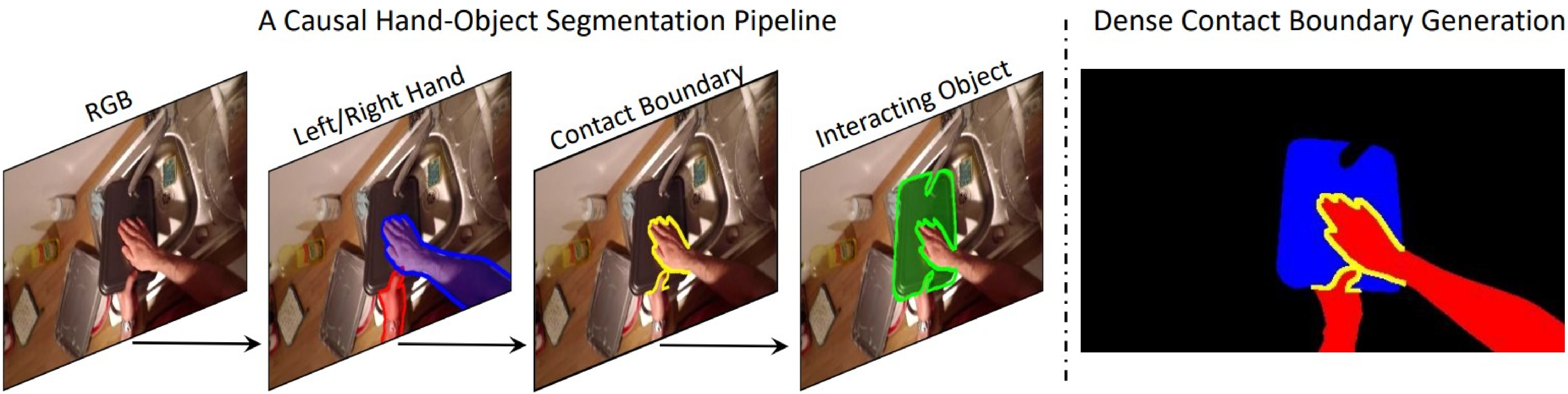
EgoHOS: Data Source

70%

Epic Kitchen | Ego4d | THU-Read | Escape Room

Zhang Lingzhi, et al. "Fine-grained egocentric hand-object segmentation: Dataset, model, and applications" (ECCV 2022)

EgoHOS: Annotations



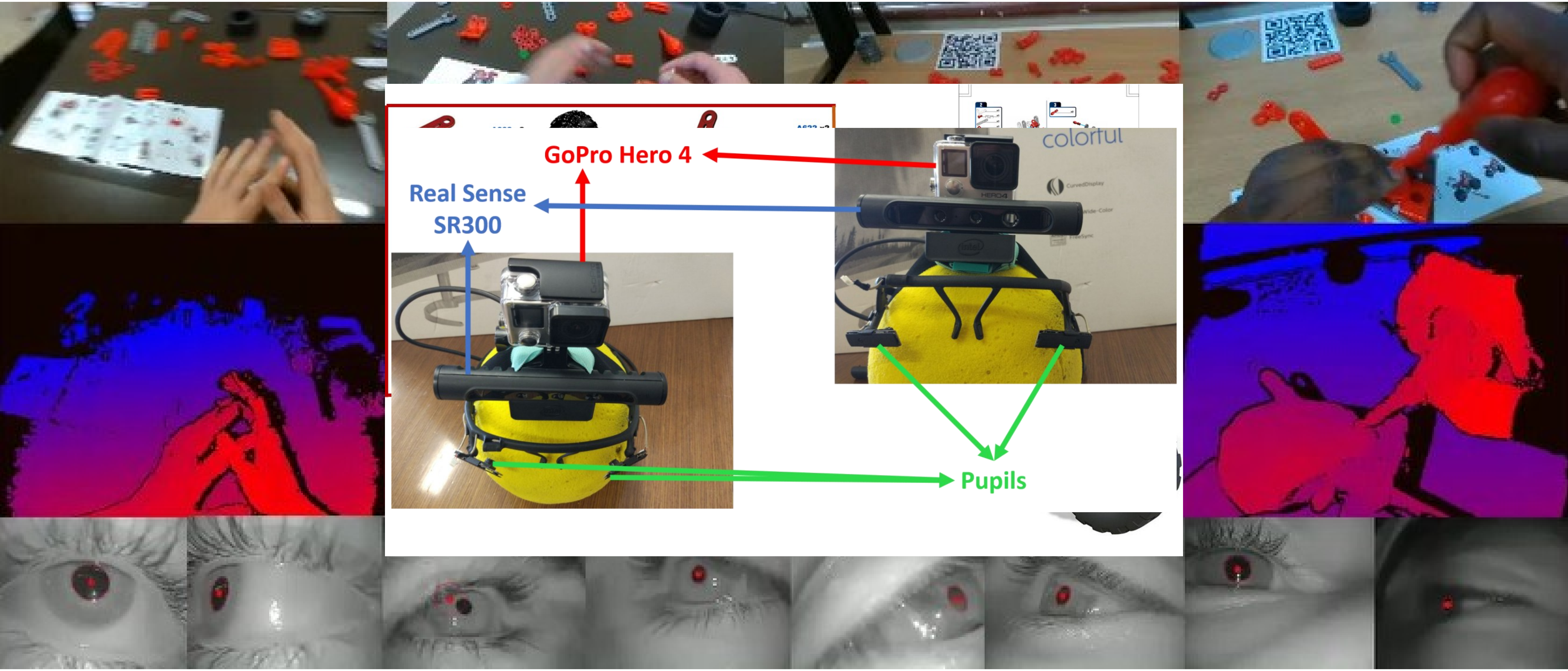
Zhang Lingzhi, et al. "Fine-grained egocentric hand-object segmentation: Dataset, model, and applications" (ECCV 2022)

EgoHOS: Hand and Interacting Object Segmentations

	Label	#Frames	#Hands	#Objects
EgoHOS	Mask	11,243	20,701	17,568

Zhang Lingzhi, et al. "Fine-grained egocentric hand-object segmentation: Dataset, model, and applications‶ (ECCV 2022)

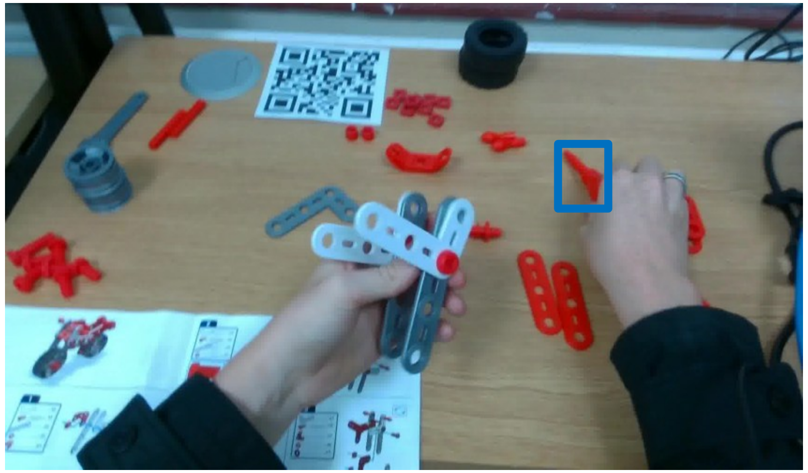
The MECCANO Dataset



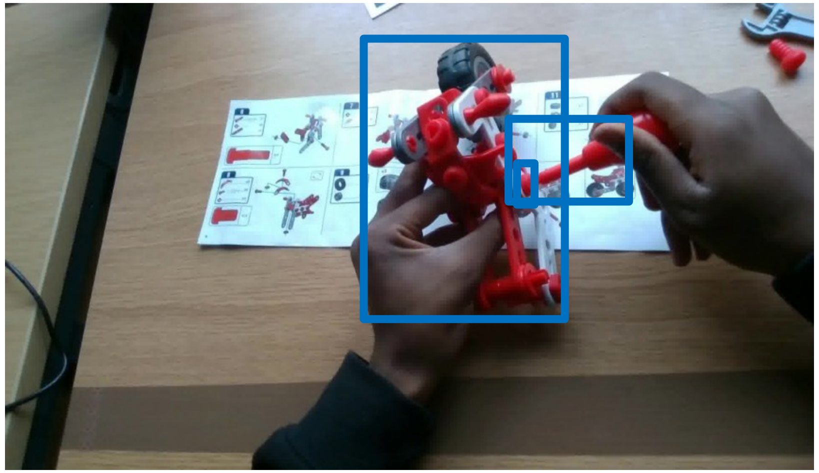
Ragusa Francesco, et al. "Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain" (CVIU 2023)

MECCANO Benchmark

1. Action Recognition
2. Active Object Detection and Recognition
3. EHOI Detection
4. Action Anticipation
5. Next-Active Object (NAO) Detection



<take, screwdriver>



<screw, {screwdriver, screw, partial_model}>

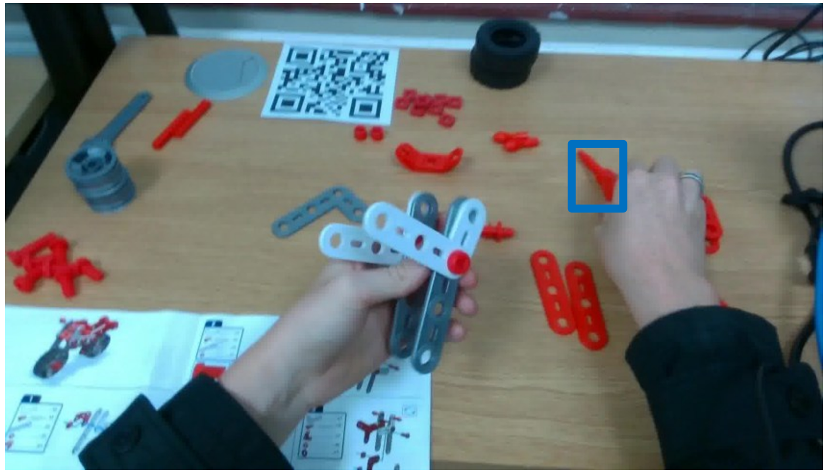
Ragusa Francesco, et al. "Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain" (CVIU 2023)

Egocentric Human-Object Interaction

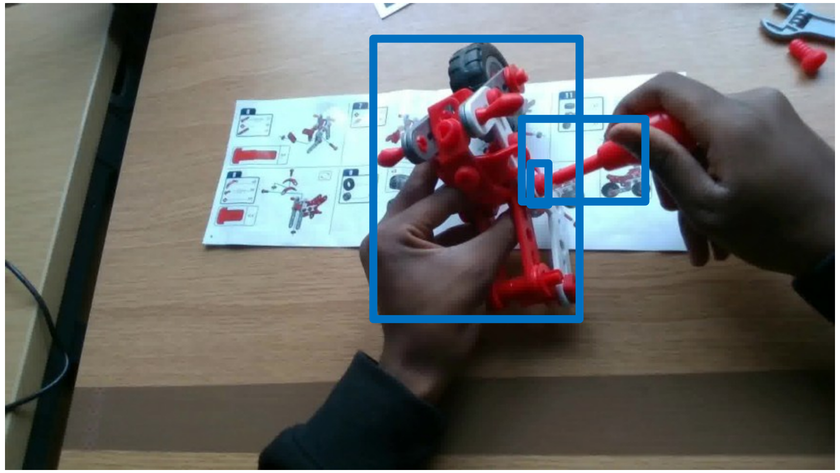
$$O = \{o_1, o_2, \ldots, o_n\}$$

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

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



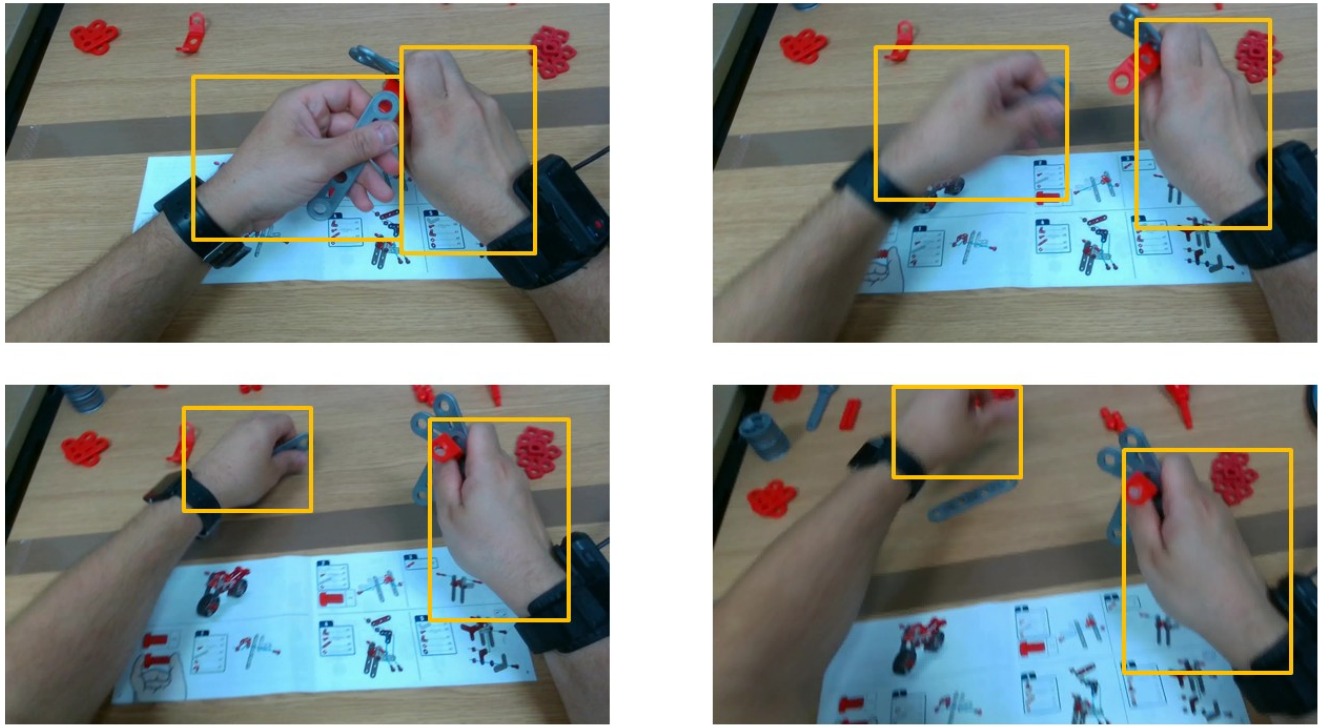
<take, screwdriver>

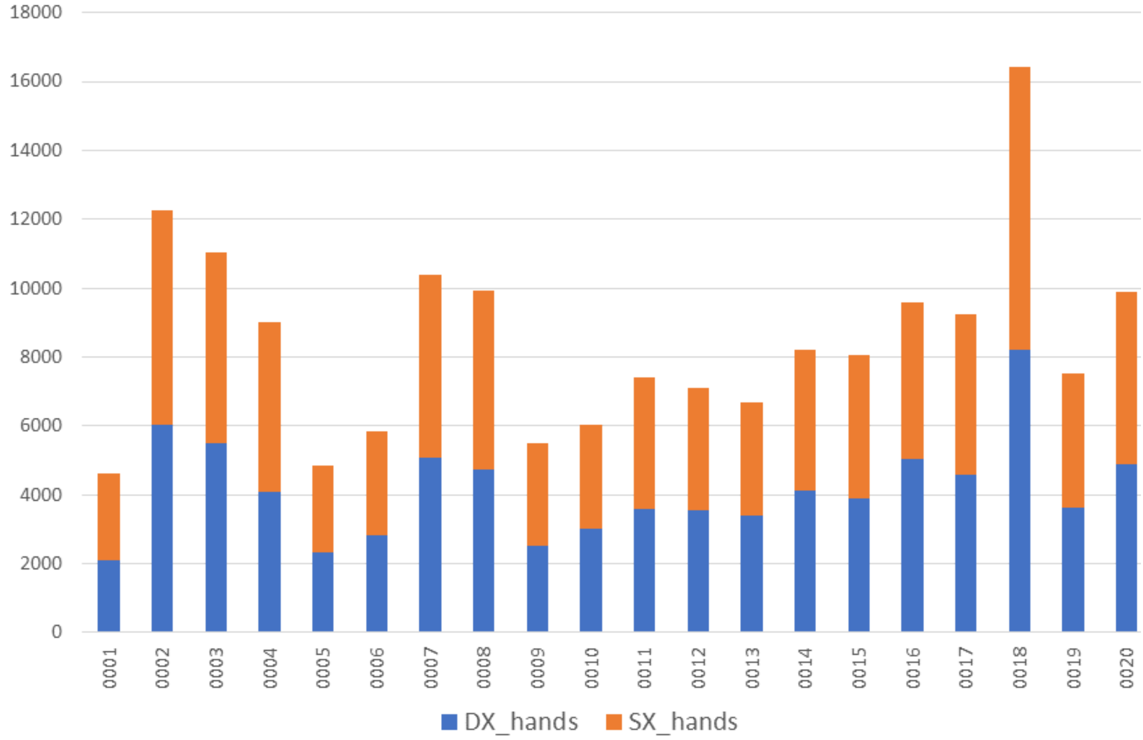


<screw, {screwdriver, screw, partial_model}>

Ragusa Francesco, et al. "Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain" (CVIU 2023)

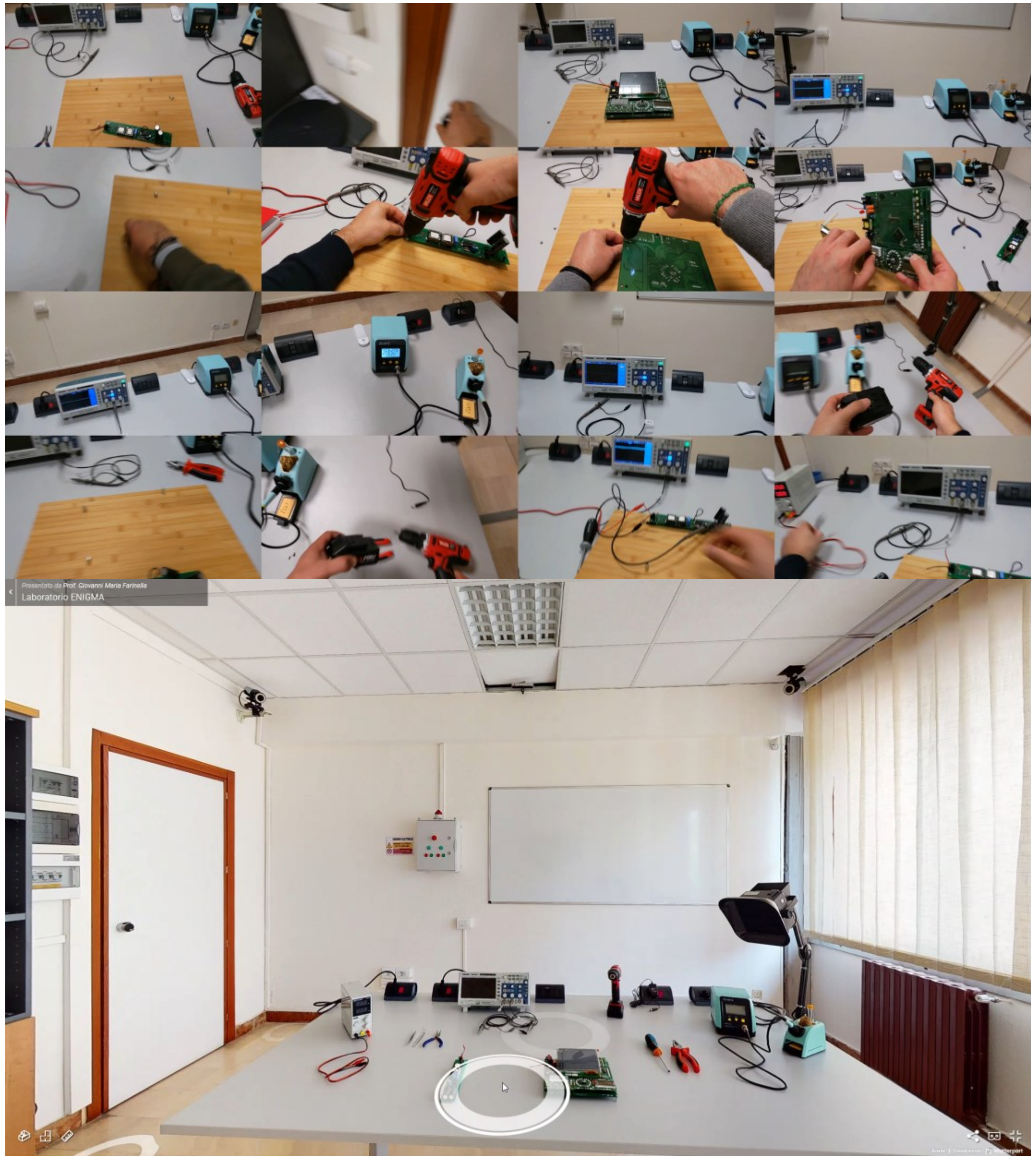
Data Annotation: Hands Annotations





Ragusa Francesco, et al. "Meccano: A multimodal egocentric dataset for humans behavior understanding in the industrial-like domain" (CVIU 2023)

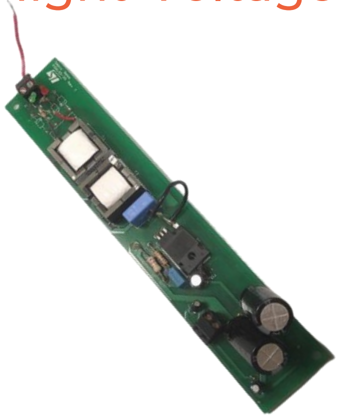
ENIGMA-51 Dataset



We designed two procedures consisting of instructions that involve humans interacting with the objects present in the laboratory to achieve the goal of repairing two electrical boards

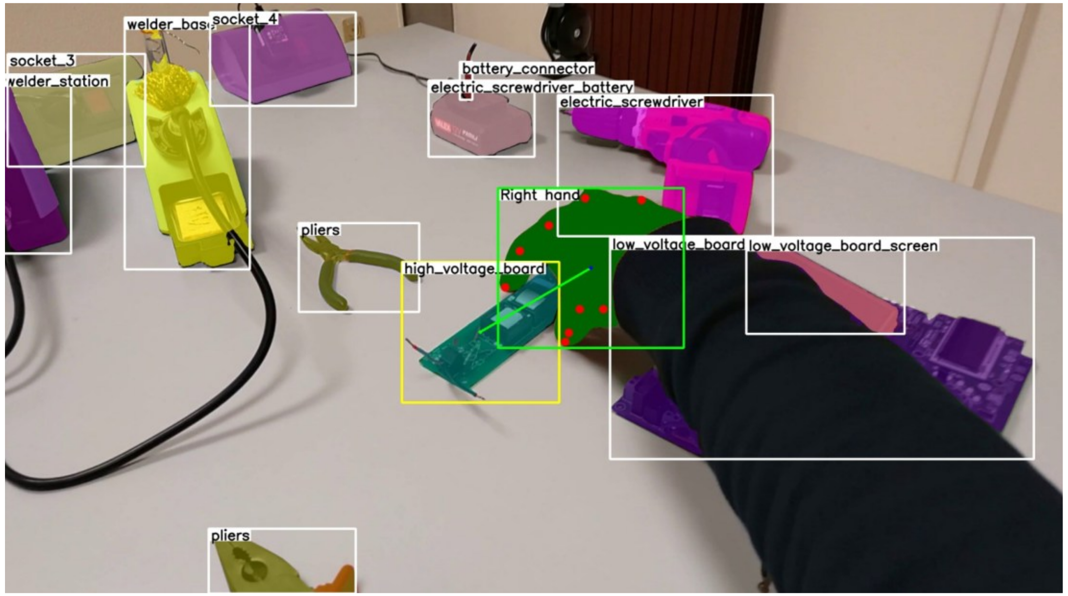
Low-Voltage

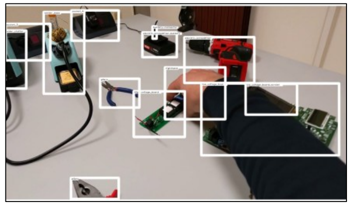
Hight-Voltage

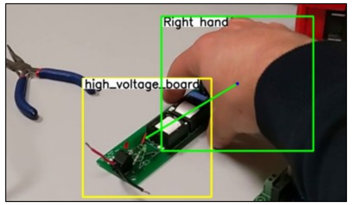


Ragusa Francesco, et al. "Enigma-51: Towards a fine-grained understanding of human behavior in industrial scenarios" (WACV 2024)

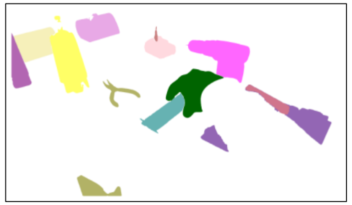
ENIGMA-51: Annotations

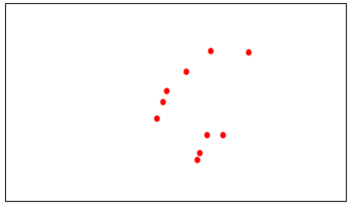



Hand-Object boxes



Human-Object Interactions


Hand-Object Masks


Hand Keypoints


Environment 3D Model


Object 3D Models

TTC: 1.20s $\Delta = 0.40$ TTC: 0.80s TTC: 0.40s first-contact high voltage board

Past Frames

Interaction Frame

Procedure :

……

4. Take the high voltage board and put it on the working area

5. Take the screwdriver

……

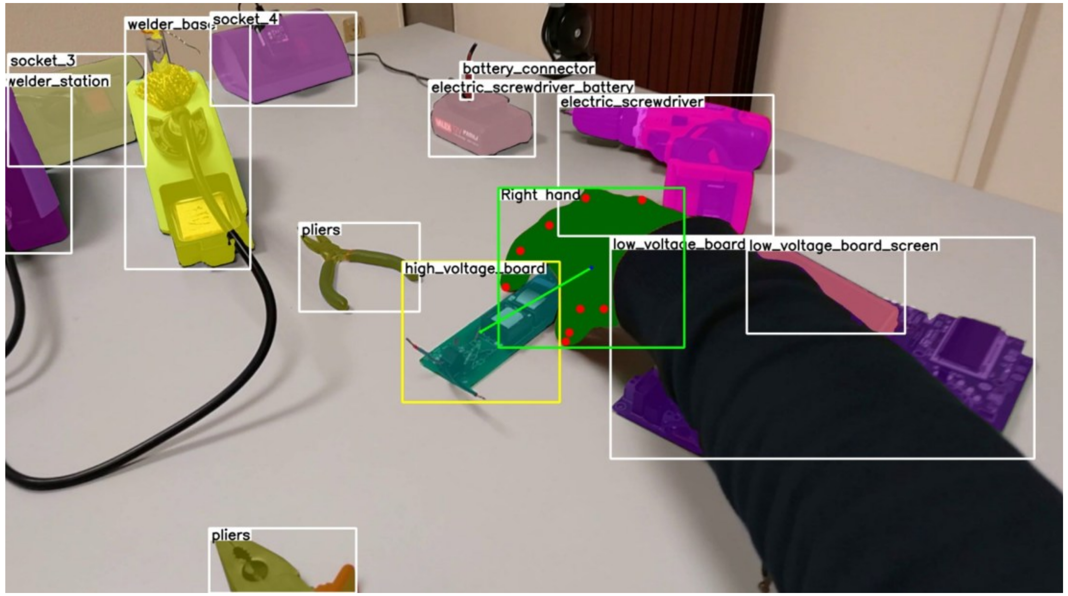
22. Turn on the welder using the switch on the corresponding socket (second from right)

23. Set the temperature of the welder to 480 °C using the yellow "UP" button

……

Ragusa Francesco, et al. "Enigma-51: Towards a fine-grained understanding of human behavior in industrial scenarios" (WACV 2024)

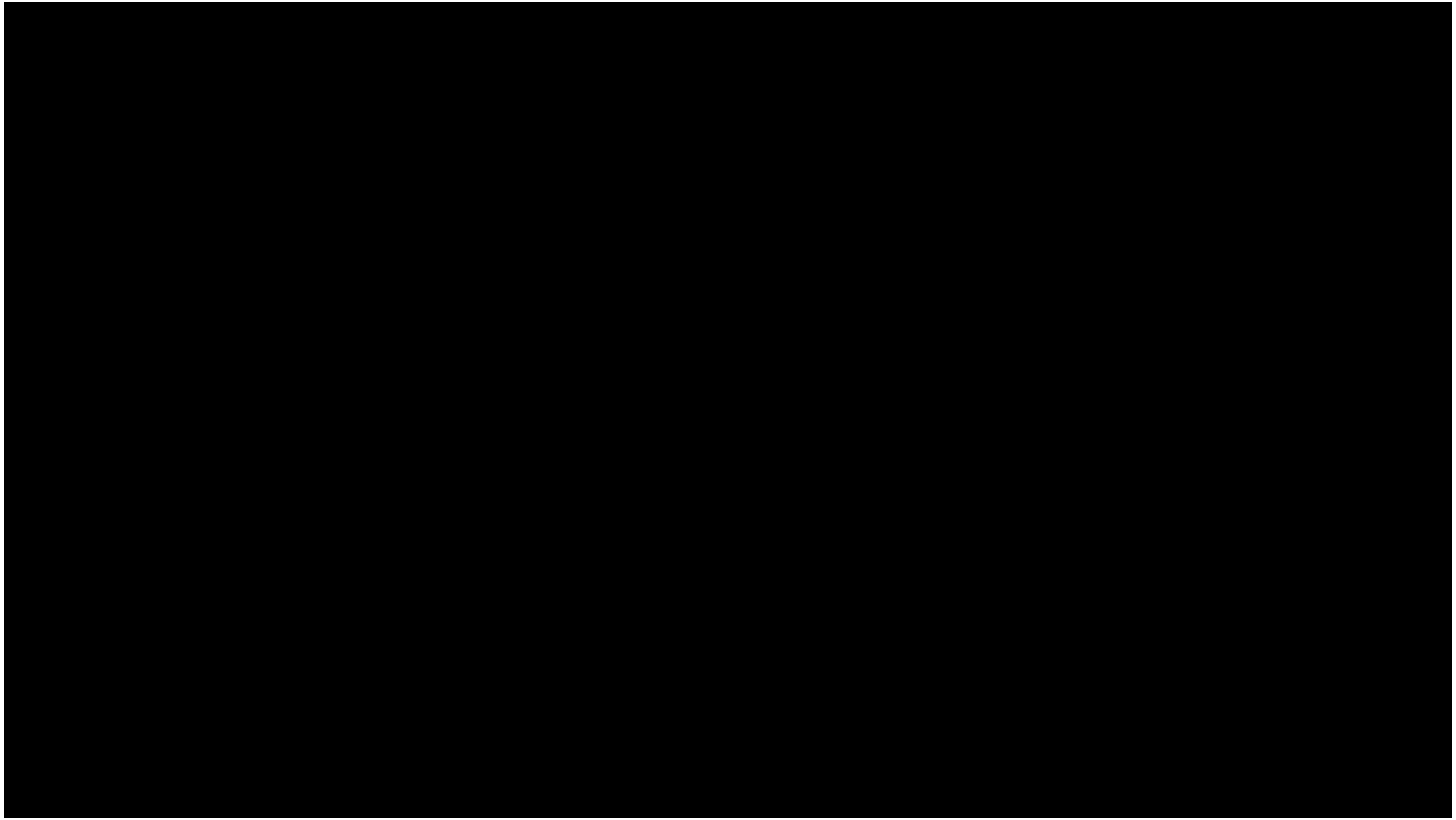
ENIGMA-51 Dataset



Splits	Train	Val	Test	Total
# Videos	27	8	16	51
# Videos Length	$\simeq$11h	$\simeq$4h	$\simeq$7h	$\simeq$22h
# Images	25,311	8,528	11,666	45,505
# Objects	152,865	53,486	68,784	275,135
# Active Objects	4,709	1,700	2,933	9,342
# Hands	31,249	11,322	13,902	56,473
# Hands in contact	5,039	1,833	3,171	10,043
# Interactions frames	6,386	2,150	4,061	12,597
# Interactions	7,133	2,406	4,497	14,036
# Past frames	19,090	6,437	7,683	33,210
# Next Object Interactions	21,535	7,280	8,499	37,314

Ragusa Francesco, et al. "Enigma-51: Towards a fine-grained understanding of human behavior in industrial scenarios" (WACV 2024)

HOT3D



Banerjee Prithviraj, et al. "Hot3d: Hand and object tracking in 3d from egocentric multi-view videos" (CVPR 2025)

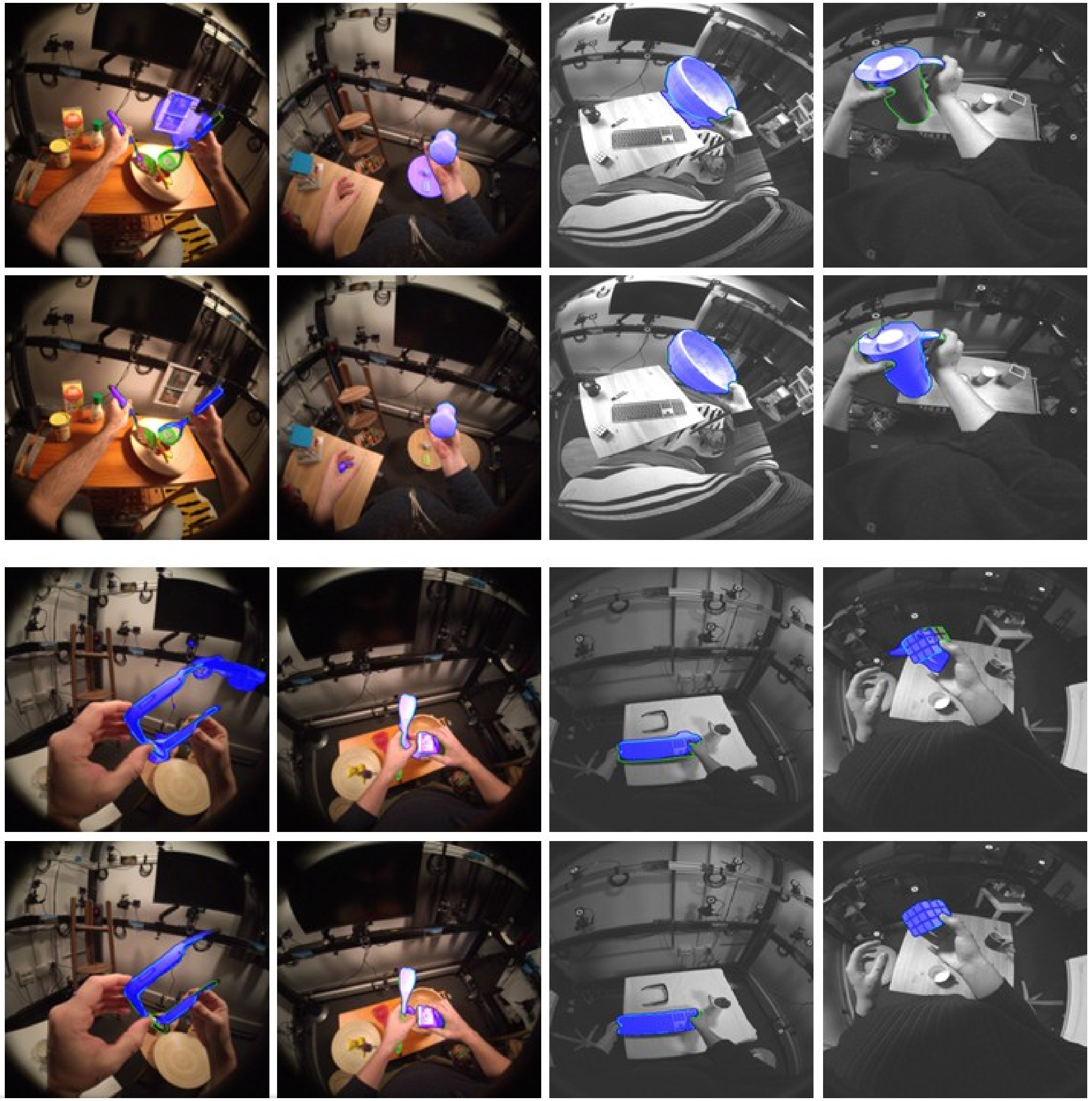
HOT3D

Dataset	Im								r scene	Gaze	Annotation
HOT3D (ours)	3.								~6	✓	Mocap
ARTIC [15]	2.								1	✗	Mocap
HOI4D [39]	2.								1–12	✗	RGB-D + man.
HO-Cap [69]	6								4	✗	RGB-D + optim.
DexYCB [8]	5								2–4	✗	Manual
HOGraspNet [9]	1.5M	RGB						30	1	✗	RGB-D + optim.
H2O-3D [24]	75K	RGB						10	1	✗	RGB-D + optim.
HO-3D [23]	78K	RGB	0 / 1	–	–	10	Single	10	1	✗	RGB-D + optim.
H2O [38]	572K	RGB-D	1 / 4	✓	Helmet	4	Both	8	1	✗	RGB-D + optim.
ContactPose [5]	2.9M	RGB-D	0 / 3	✗	–	50	Both	25	1	✗	RGB-D + mocap
ObMan [27]	147K	Synth. RGB	0 / 1	–	–	20	Single	2772	1	✗	Synthetic

Banerjee Prithviraj, et al. "Hot3d: Hand and object tracking in 3d from egocentric multi-view videos" (CVPR 2025)

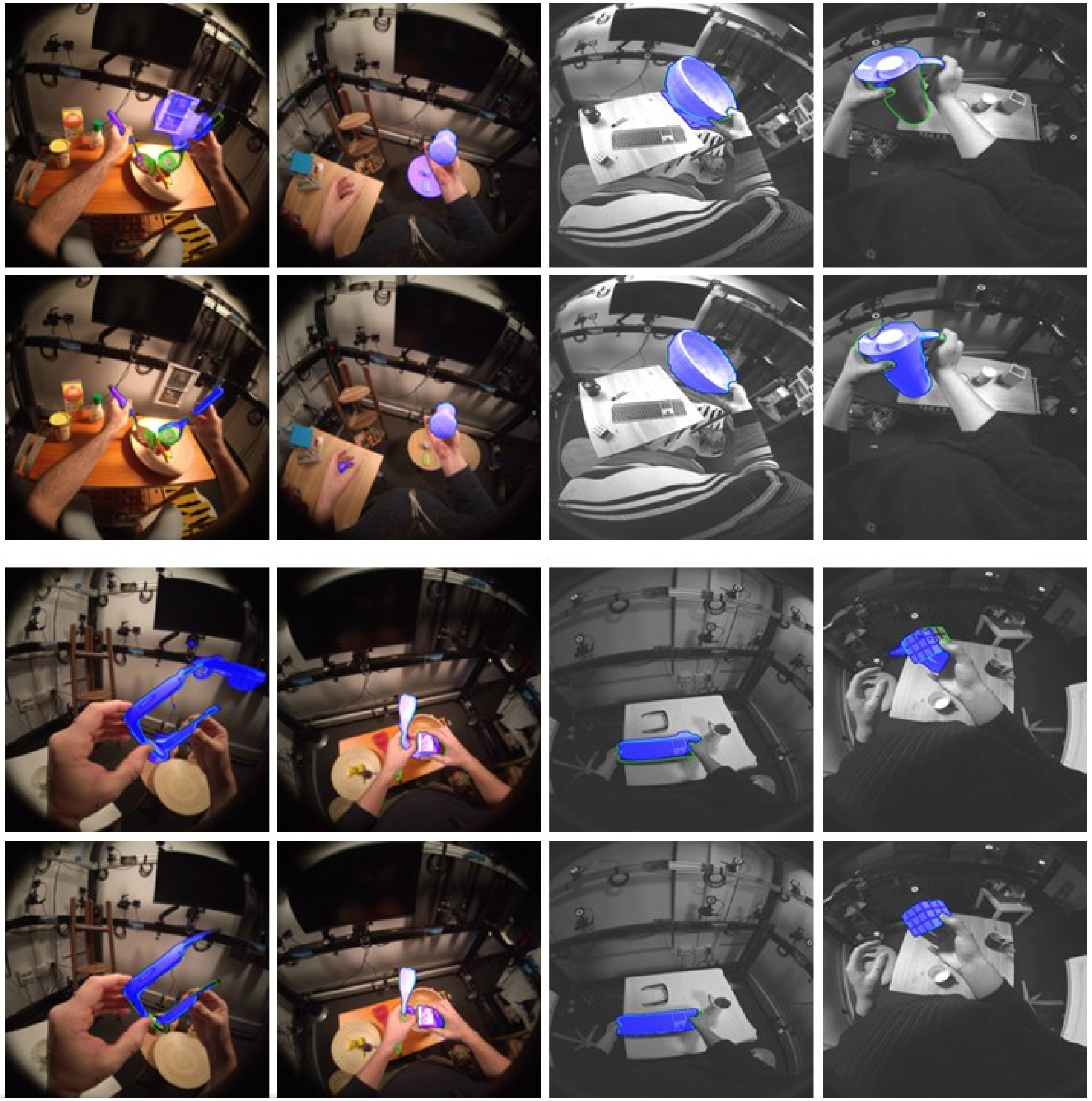
HOT3D: Tasks

1. 3D hand pose tracking
2. 6DoF object pose estimation
3. 2D segmentation of in-hand objects
4. 3D lifting of in-hand objects



Banerjee Prithviraj, et al. "Hot3d: Hand and object tracking in 3d from egocentric multi-view videos" (CVPR 2025)

HOT3D: Tasks

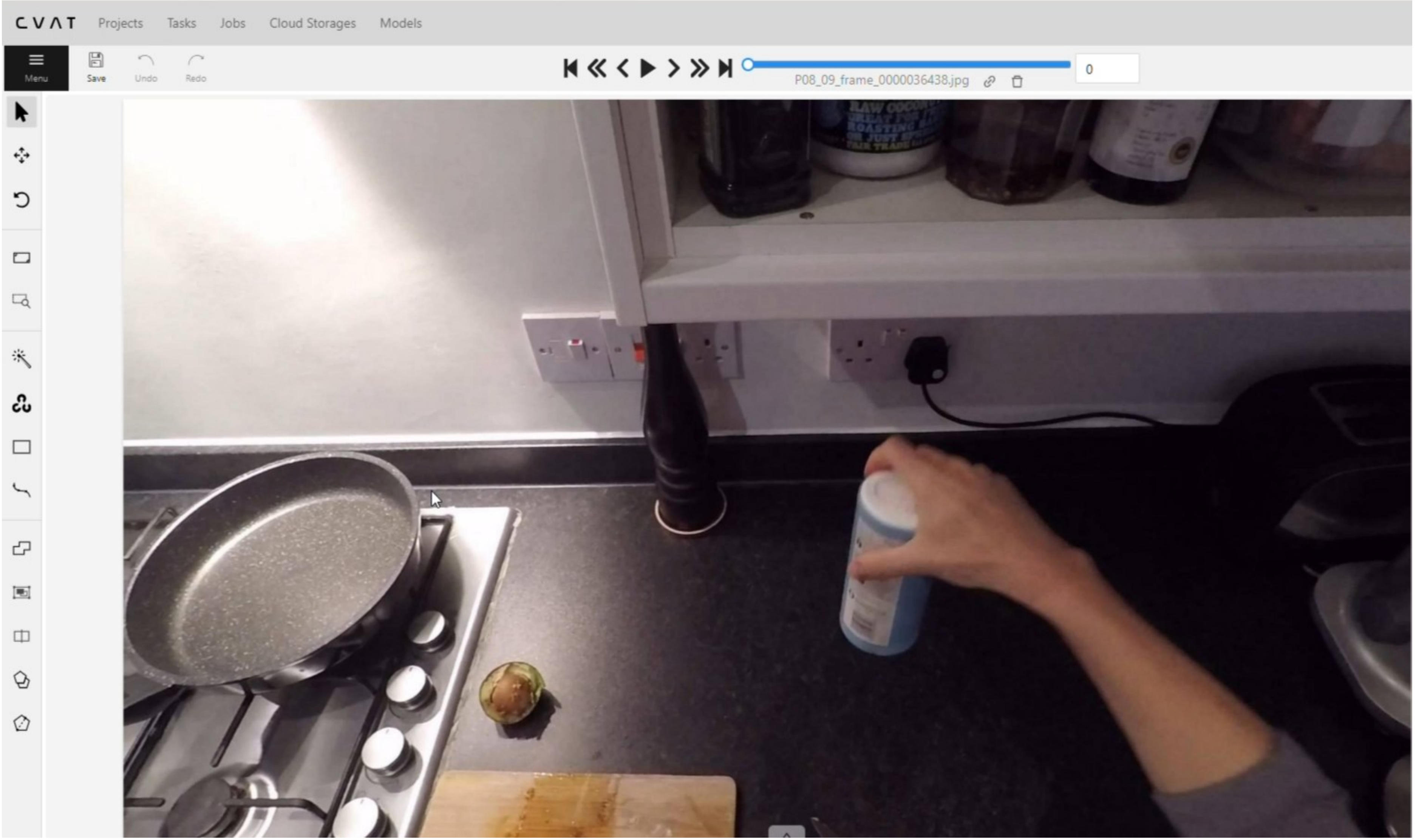


Method	Test dataset	Object in hand (mIoU ↑): Either	Left	Right	Both
EgoHOS [75]	EgoHOS	–	62.2	44.4	52.8
EgoHOS [75]	HOT3D-Aria	42.6	21.0	26.3	32.5
MRCNN	HOT3D-Aria	47.1	–	–	–
MRCNN-DA	HOT3D-Aria	**55.2**	–	–	–
EgoHOS [75]	HOT3D-Quest3	33.1	13.5	14.4	24.8
MRCNN	HOT3D-Quest3	37.8	–	–	–
MRCNN-DA	HOT3D-Quest3	**54.7**	–	–	–

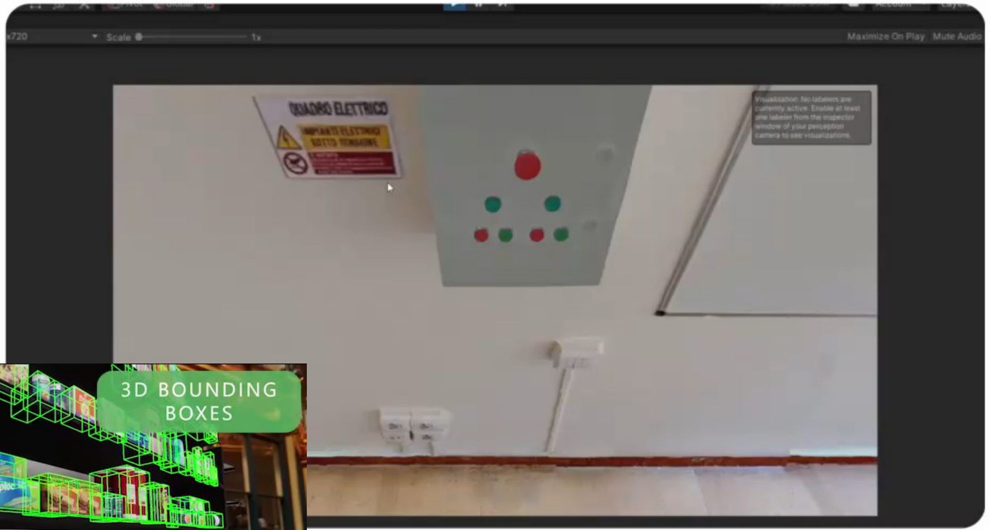
Table 4. **2D segmentation of in-hand objects.** EgoHOS [75] trained on the EgoHOS dataset is compared with our baselines based on Mask R-CNN [28] and trained on our in-house dataset of images from Aria. We observe a large accuracy drop of the EgoHOS model on HOT3D.

Banerjee Prithviraj, et al. "Hot3d: Hand and object tracking in 3d from egocentric multi-view videos" (CVPR 2025)

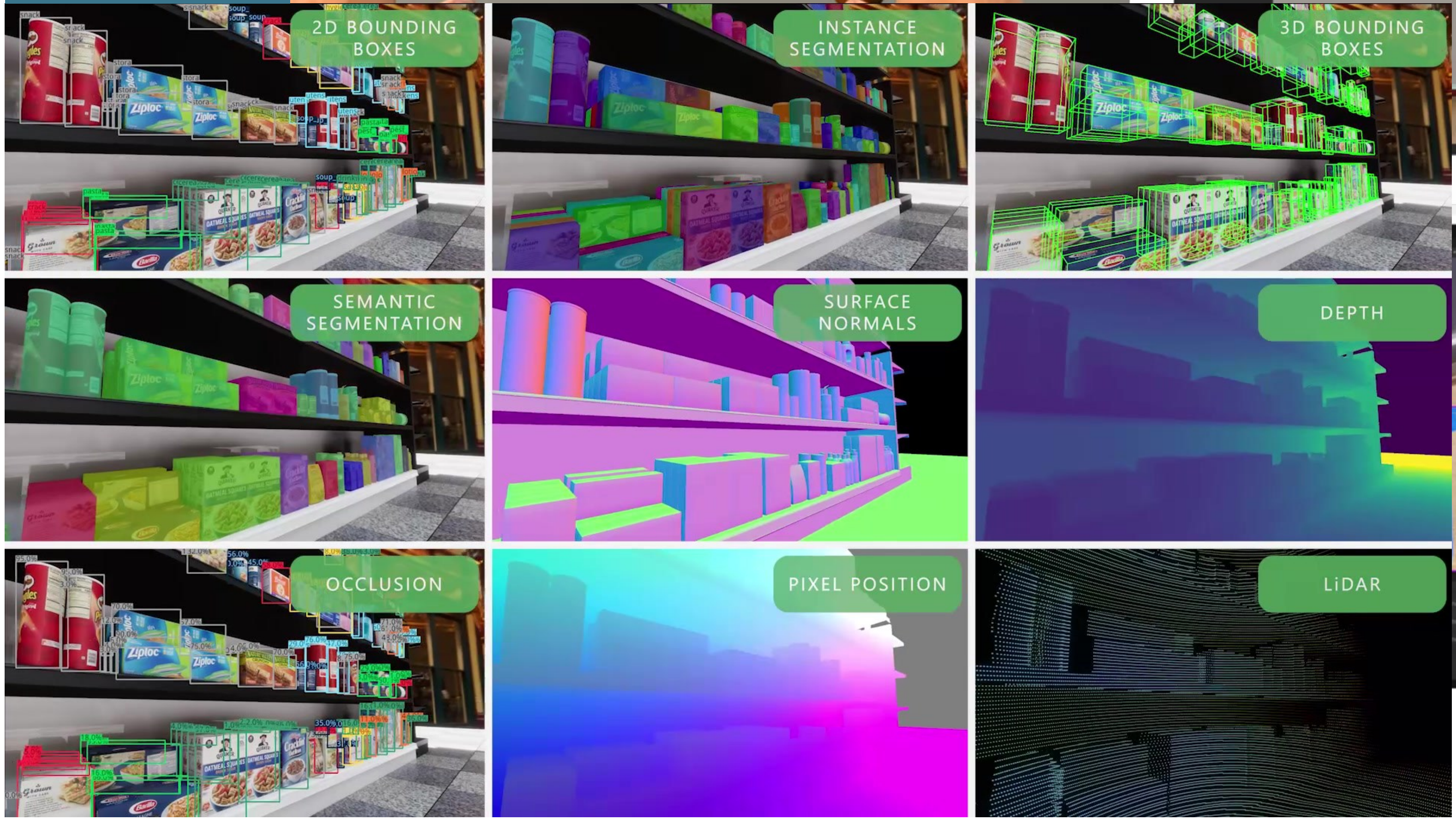
Challenges in dataset collection and annotation

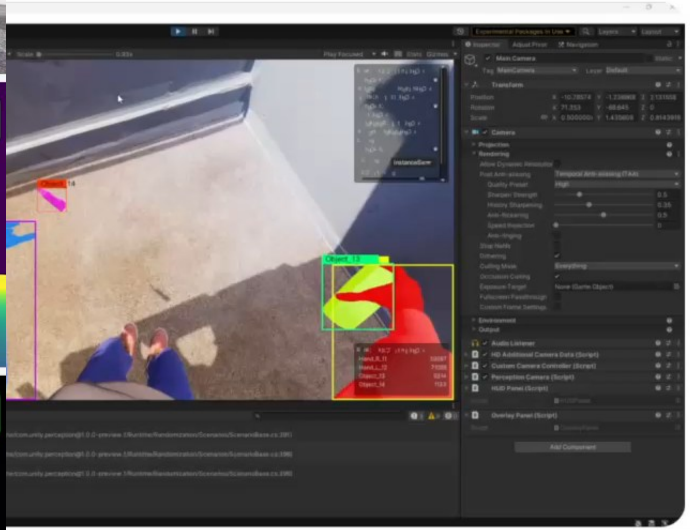


Synthetic Data: Automatic Labeling

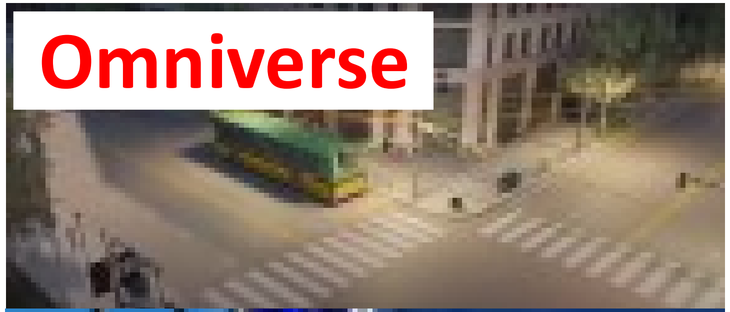


Labels





Data Generation Pipeline



Generic vs In-Domain Data

"Ideal" Synthetic Data should cover a wide range of scenarios and conditions, accurately representing the variability and complexity present in real-world data.

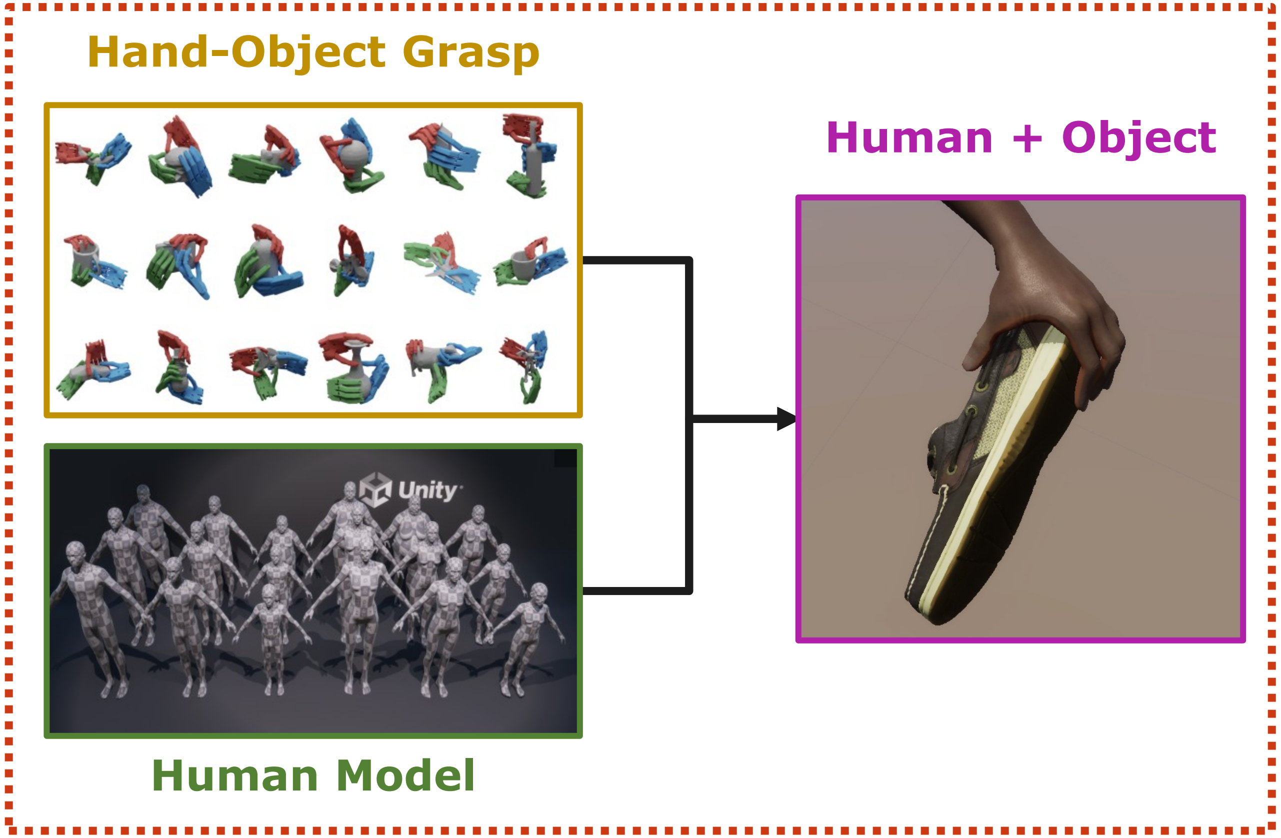
Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

HOI-Synth: Data Generation Pipeline

Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

HOI-Synth: Data Generation Pipeline

(a) Hand-Object Interaction Generation



DexGraspNet

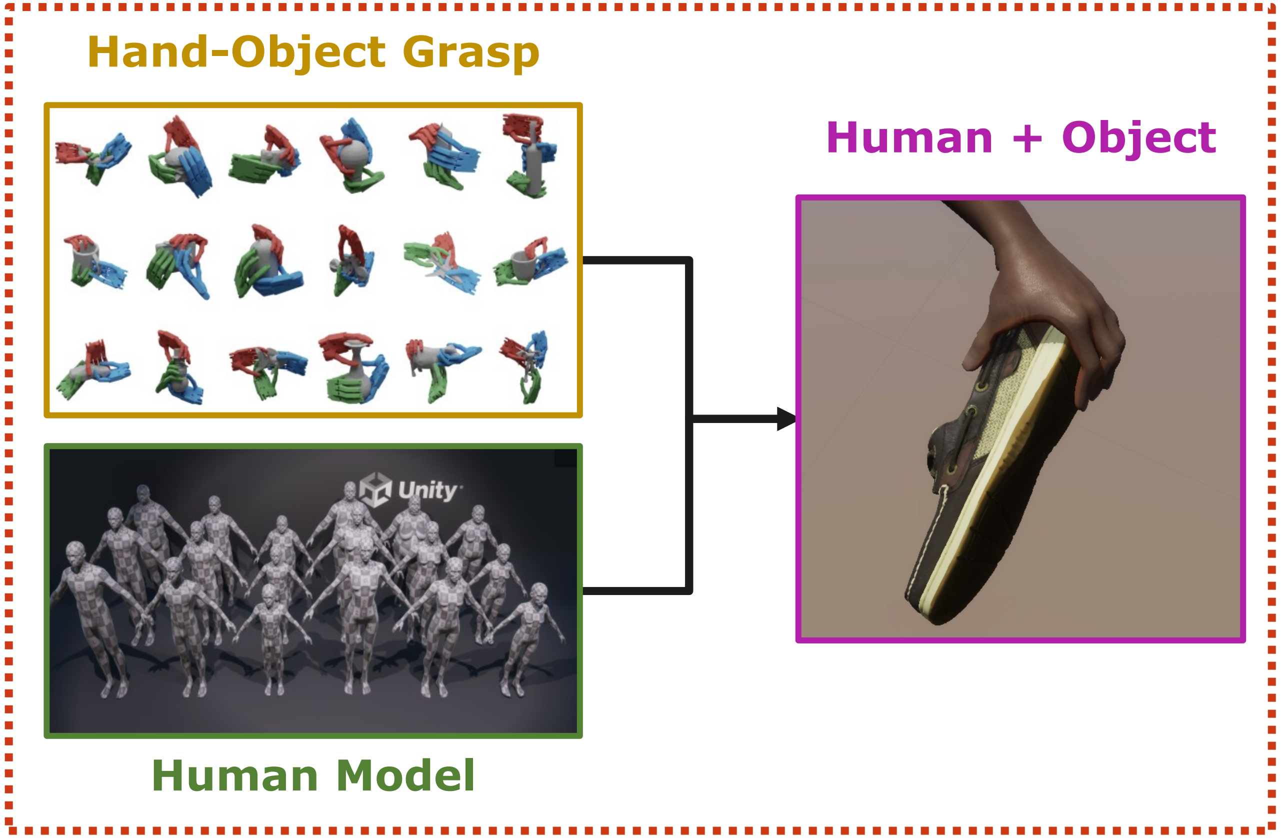
- 1.32 million grasps
- 5355 objects

Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

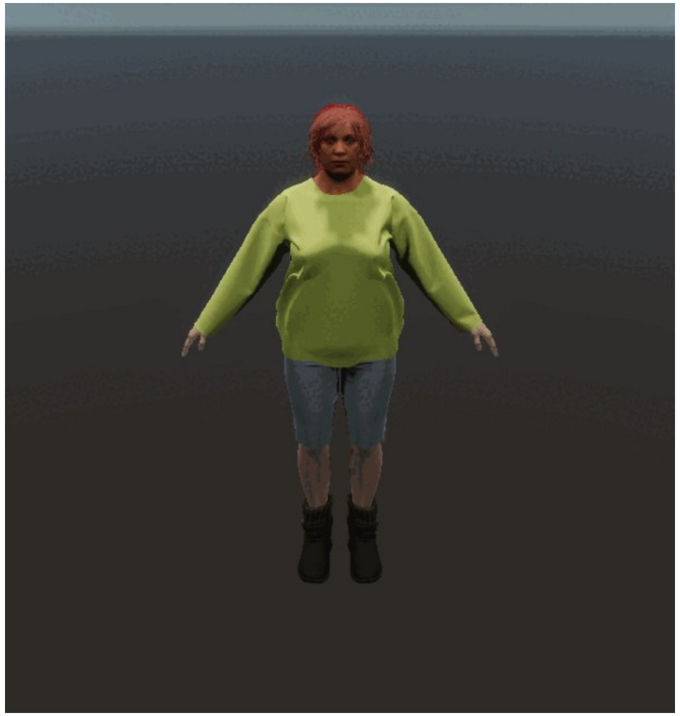
Università di Catania

HOI-Synth: Data Generation Pipeline

(a) Hand-Object Interaction Generation



SyntheticHumans



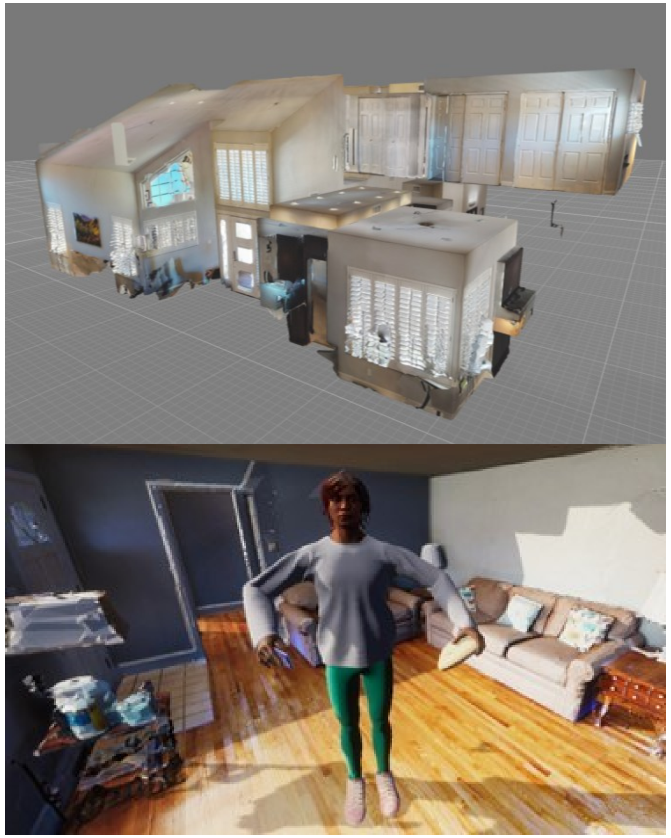
Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

HOI-Synth: Data Generation Pipeline

Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

HOI-Synth: Data Generation Pipeline

(b) Environment Selection and Human Placement

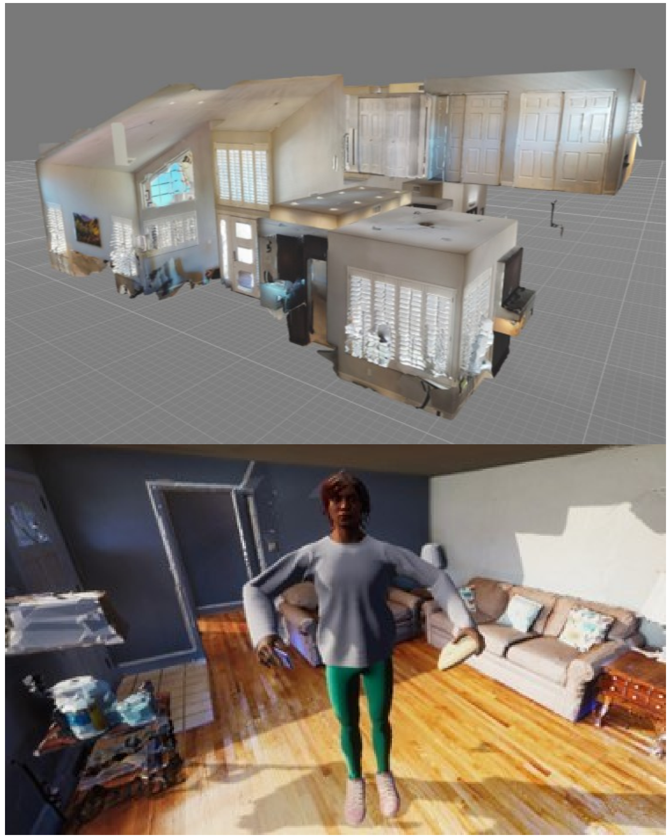


Habitat - Matterport 3D Research Dataset (HM3D)

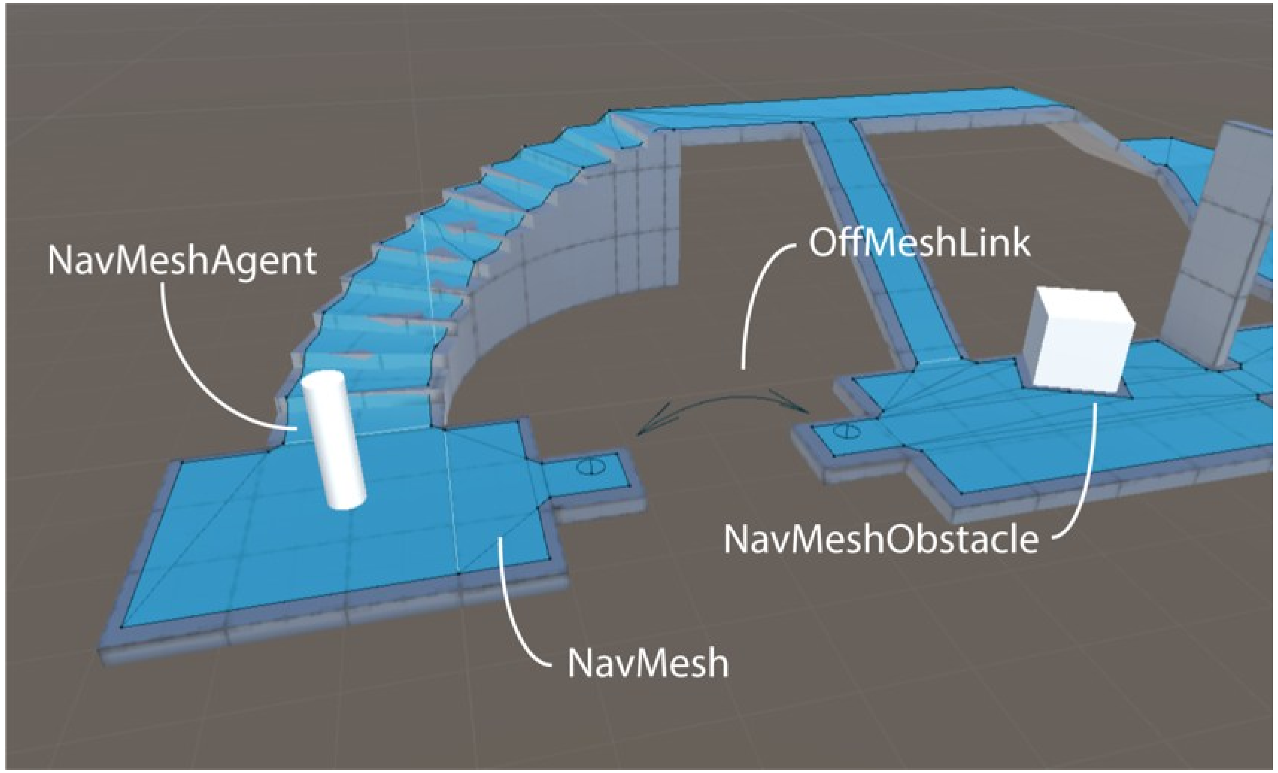
Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

HOI-Synth: Data Generation Pipeline

(b) Environment Selection and Human Placement



Navigation (NavMesh) System (Unity)



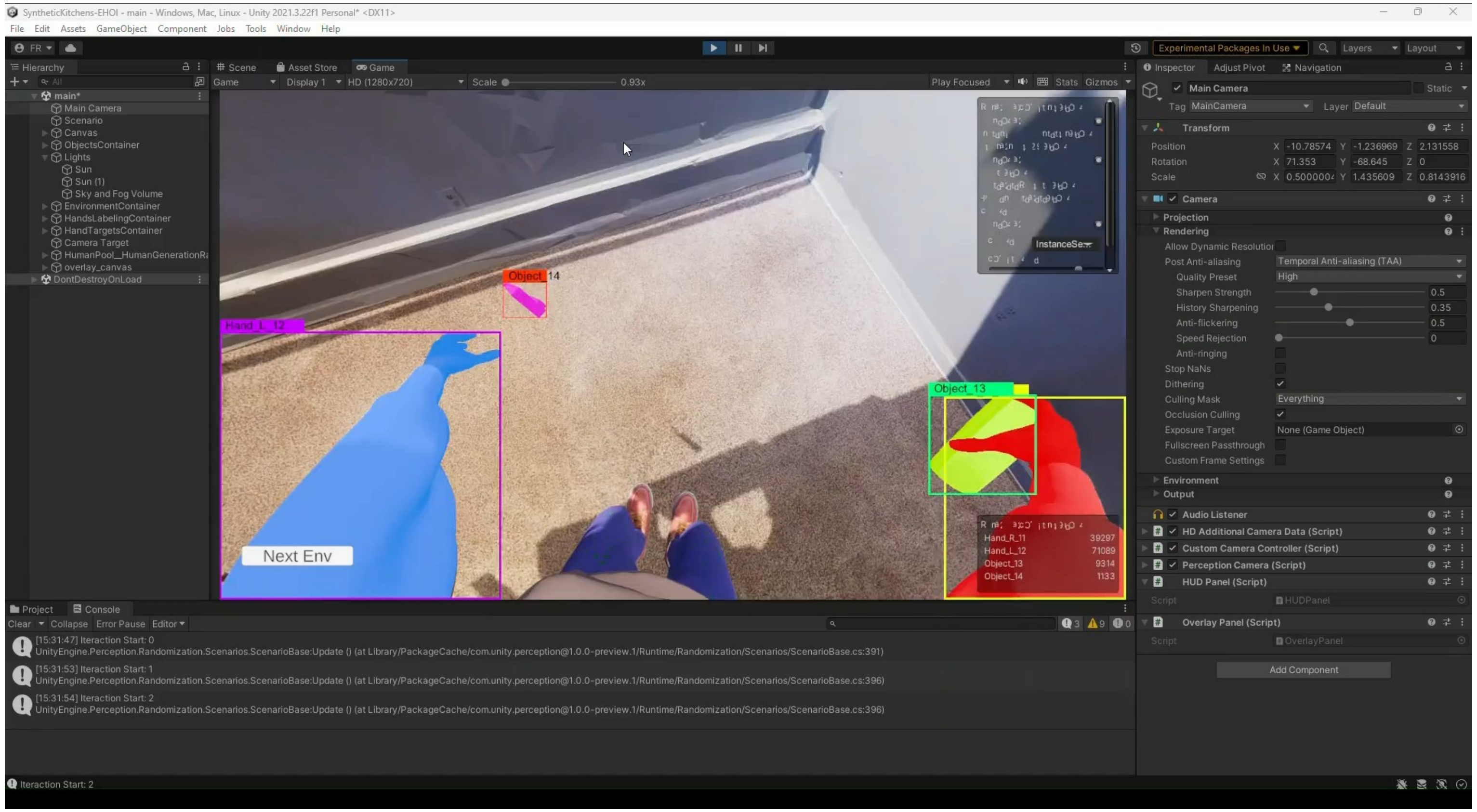
Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

HOI-Synth: Data Generation Pipeline

Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

HOI-Synth: Data Generation Pipeline

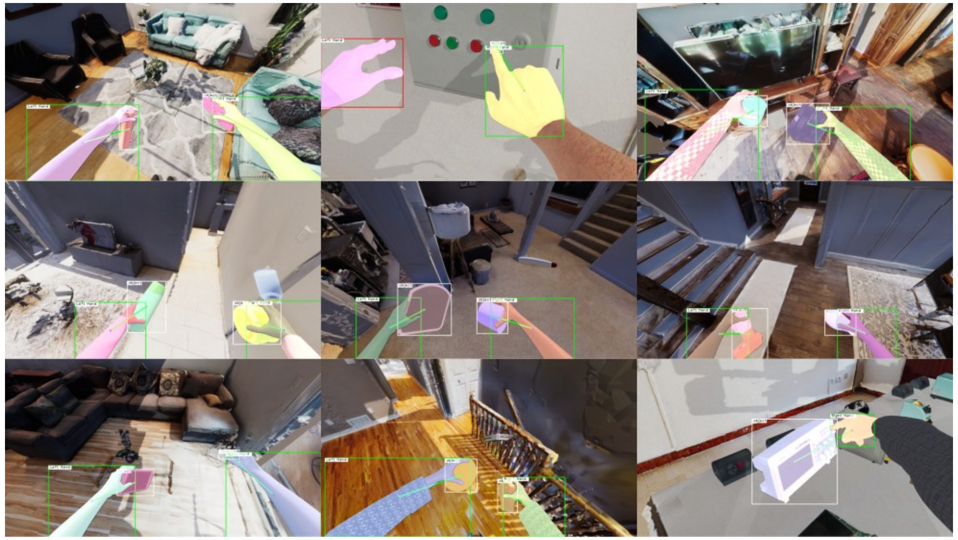
(c) Generated Labeled Egocentric Data



Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

HOI-Synth Benchmark

GENERATED EGOCENTRIC SYNTHETIC DATA

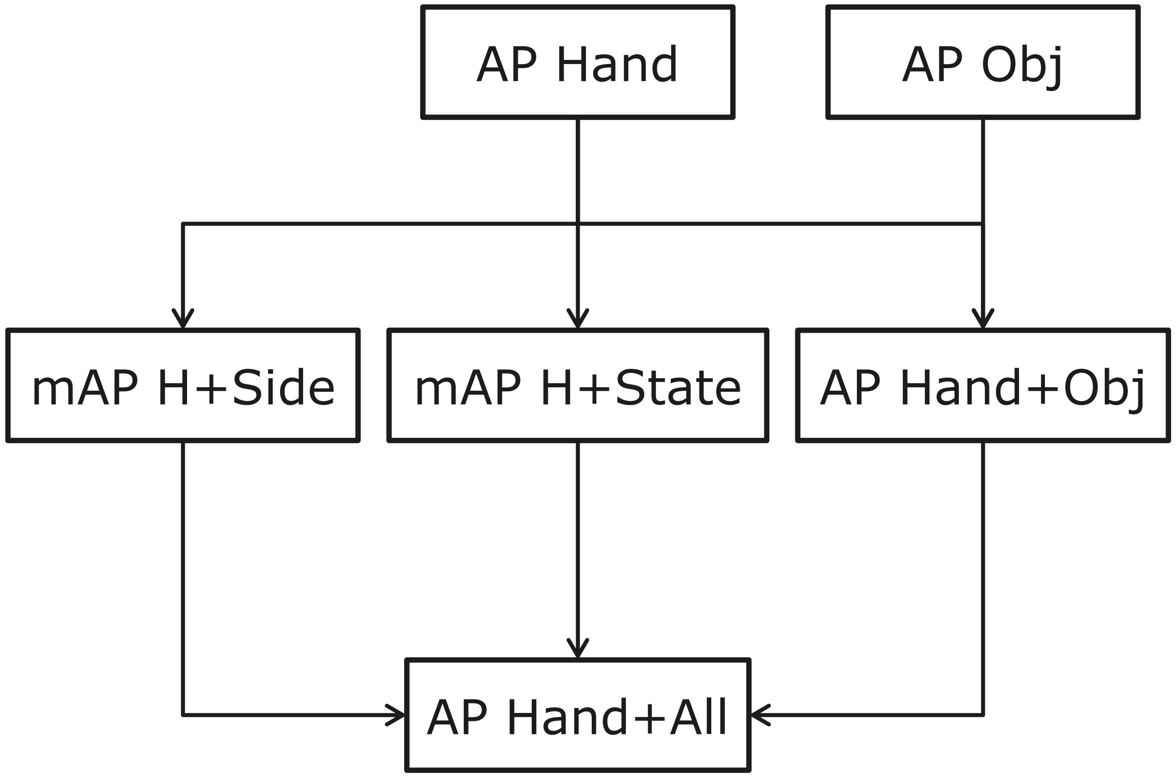


- **Epic-Kitchens VISOR**
  - 32,857 real + 30,259 synthetic images
- **EgoHOS**
  - 8,107 real + 8,107 synthetic images
- **ENIGMA-51**
  - 3,479 real images
  - In-Domain 16,773 + Out-domain 20,321 synthetic images

Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

3. Models and Architectures for Hand-Object Interactions Detection

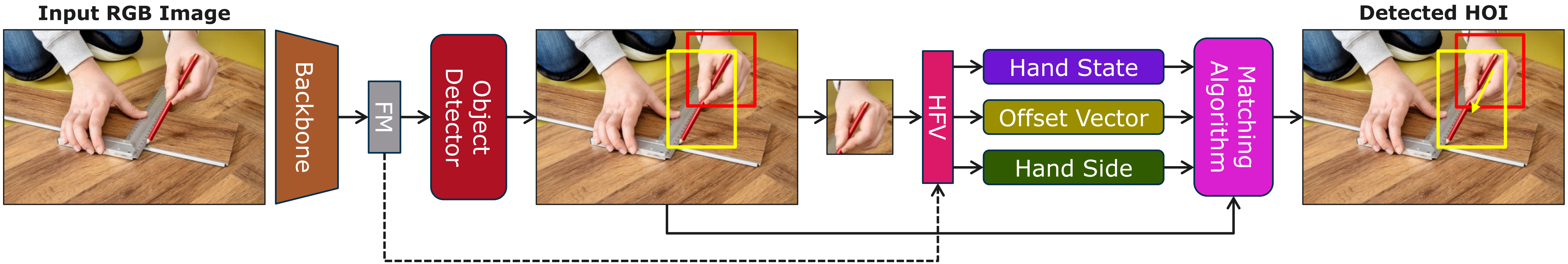
Evaluation Protocols: mAP with criterion



Shan Dandan, et al. "Understanding human hands in contact at internet scale" (CVPR 2020)

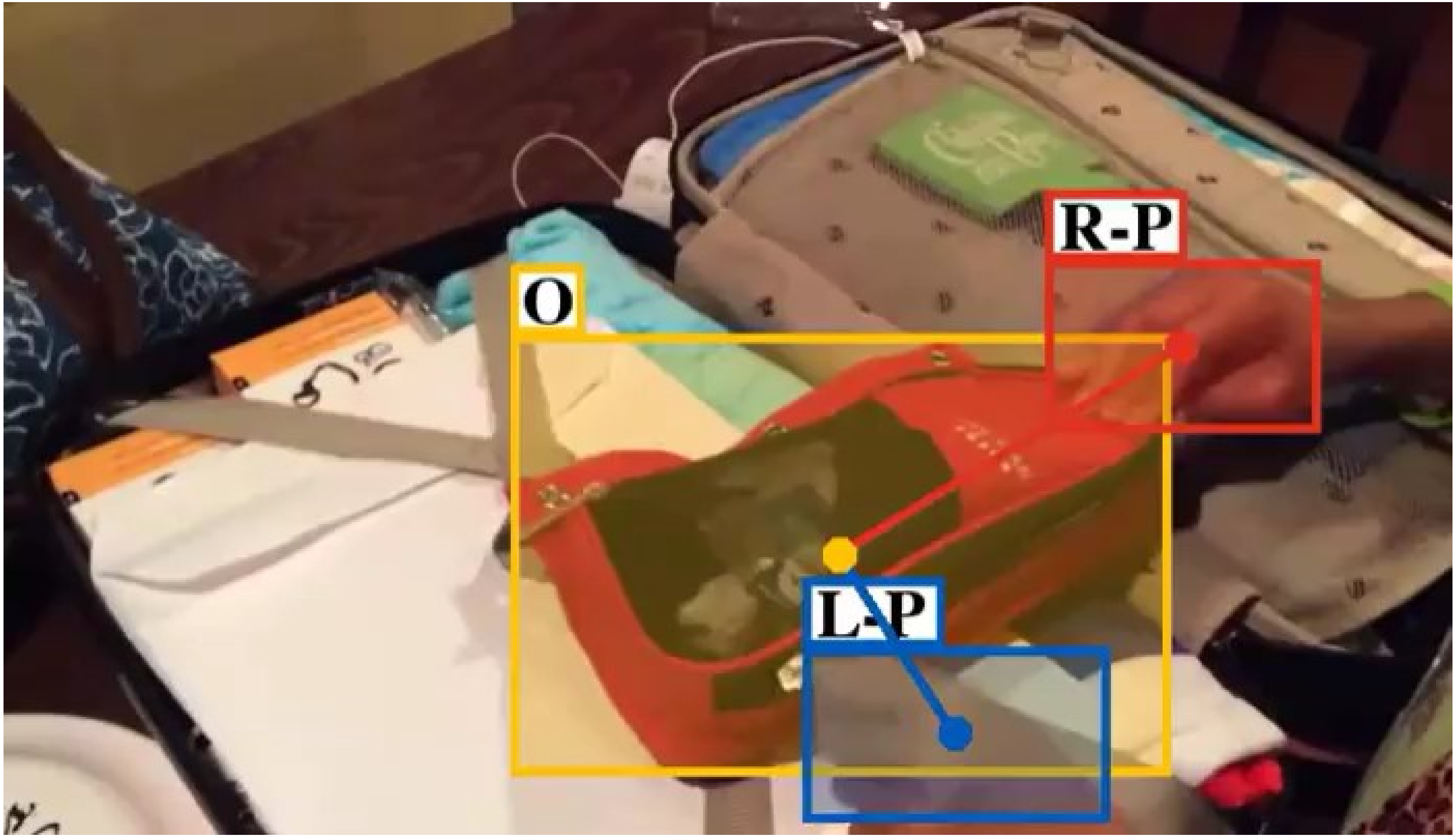
FM Features Map

HFV Hand Features Map



Shan Dandan, et al. "Understanding human hands in contact at internet scale" (CVPR 2020)

100 Days of Hands: Qualitative Results



Shan Dandan, et al. "Understanding human hands in contact at internet scale" (CVPR 2020)

100 Days of Hands: Quantitative Results

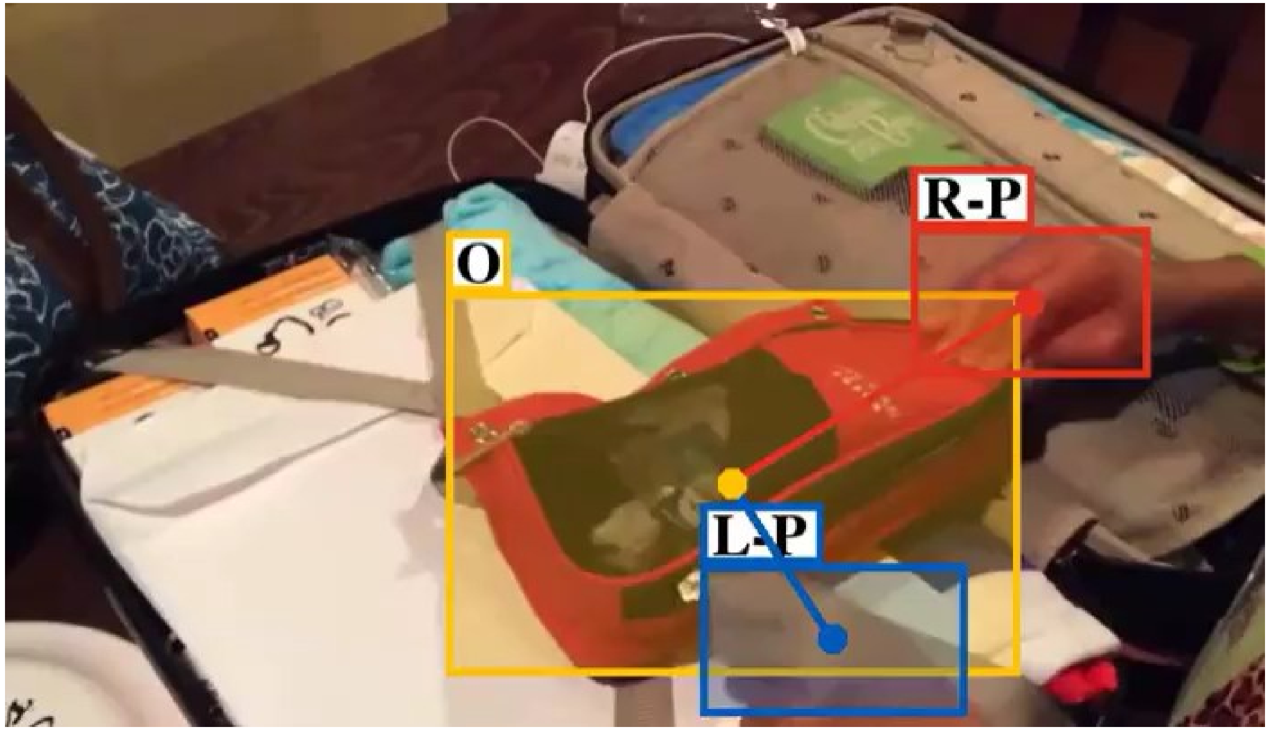
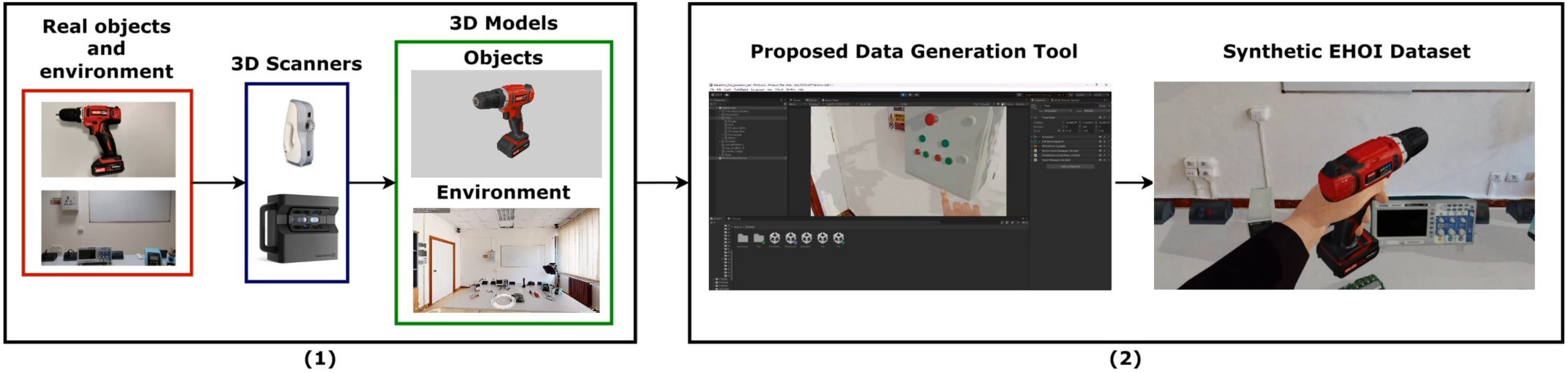


Table 4. Average Precision when we vary definitions of correct detection. Using 15K samples dramatically degrades performance on any category, and 45K samples produces a large drop on getting all outputs correct.

	Hand	Obj	H+Side	H+State	H+O	All
Full	89.6	63.9	78.9	64.0	46.9	38.5
45K	88.4	61.0	77.3	62.7	39.3	31.3
15K	80.9	54.2	66.8	53.3	27.3	19.7

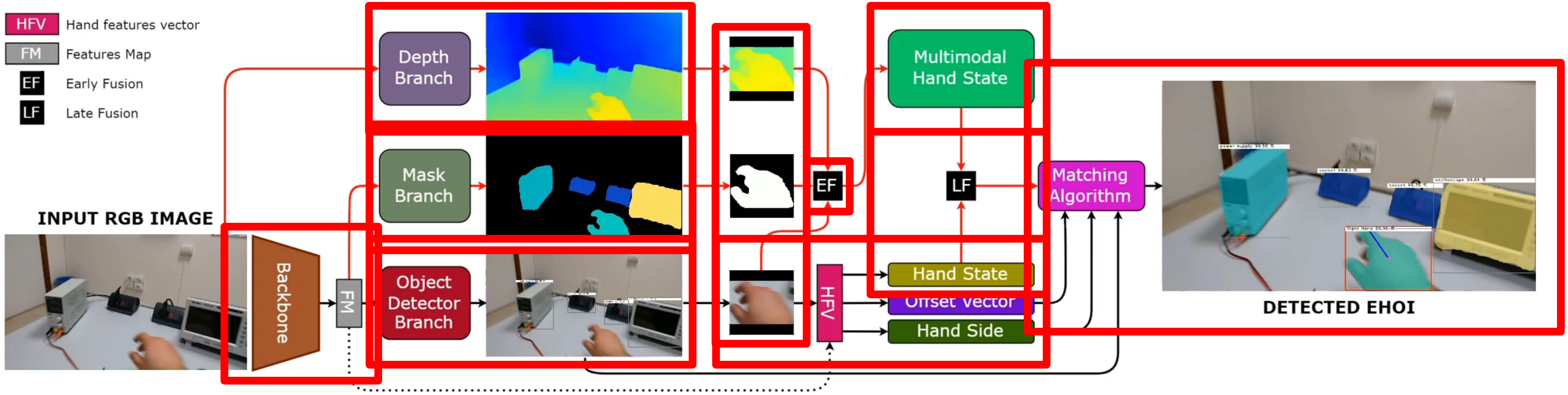
Shan Dandan, et al. "Understanding human hands in contact at internet scale" (CVPR 2020)

Proposed Synthetic Data Generation Pipeline



Leonardi Rosario, et al. "Exploiting multimodal synthetic data for egocentric human-object interaction detection in an industrial scenario" (CVIU 2024)

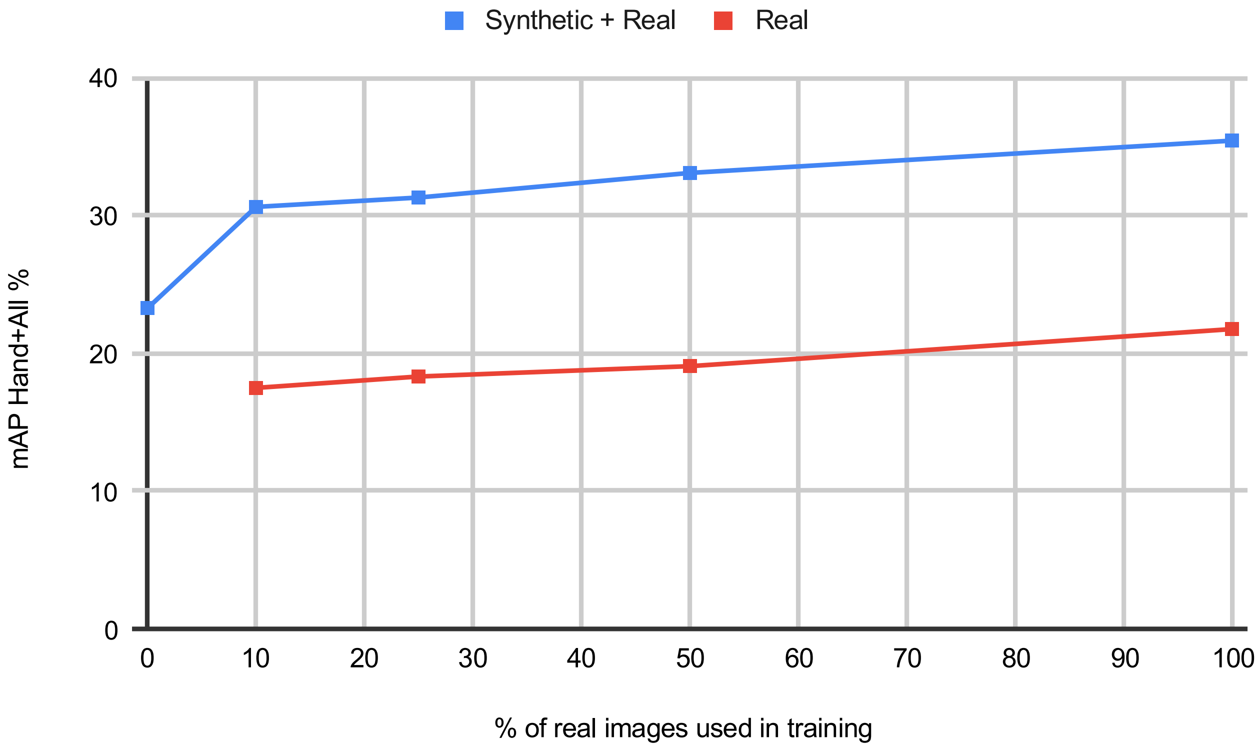
Proposed Approach



Leonardi Rosario, et al. "Exploiting multimodal synthetic data for egocentric human-object interaction detection in an industrial scenario" (CVIU 2024)

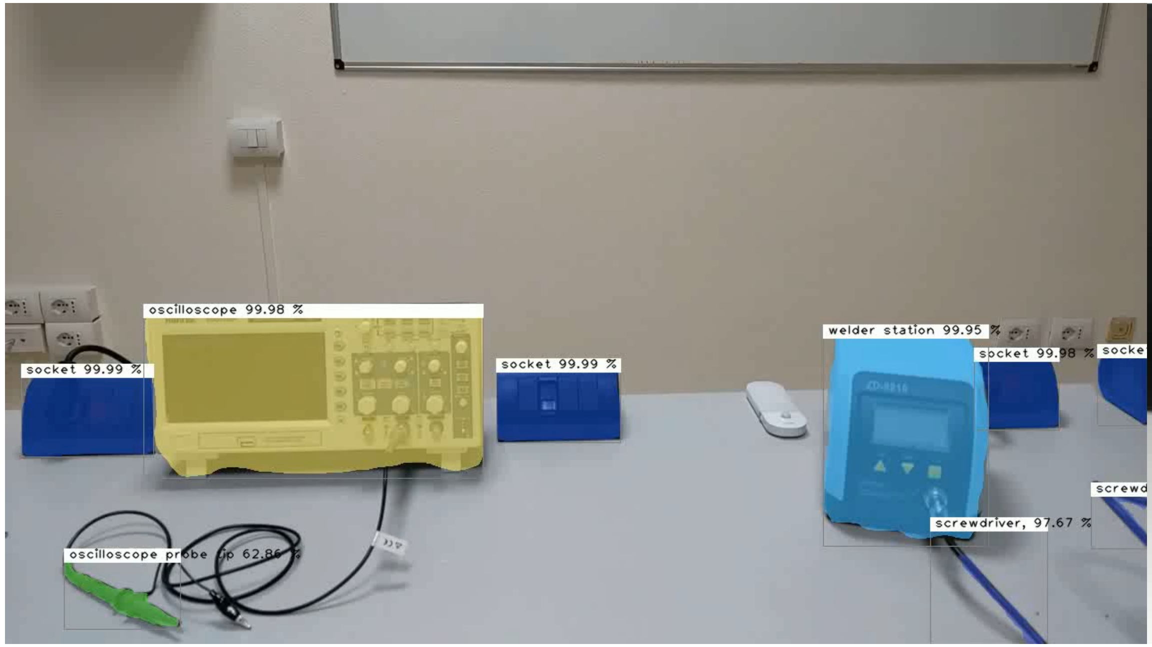
Quantitative Results

Contact state	MHS Input Modalities	AP H+State	mAP H+All
HS	-	56.88	35.47
MHS	RGB+DEPTH+MASK	57.56	35.81
HS+MHS	RGB	58.29	35.71
HS+MHS	RGB+DEPTH	58.37	35.92
HS+MHS	RGB+MASK	58.30	35.34
FULL	RGB+DEPTH+MASK	**58.40**	**36.51**

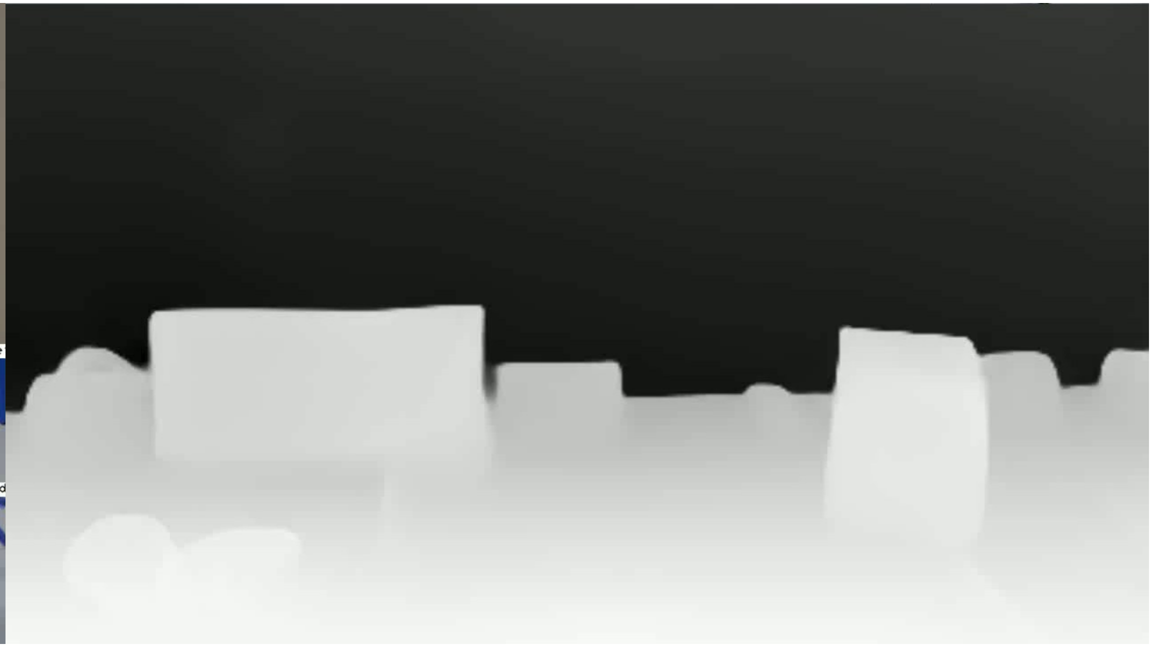


Leonardi Rosario, et al. "Exploiting multimodal synthetic data for egocentric human-object interaction detection in an industrial scenario" (CVIU 2024)

Qualitative Results



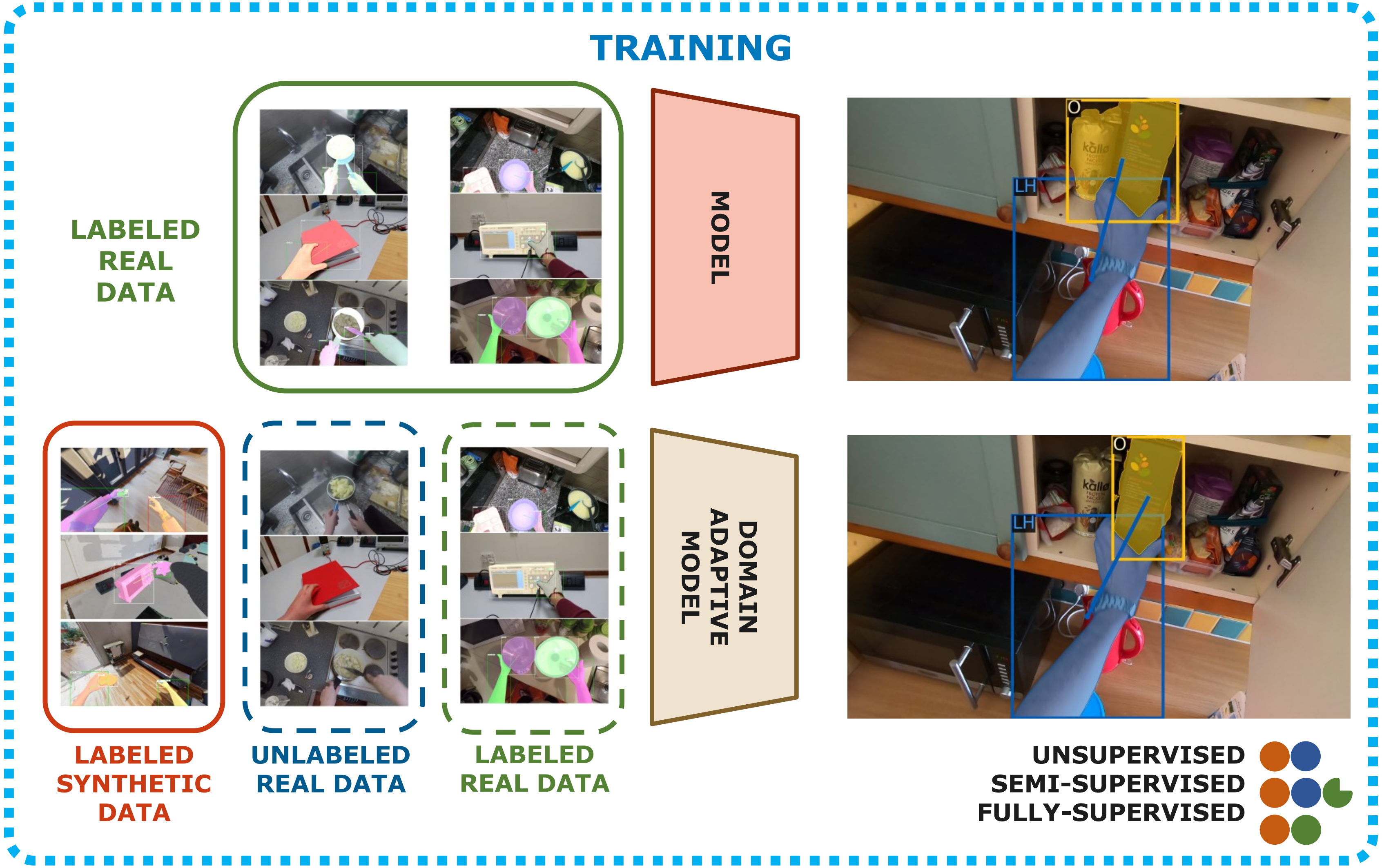
RGB



Estimated Monocular Depth Estimation

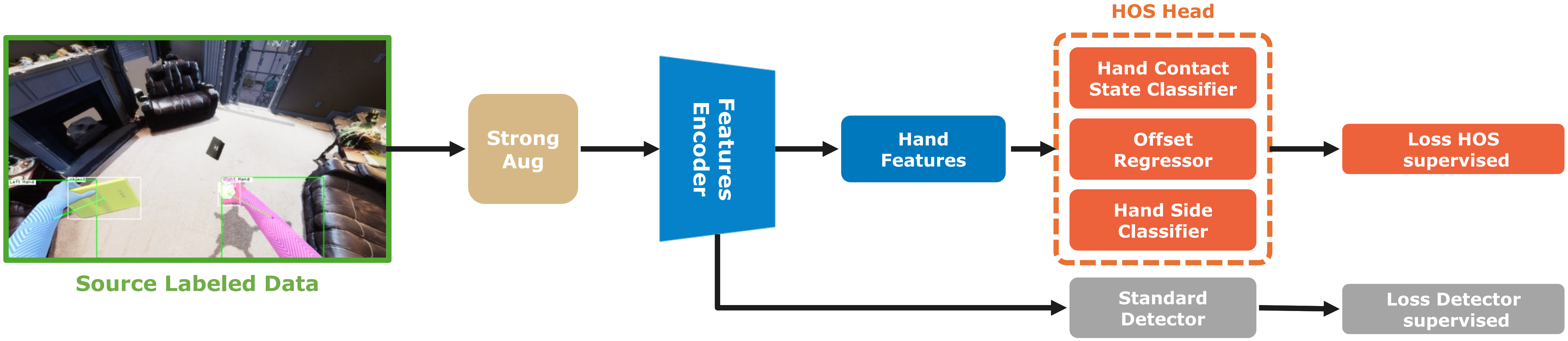
Leonardi Rosario, et al. "Exploiting multimodal synthetic data for egocentric human-object interaction detection in an industrial scenario" (CVIU 2024)

Domain Adaptation



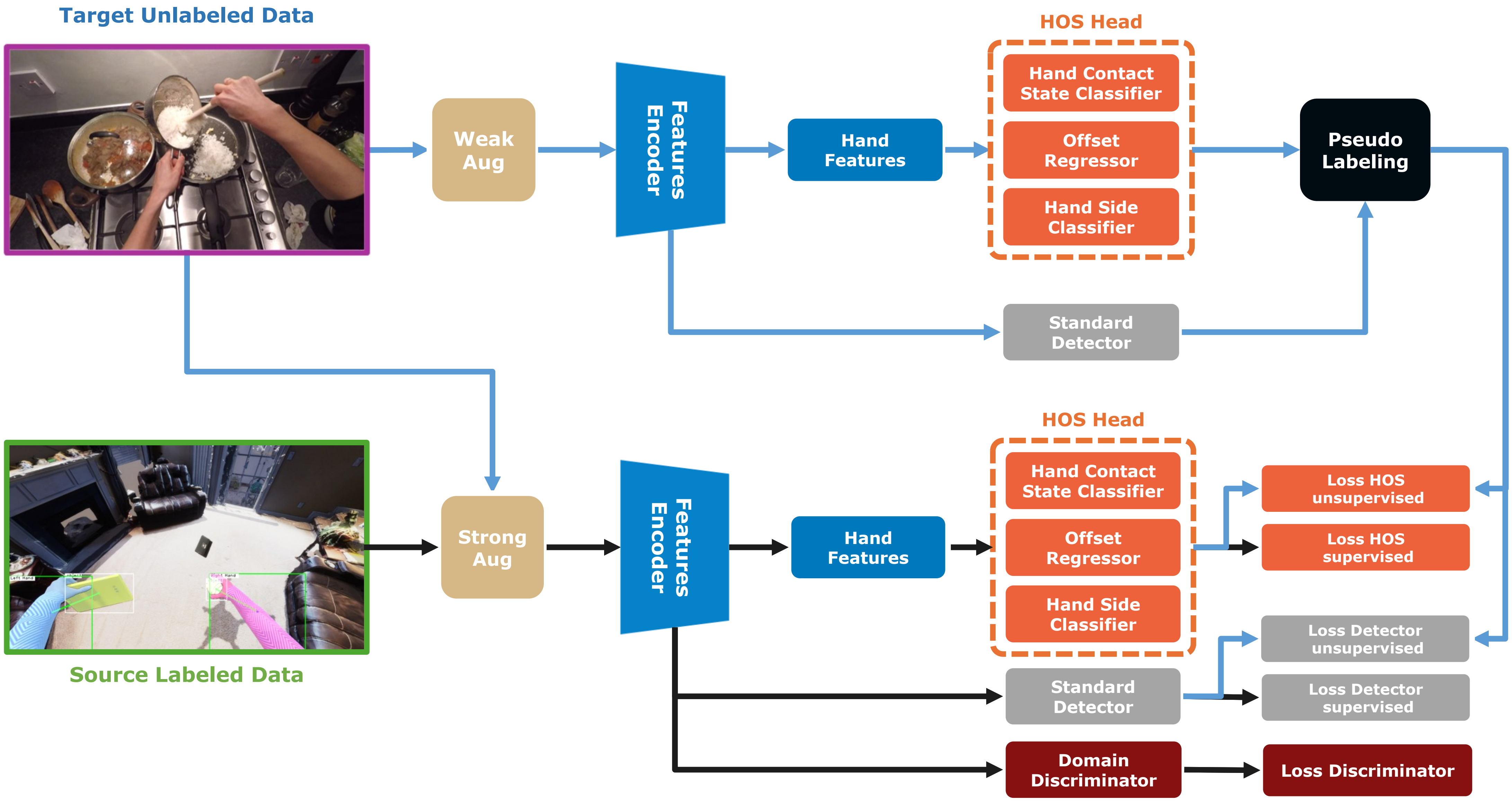
Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

Proposed Approach: Burn-in



Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

Proposed Approach: Teacher->Student



Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

Quantitative Results: Epic-Kitchens VISOR

a) Unsupervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
0%	Synthetic-Only	09.88	28.41	24.89	08.64	01.23
	UDA	**33.33**	**80.16**	**65.98**	**33.47**	**08.35**
Absolute Improvement		**+23.45**	+51.75	+41.09	+24.83	+7.12

b) Semi-supervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
25% (8,215 images)	Real-Only	37.90	90.14	**85.66**	53.99	17.85
	Synthetic+Real	38.19	89.98	84.67	**55.88**	18.49
	SSDA	**45.55**	**90.37**	84.42	52.59	**22.15**
Absolute Improvement		**+7.65**	+0.23	−0.99	+1.89	+4.30

C) Fully-supervised Setting

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

Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

Quantitative Results: Epic-Kitchens VISOR

a) Unsupervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
0%	Synthetic-Only	09.88	28.41	24.89	08.64	01.23
	UDA	**33.33**	**80.16**	**65.98**	**33.47**	**08.35**
Absolute Improvement		**+23.45**	+51.75	+41.09	+24.83	+7.12

b) Semi-supervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
25% (8,215 images)	Real-Only	37.90	90.14	**85.66**	53.99	17.85
	Synthetic+Real	38.19	89.98	84.67	**55.88**	18.49
	SSDA	**45.55**	**90.37**	84.42	52.59	**22.15**
Absolute Improvement		**+7.65**	+0.23	−0.99	+1.89	+4.30

C) Fully-supervised Setting

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

Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

Quantitative Results: Epic-Kitchens VISOR

a) Unsupervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
0%	Synthetic-Only	09.88	28.41	24.89	08.64	01.23
	UDA	**33.33**	**80.16**	**65.98**	**33.47**	**08.35**
Absolute Improvement		**+23.45**	+51.75	+41.09	+24.83	+7.12

b) Semi-supervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
25% (8,215 images)	Real-Only	37.90	90.14	**85.66**	53.99	17.85
	Synthetic+Real	38.19	89.98	84.67	**55.88**	18.49
	SSDA	**45.55**	**90.37**	84.42	52.59	**22.15**
Absolute Improvement		**+7.65**	+0.23	−0.99	+1.89	+4.30

C) Fully-supervised Setting

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

Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

Quantitative Results: EgoHOS

a) Unsupervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
0%	Synthetic-Only	07.16	18.25	15.93	05.33	01.24
	UDA	**28.16**	**70.30**	**59.21**	**20.84**	**09.65**
Absolute Improvement		**+21.00**	+52.05	+43.28	+15.51	+8.41

b) Semi-supervised Setting

% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
25% (2,026 images)	Real-Only	33.73	78.94	70.62	41.67	21.83
	Synthetic+Real	33.78	79.60	71.61	46.11	19.87
	SSDA	**37.16**	**83.79**	**74.28**	**49.00**	**23.82**
Absolute Improvement		**+3.43**	+4.85	+3.66	+7.33	+1.99

C) Fully-supervised Setting

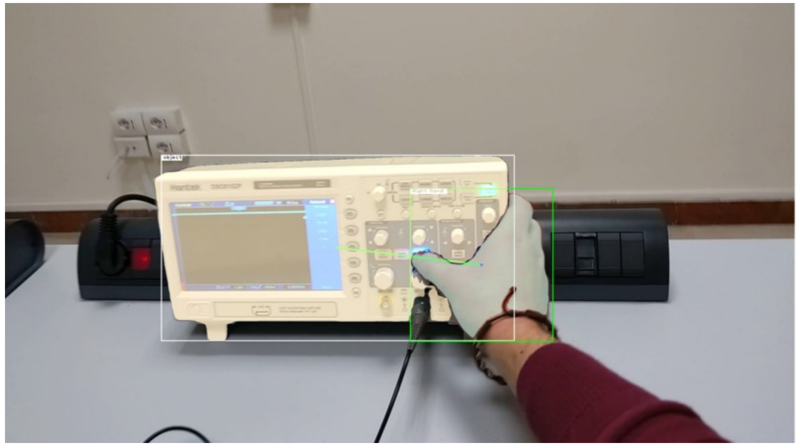
% Real Labeled Data	Approach	Overall	H	H+S	H+C	O
100% (8,758 images)	Real-Only	36.16	84.39	76.24	51.81	26.46
	Synthetic+Real	34.68	84.56	71.56	49.72	23.16
	FSDA	**39.61**	**85.58**	**76.80**	**51.99**	**27.05**
Absolute Improvement		**+3.45**	+1.19	+0.56	+0.18	+0.59

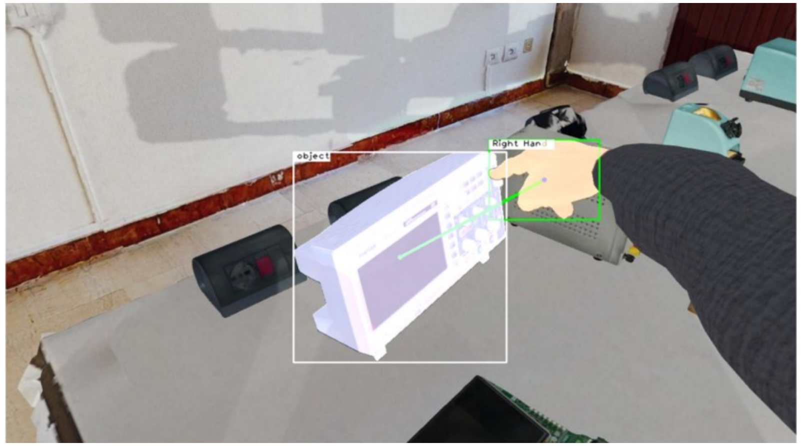
Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

Quantitative Results: ENIGMA-51

a) Unsupervised Setting

% Real Labeled Data	Approach	In-domain	Overall	H	H+C	O
0%	Synthetic-Only		5.67	15.78	2.66	2.31
	Synthetic-Only	✓	12.85	56.05	15.24	4.79
	UDA		18.57	58.17	17.74	13.15
	UDA	✓	**34.78**	**78.83**	**28.14**	**25.84**
Absolute Improvement			+16.21	+20.66	+10.40	+12.69





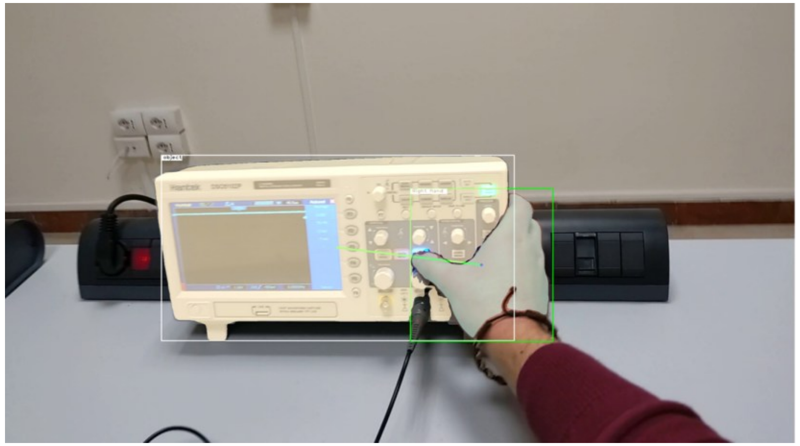
Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

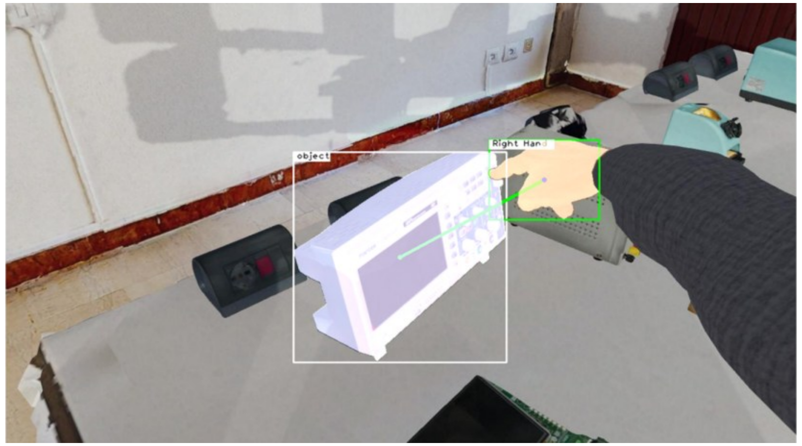
Università di Catania

Quantitative Results: ENIGMA-51

a) Supervised Setting

% Real Labeled Data	Approach	In-domain	Overall	H	H+C	O
100%	Real-only	✓	63.84	85.01	52.32	51.35
	FSDA		**64.41**	**85.94**	**54.13**	52.50
	FSDA	✓	64.20	85.37	51.60	**53.30**
Absolute Improvement			+0.57	+0.93	+1.81	+1.95





Leonardi Rosario, et al. "Are Synthetic Data Useful for Egocentric Hand-Object Interaction Detection?" (ECCV 2024)

4. Open Challenges

Open Challenges

- Standardization of evaluation protocols and task definitions!!!!

mIoU Hand & Object Mask

Hand Object Interaction Detection

Egocentric Hand Object Interaction Detection

mAP All

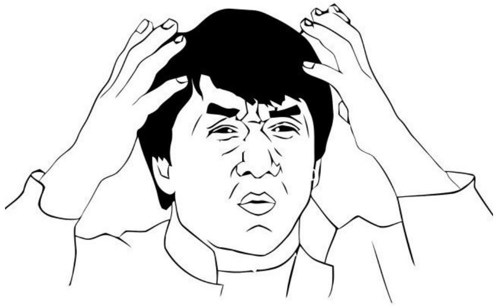
Egocentric Hand-Object Segmentation

Egocentric Human Object Interaction Detection

mAP Hand + All

Hand Object Segmentation

- Updating and improving pre-existing architectures
- Handling occlusions and complex interactions
- Integration of multimodal and temporal data



Open Challenges

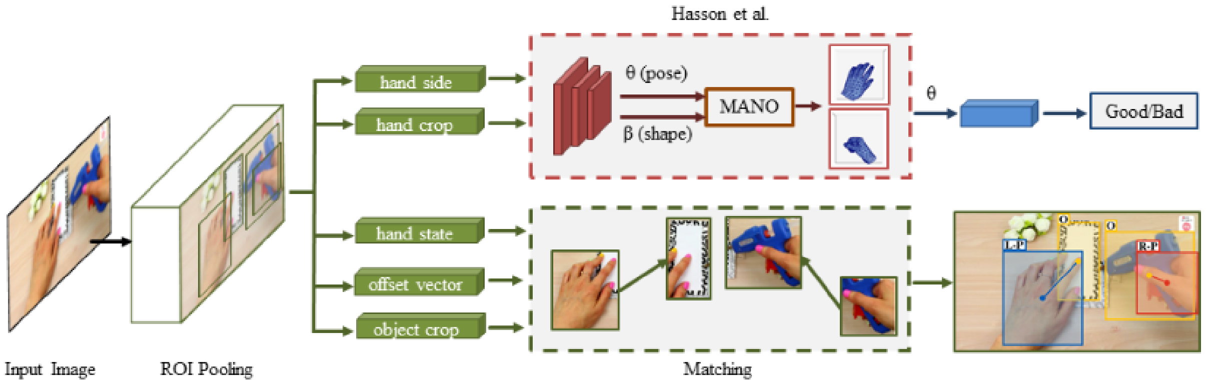
- Standardization of evaluation protocols and task definitions!!!!

- Updating and improving pre-existing architectures

Hand Object Detector

This is the code for our paper *Understanding Human Hands in Contact at Internet Scale* (CVPR 2020, **Oral**).

Dandan Shan, Jiaqi Geng*, Michelle Shu*, David F. Fouhey



Watch 5 | Fork 70 | Star 274

2020

- Handling occlusions and complex interactions
- Integration of multimodal and temporal data

Open Challenges

- Standardization of evaluation protocols and task definitions!!!!

- Updating and improving pre-existing architectures

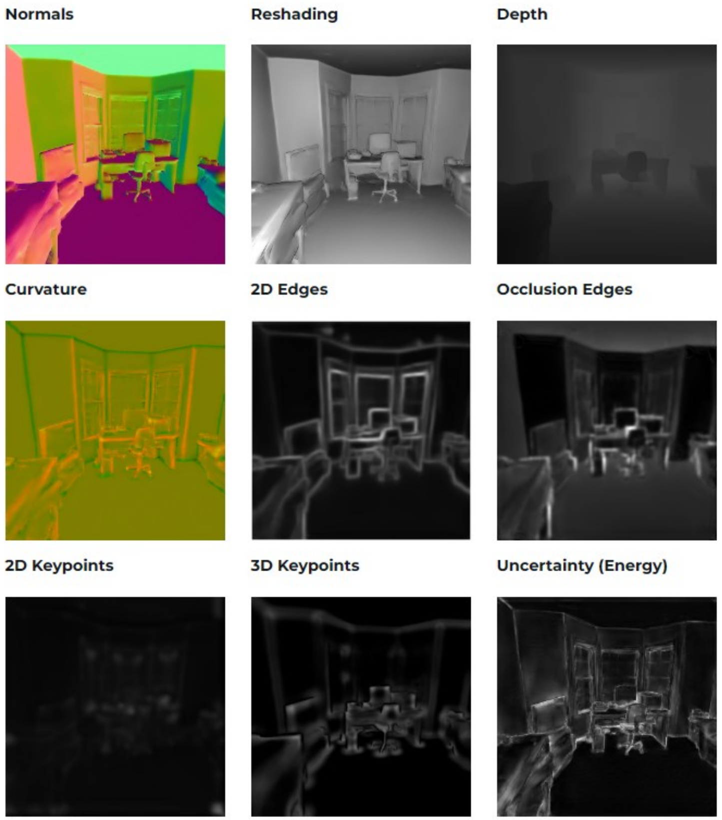
- Handling occlusions and complex interactions

- Integration of multimodal and temporal data

Open Challenges

- Standardization of evaluation protocols and task definitions!!!!
- Updating and improving pre-existing architectures
- Handling occlusions and complex interactions
- Integration of multimodal and temporal data

---------->



Open Challenges

- Standardization of evaluation protocols and task definitions!!!!
- Updating and improving pre-existing architectures
- Handling occlusions and complex interactions

- Integration of multimodal and temporal data

Static Frame **----------->** **Video Shot**

Hand-Object Interactions in Egocentric Vision

Thank you!

Rosario Leonardi

LIVE Group @ UNICT - https://iplab.dmi.unict.it/live/

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

Department of Mathematics and Computer Science - University of Catania

rosario.leonardi@unict.it