Understanding and Reconstructing Hand-Object Interactions (HOIs) from Egocentric Videos

University of BRISTOL

Dima Damen
Rhobin W @CVPR2024

Egocentric Videos?

In this talk

HOI in 2D

- VISOR (masks and hand-interactions)
- HOI-Ref
- GenHowTo

HOI 3D Reconstruction in view

- Get a Grip

HOI 3D Reconstruction in and out of view

- EPIC Fields - Scene reconstruction from egocentric views
- OSNOM - 3D tracking of HOI in world coordinate frames

In this talk

HOI in 2D

- VISOR (masks and hand-interactions)
- HOI-Ref
- GenHowTo

HOI 3D Reconstruction in view

- Get a Grip

HOI 3D Reconstruction in and out of view

- EPIC Fields - Scene reconstruction from egocentric views
- OSNOM - 3D tracking of HOI in world coordinate frames

EPIC-KITCHENS VISOR

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma,
Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

Dima Damen
Rhobin W @CVPR2024

VISOR is….

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler, Dima Damen

pour spice

left hand
right hand
hob
saucepan
spice
spice container
spoon
soup
pepper container lid

VISOR is….

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler, Dima Damen

pour spice

left hand
right hand
hob
saucepan
spice
spice container
spoon
soup
pepper container lid

saucepan → pan → cookware

spoon → spoon → cutlery

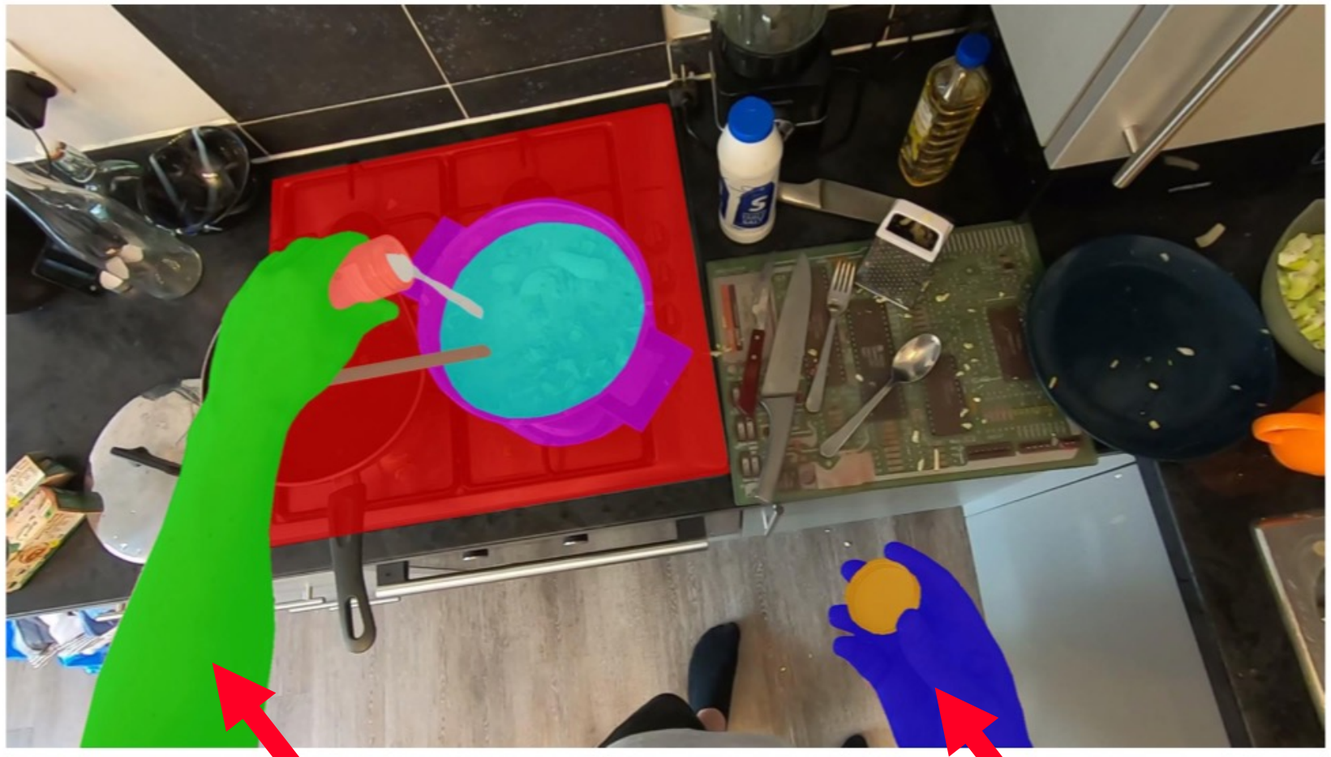
A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

VISOR is....

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler, Dima Damen

action

pour spice



left hand

right hand

hob

saucepan

spice

spice container

spoon

soup

pepper container lid

in-contact (spice container)

in-contact (container lid)

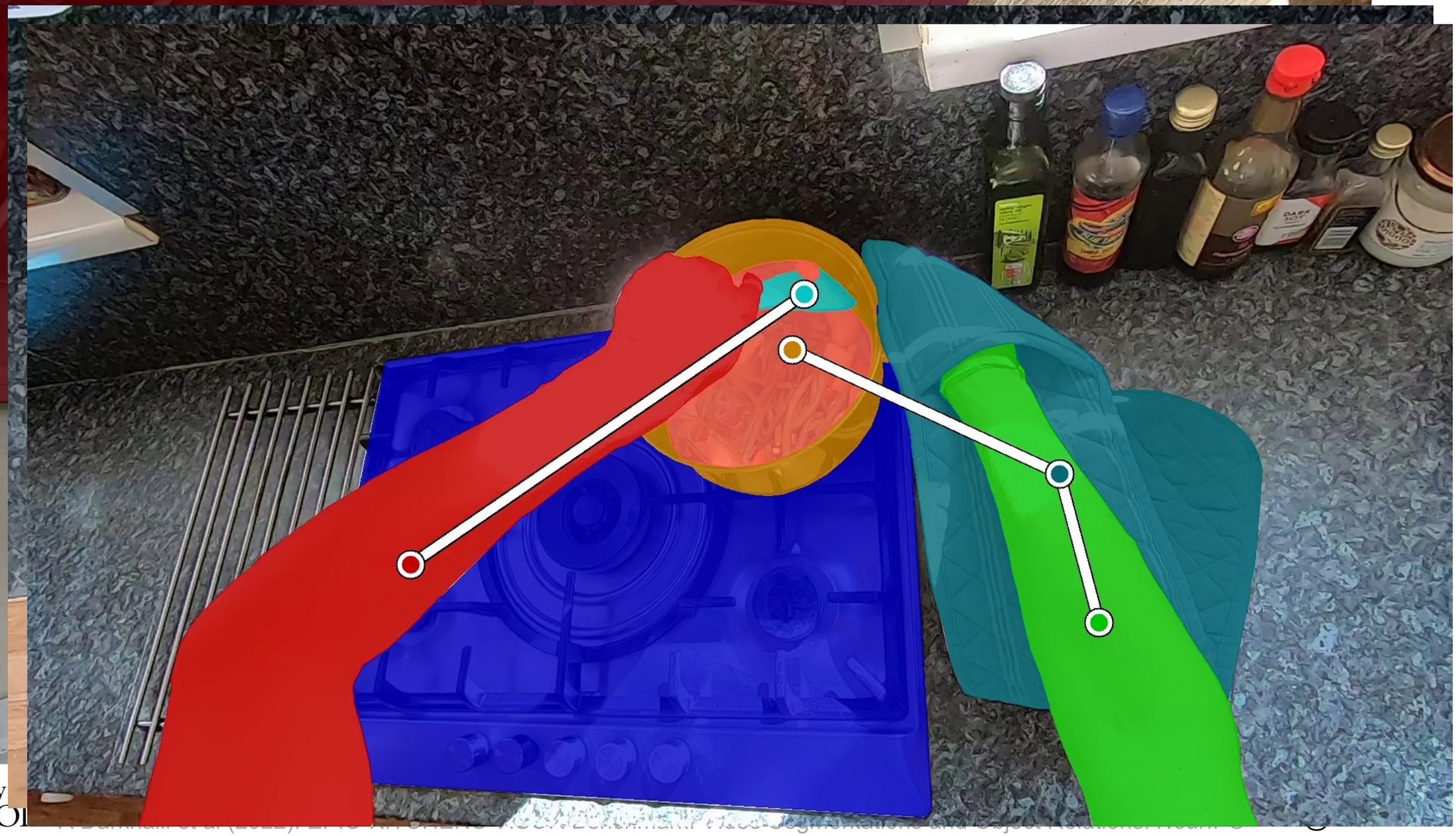
Comparative Stats

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler, Dima Damen

	Basic Statistics		Pixel-Level Annotations	Action Annotations		
Dataset	Total Mins	Avg Seq Ln	Total Masks	Actions	#Action Classes	#Entity Classes
EgoHand [3]	72	-	15.1K	-	-	2
DAVIS [6]	8	3s	32.0K	-	-	-
YTVOS [43]	335	5s	197.2K	-	-	94
UVOv0.5 (Sparse) [41]	511	3s	*200.6K	10,213	300	-
VISOR (Ours)	**2,180**	**12s**†	**271.6K**	**27,961**	**2,594**	**257**

VISOR Relations

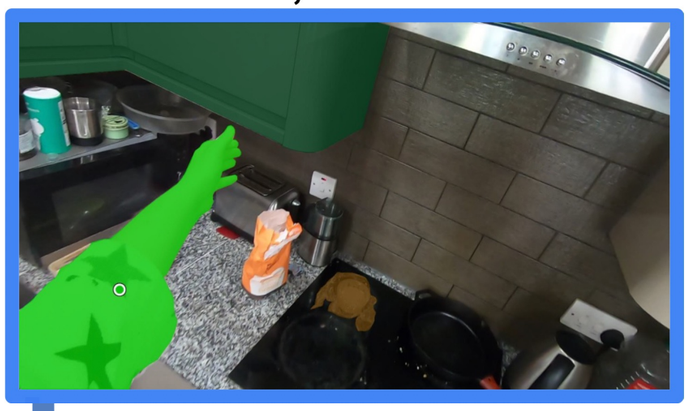
with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar,
Richard Higgins, David Fouhey, Sanja Fidler, Dima Damen



Object relation stats

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler, Dima Damen

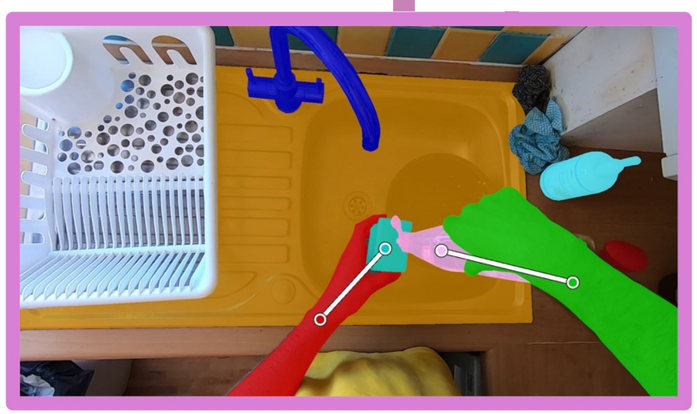
1 Hand, No Contact



2 Hands, No Contact

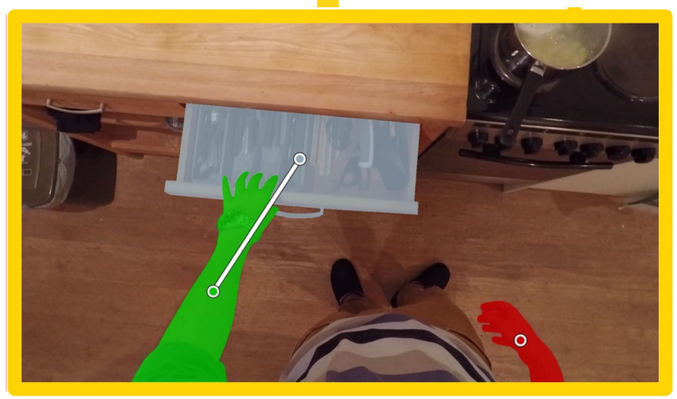
1 Hand, In Contact

2.7% 41.5% 0.7% 19.4% 27.2% 8.5%



2 Hands, 2 Obj Contacts

2 Hands, Same Contact



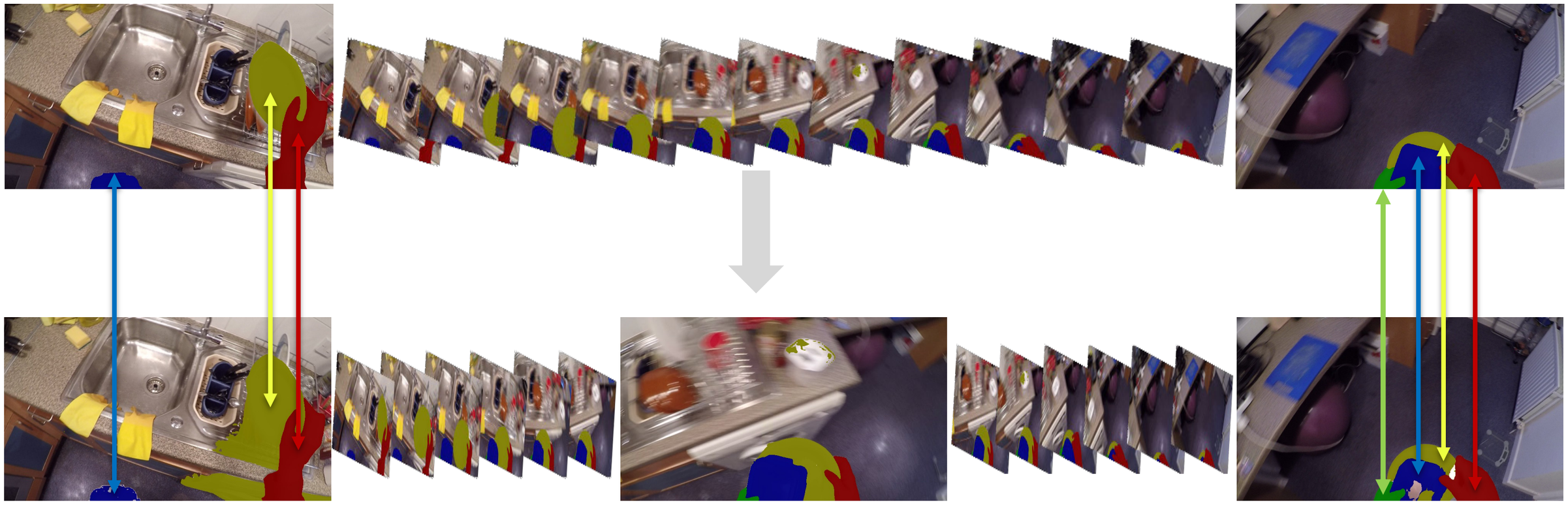
2 Hands, 1 In Contact

Dense Annotations

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler, Dima Damen

Dense Annotations

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma, Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler, Dima Damen



EPIC-KITCHENS VISOR

with: Ahmad Darkhalil, Dandan Shan, Bin Zhu, Jian Ma,
Amlan Kar, Richard Higgins, David Fouhey, Sanja Fidler

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations.
NeurIPS

Dima Damen
Rhobin W @CVPR2024

Ego-Exo4D

Fri Oral + Poster (Session 5 #292)

with: Kristen Grauman
+102 authors

Ego-Exo Relation

In this talk

HOI in 2D

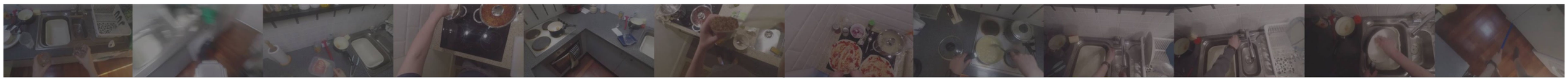
- VISOR (masks and hand-interactions)
- HOI-Ref
- GenHowTo

HOI 3D Reconstruction in view

- Get a Grip

HOI 3D Reconstruction in and out of view

- EPIC Fields - Scene reconstruction from egocentric views
- OSNOM - 3D tracking of HOI in world coordinate frames



HOI-Ref:
Hand-Object Interaction Referral in Egocentric Vision

Siddhant Bansal, Michael Wray, Dima Damen

University of BRISTOL

Dima Damen
Rhobin W @CVPR2024

HOI-Ref

with: Siddhant Bansal
Michael Wray

[refer] where is the left hand of the person?

[1]

[identify] cheese

[1]

[1] MiniGPT-v2: Large Language Model as a Unified Interface for Vision-Language Multi-task Learning by Jun Chen, Deyao Zhu , Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, Mohamed Elhoseiny.

HOI-Ref

with: Siddhant Bansal
Michael Wray

Available Annotations

Narration

#C drops the polaroid camera on the table

Generated QA Pairs

Q: [caption] What is happening in the photo?

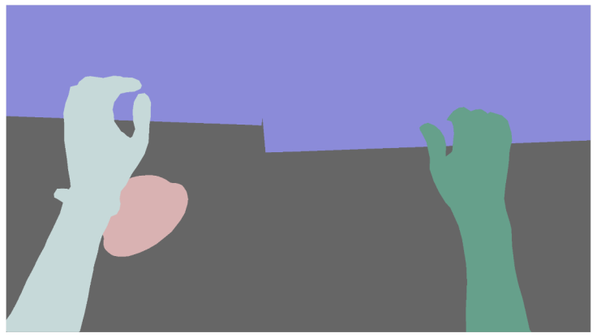
A: The person drops the polaroid camera on the table

Hand/Object Bounding Boxes

Q: [refer] Where are the hands of the person in the image?

A: The hands are here: {<33><58><43><76>} and {<54><30><65><52>}

Hand/Object Segments



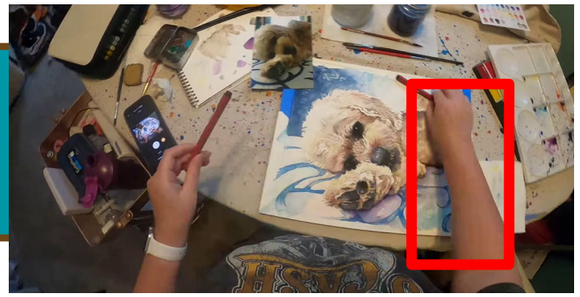
Q: [refer] Where is the cupboard?

A: {<0><0><100><36>}

HOI-Ref

with: Siddhant Bansal
Michael Wray

{<68><22><90><96>}

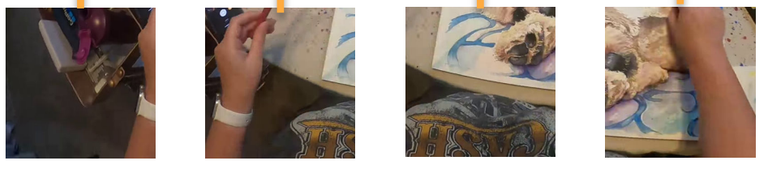


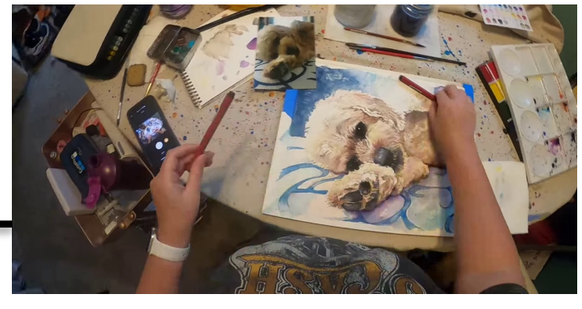
Large Language Model

Projection Layer

Vision Encoder

[refer] Where is the right hand of the person?





HOI-Ref

with: Siddhant Bansal
Michael Wray

[refer] where is the left hand of the person?

VLM4HOI

[identify] carrot

VLM4HOI

[identify] cheese

VLM4HOI

HOI-Ref

with: Siddhant Bansal
Michael Wray

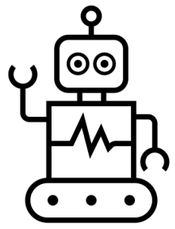
Egocentric Vision is not about static objects but about hands ***interacting*** with objects.

Can current VLMs understand this relationship?

S Bansal, M Wray, D Damen (2024). HOI-Ref: Hand-Object Interaction Referral in Egocentric Vision. ArXiv

HOI-Ref

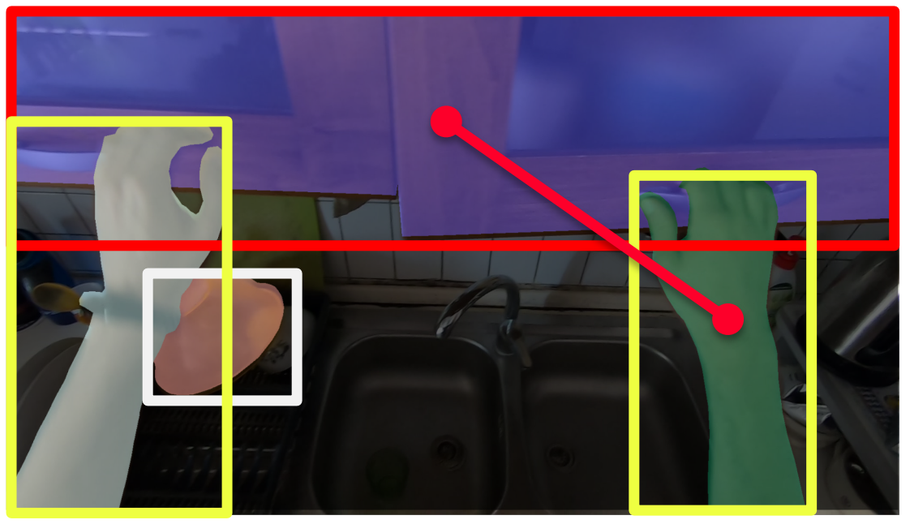
with: Siddhant Bansal
Michael Wray



[refer] Locate the object being manipulated by left hand

HOI-Ref

with: Siddhant Bansal
Michael Wray



Q: [refer] Where is the object manipulated by the right hand?

A: {<0><0><100><36>}

HOI-Ref

with: Siddhant Bansal
Michael Wray

HOI-QA: 3.9M Pairs

Shikra-RD[2]: 5922 Pairs

Flicker30k [1]: 2.5K Pairs

Multi-task Conversation [1] : 60.9K Pairs

[1] MiniGPT-v2: Large Language Model as a Unified Interface for Vision-Language Multi-task Learning by Jun Chen, Deyao Zhu , Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, Mohamed Elhoseiny.
[2] Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic by Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, Rui Zhao

HOI-Ref

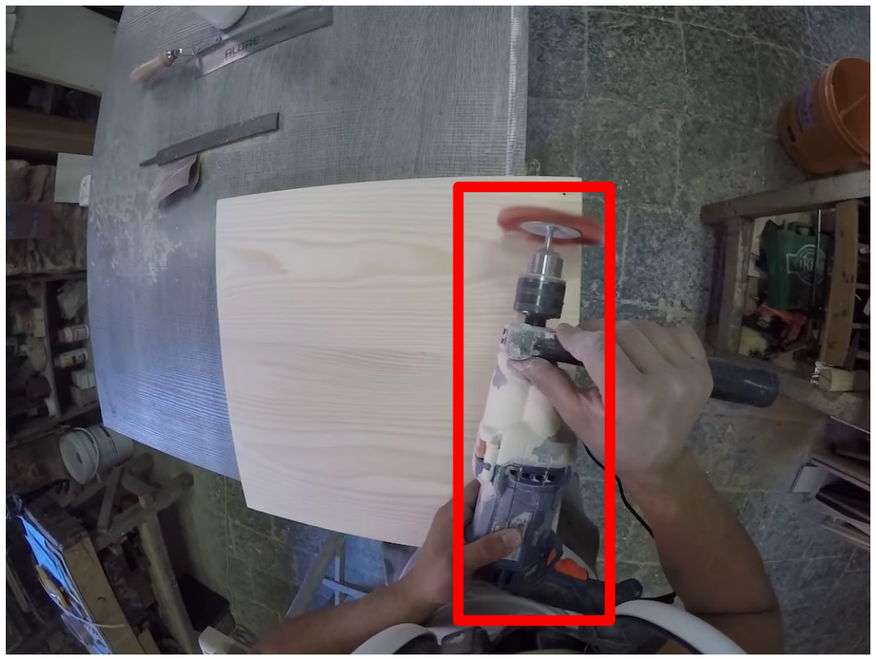
with: Siddhant Bansal
Michael Wray

VLM4HOI

[refer] Locate the object being manipulated by left hand

HOI-Ref

with: Siddhant Bansal
Michael Wray



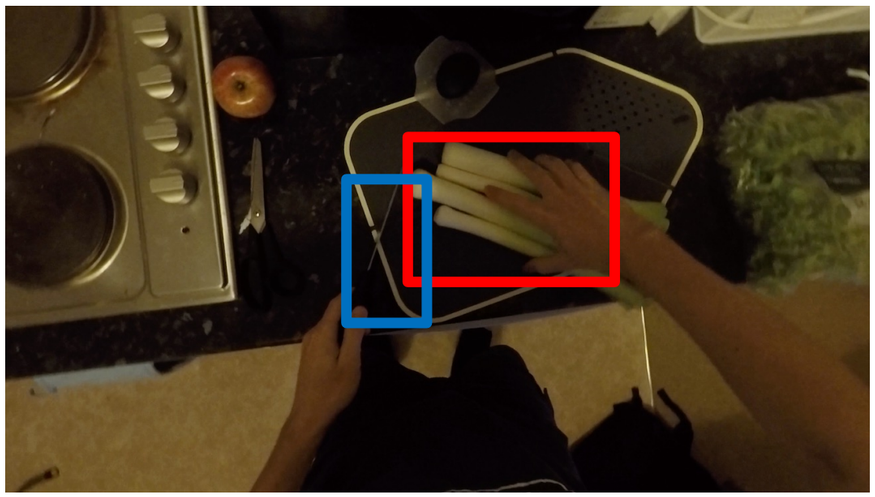
Q: [refer] where is the manipulated object?
A: (53, 30, 68, 92)

Q: [identify] what is this (6, 0, 48, 95)
A: That is a wagon in the image.

HOI-Ref

with: Siddhant Bansal
Michael Wray

Q: [detection] shoes
A: shoes (28, 22, 32, 26)



Q: [grounding] describe the actions done by hand
A: The person slices the celery with the knife

In this talk

HOI in 2D

- VISOR (masks and hand-interactions)
- HOI-Ref
- GenHowTo

HOI 3D Reconstruction in view

- Get a Grip

HOI 3D Reconstruction in and out of view

- EPIC Fields - Scene reconstruction from egocentric views
- OSNOM - 3D tracking of HOI in world coordinate frames

Wed (Session 2)
Poster # 172

CVPR
JUNE 17-21, 2024
SEATTLE, WA

GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos

Tomáš Souček

Dima Damen

Michael Wray

Ivan Laptev

Josef Šivic

University of BRISTOL

CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

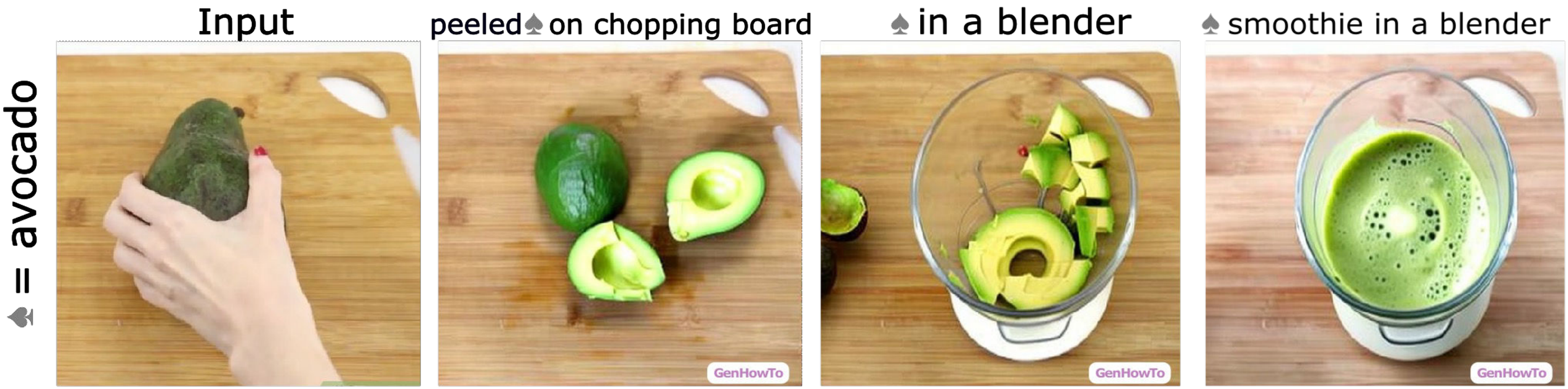
University of BRISTOL

Dima Damen
Rhobin W @CVPR2024

GenHowTo…

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

- Hands transform objects….



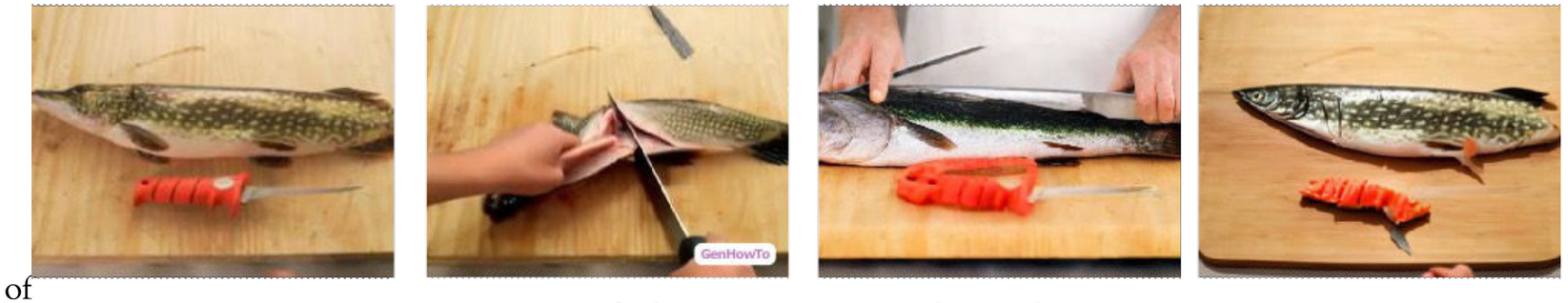
T Soucek et al (2023). GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. ArXiv

GenHowTo…

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

Prompt: a frosted cake with strawberries around the top

Prompt: a person kneading dough on a cutting board



Prompt: a person cutting a fish on a cutting board

University of BRISTOL

Dima Damen
Rhobin W @CVPR2024

GenHowTo…

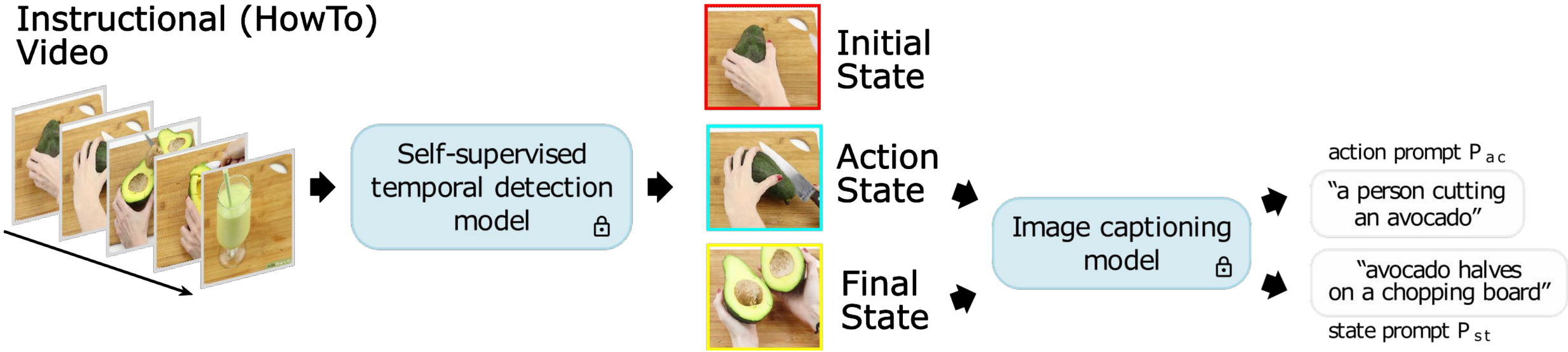
with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

- Two contributions…. Dataset & Method

GenHowTo…

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

- Two contributions…. Dataset & Method

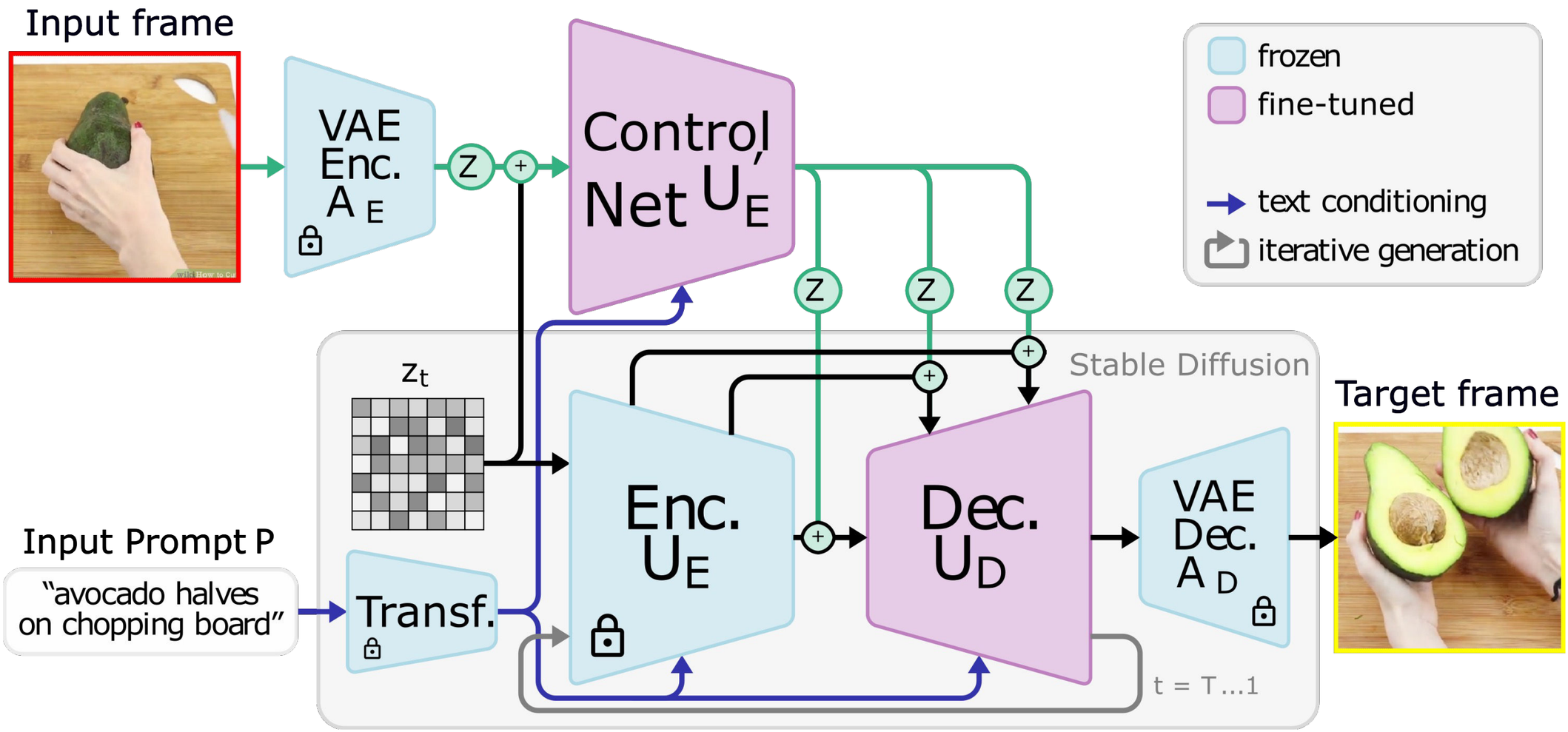


Tomas Soucek, Jean-Baptiste Alayrac, Antoine Miech, Ivan Laptev, and Josef Sivic (2022). Multi-task learning of object state changes from uncurated videos.

GenHowTo…

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

- Two contributions…. Dataset & Method



T Soucek et al (2023). GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. ArXiv

GenHowTo…

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

Input *less noise* ⟷ *more noise*

T Soucek et al (2023). GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. ArXiv

Dima Damen
Rhobin W @CVPR2024

GenHowTo…

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

- Qualitative Evaluation…
  - Initial vs Final State
  - Binary Classifier

	Method	Acc_{ac} ↑	Acc_{st} ↑
	test set categories unseen during training		
(a)	Stable Diffusion	0.51	0.50
(b)	Edit Friendly DDPM	0.60	0.61
(c)	InstructPix2Pix	0.55	0.63
(d)	*CLIP (manual prompts)*	0.52	0.62
(e)	**GenHowTo**	**0.66**	**0.74**
	test set categories seen during training		
(f)	Edit Friendly DDPM†	0.69	0.80
(g)	**GenHowTo†**	**0.77**	**0.88**
(h)	*Real images*	0.96	0.97

† Models trained also on the test set *categories*.

T Soucek et al (2023). GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. ArXiv

GenHowTo…

with: Tomas Soucek
Ivan Laptev
Michael Wray
Josef Sivic

a person is wrapping a tortilla on a plate

REAL IMAGE —— GENERATED

a plate with two burritos on it

REAL IMAGE —— GENERATED

a man pouring beer into a glass

REAL IMAGE —— GENERATED

a man sitting at a table holding a glass of beer

REAL IMAGE —— GENERATED

University of BRISTOL

T Soucek et al (2023). GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos. ArXiv

Dima Damen
Rhobin W @CVPR2024

Wed (Session 2)
Poster # 172

CVPR
JUNE 17-21, 2024
SEATTLE, WA

GenHowTo: Learning to Generate Actions and State Transformations from Instructional Videos

Tomáš Souček

Dima Damen

Michael Wray

Ivan Laptev

Josef Šivic

University of BRISTOL

CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

University of BRISTOL

Dima Damen
Rhobin W @CVPR2024

In this talk

HOI in 2D

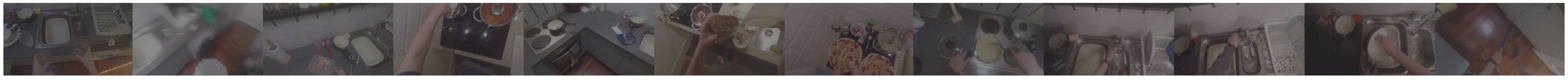
- VISOR (masks and hand-interactions)
- HOI-Ref
- GenHowTo

HOI 3D Reconstruction in view

- Get a Grip

HOI 3D Reconstruction in and out of view

- EPIC Fields - Scene reconstruction from egocentric views
- OSNOM - 3D tracking of HOI in world coordinate frames



Get a Grip

Reconstructing Hand-Object Stable Grasps in Egocentric Videos

Zhifan Zhu and Dima Damen

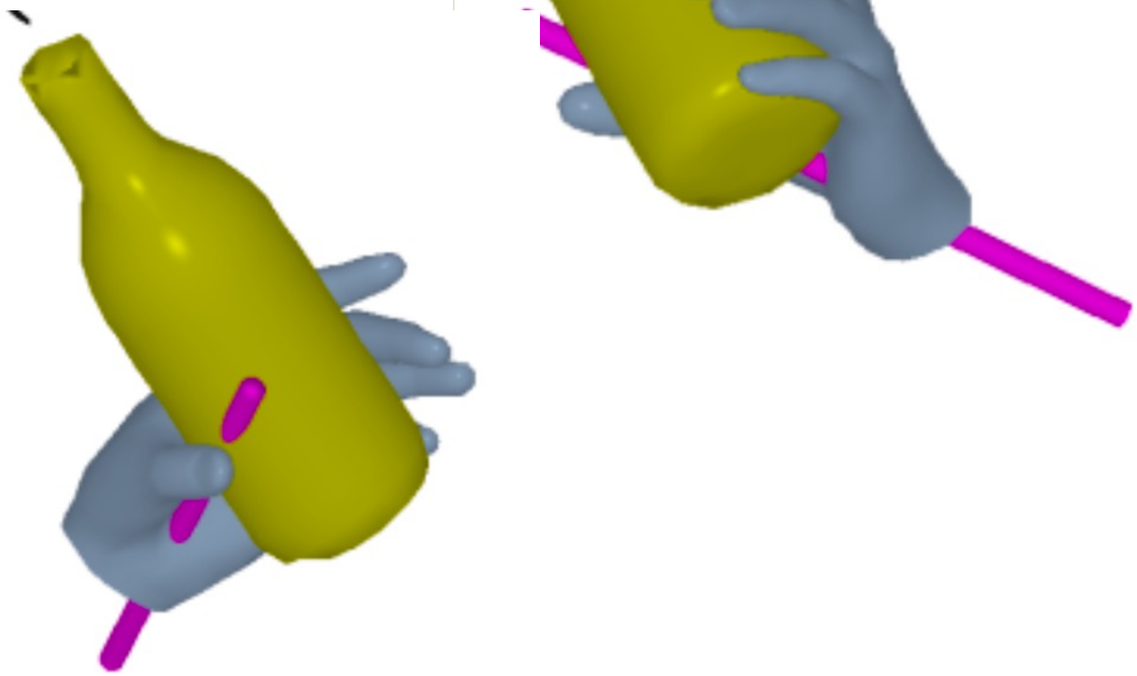
University of BRISTOL

Get a Grip

with: Zhifan Zhu

left hand

bottle



University of BRISTOL

Z Zhu and D Damen (2024). Get a Grip: Reconstructing Hand-Object Stable Grasps in Egocentric Videos. *ArXiv*

Dima Damen
Rhobin W @CVPR2024

Get a Grip

with: Zhifan Zhu

Non-Ego Views

Ego Views

University of BRISTOL

Z Zhu and D Damen (2024). Get a Grip: Reconstructing Hand-Object Stable Grasps in Egocentric Videos. *ArXiv*

Dima Damen
Rhobin W @CVPR2024

Get a Grip

with: Zhifan Zhu

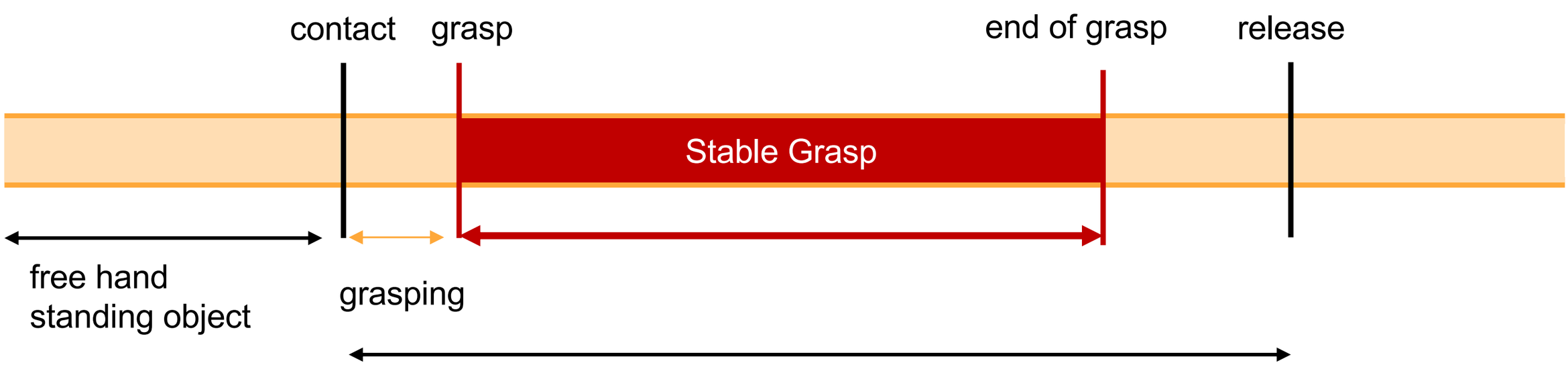
University of BRISTOL

Z Zhu and D Damen (2024). Get a Grip: Reconstructing Hand-Object Stable Grasps in Egocentric Videos. *ArXiv*

Dima Damen
Rhobin W @CVPR2024

Get a Grip

with: Zhifan Zhu

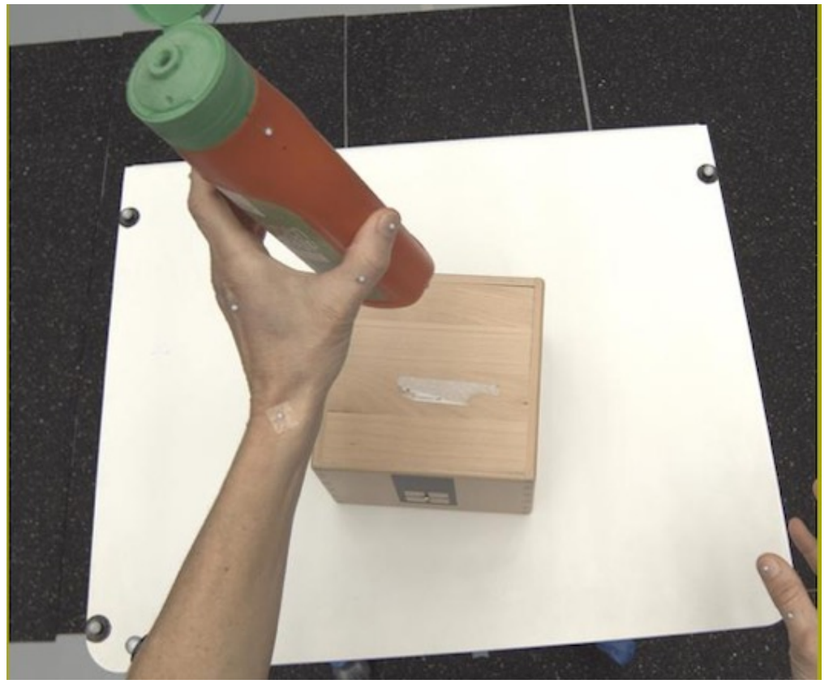


Z Zhu and D Damen (2024). Get a Grip: Reconstructing Hand-Object Stable Grasps in Egocentric Videos. *ArXiv*

Get a Grip

with: Zhifan Zhu

ARCTIC (CVPR 2023)



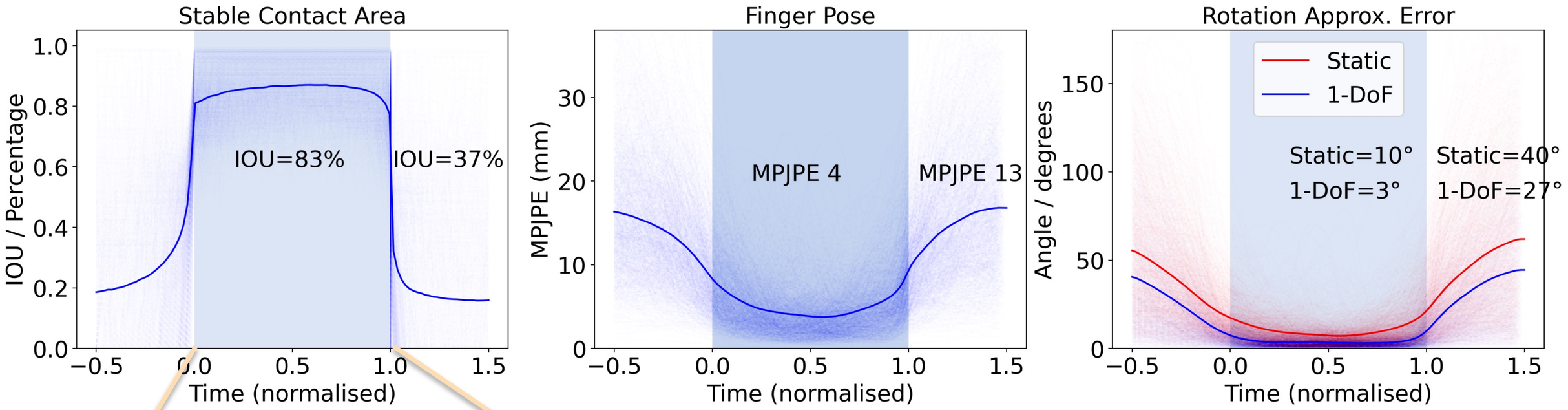
Z Fan, et al.
ARCTIC: A dataset for dexterous bimanual hand- object manipulation. **CVPR 2023**

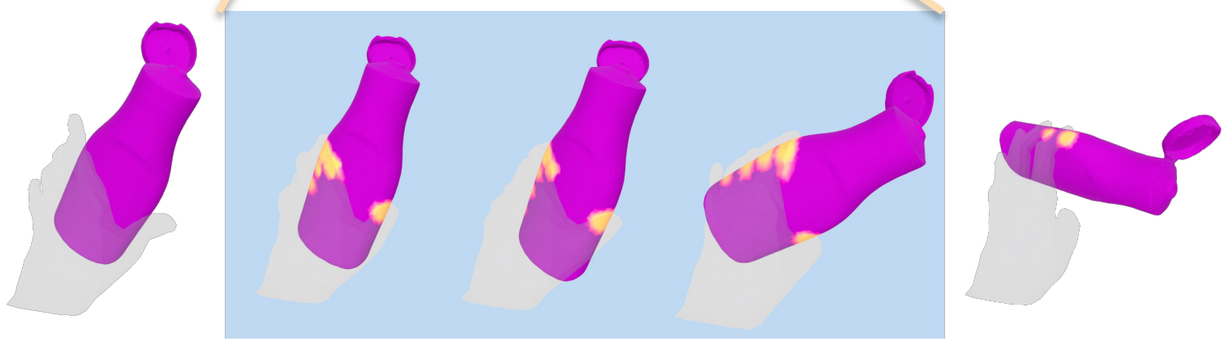
HOI4D (CVPR 2022)

Yunze Liu, et al.
HOI4D: A 4D Egocentric Dataset for Category-Level Human-Object Interaction.
CVPR 2022

Get a Grip

with: Zhifan Zhu





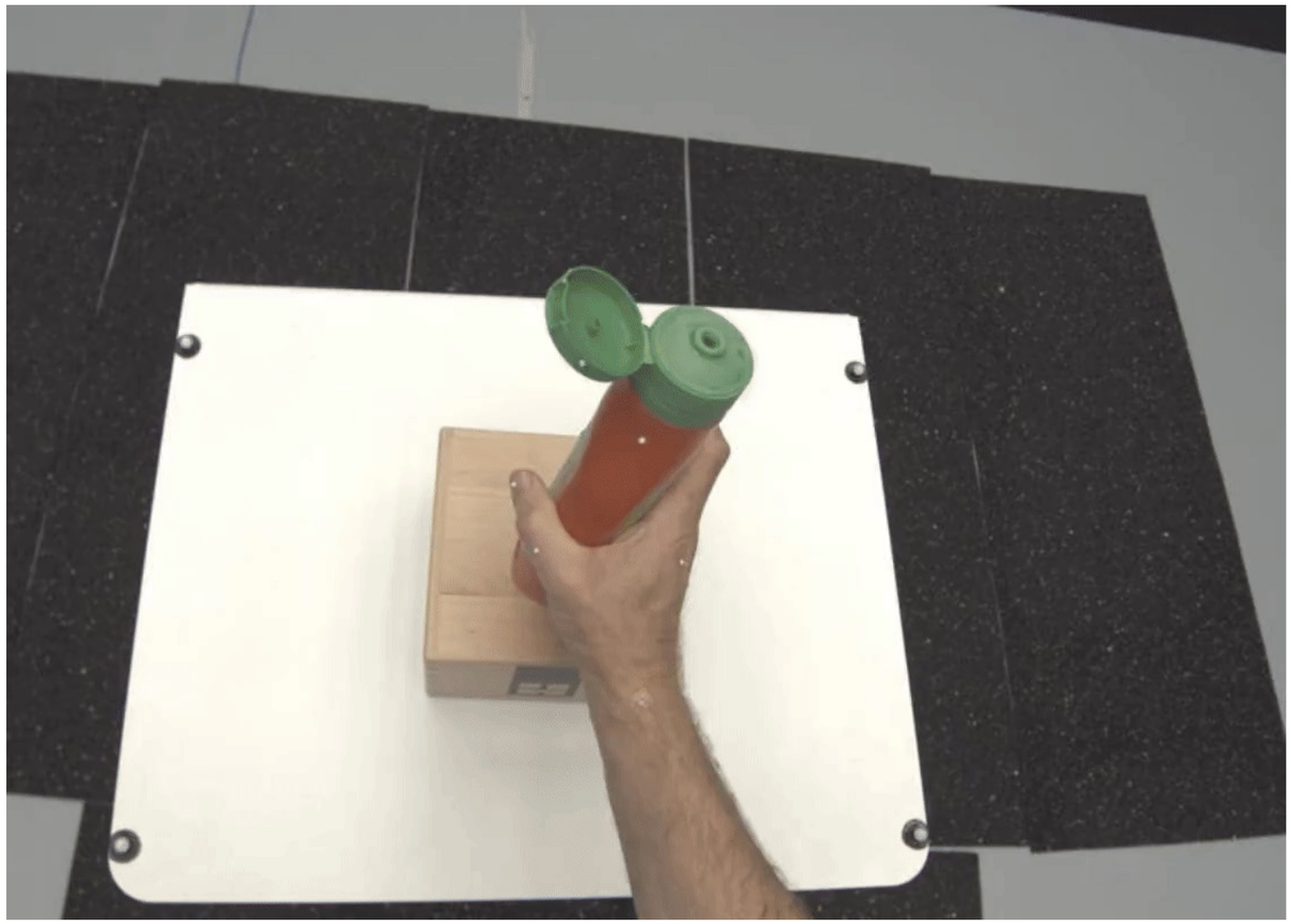
University of BRISTOL

Z Zhu and D Damen (2024). Get a Grip: Reconstructing Hand-Object Stable Grasps in Egocentric Videos. *ArXiv*

Dima Damen
Rhobin W @CVPR2024

Get a Grip

with: Zhifan Zhu



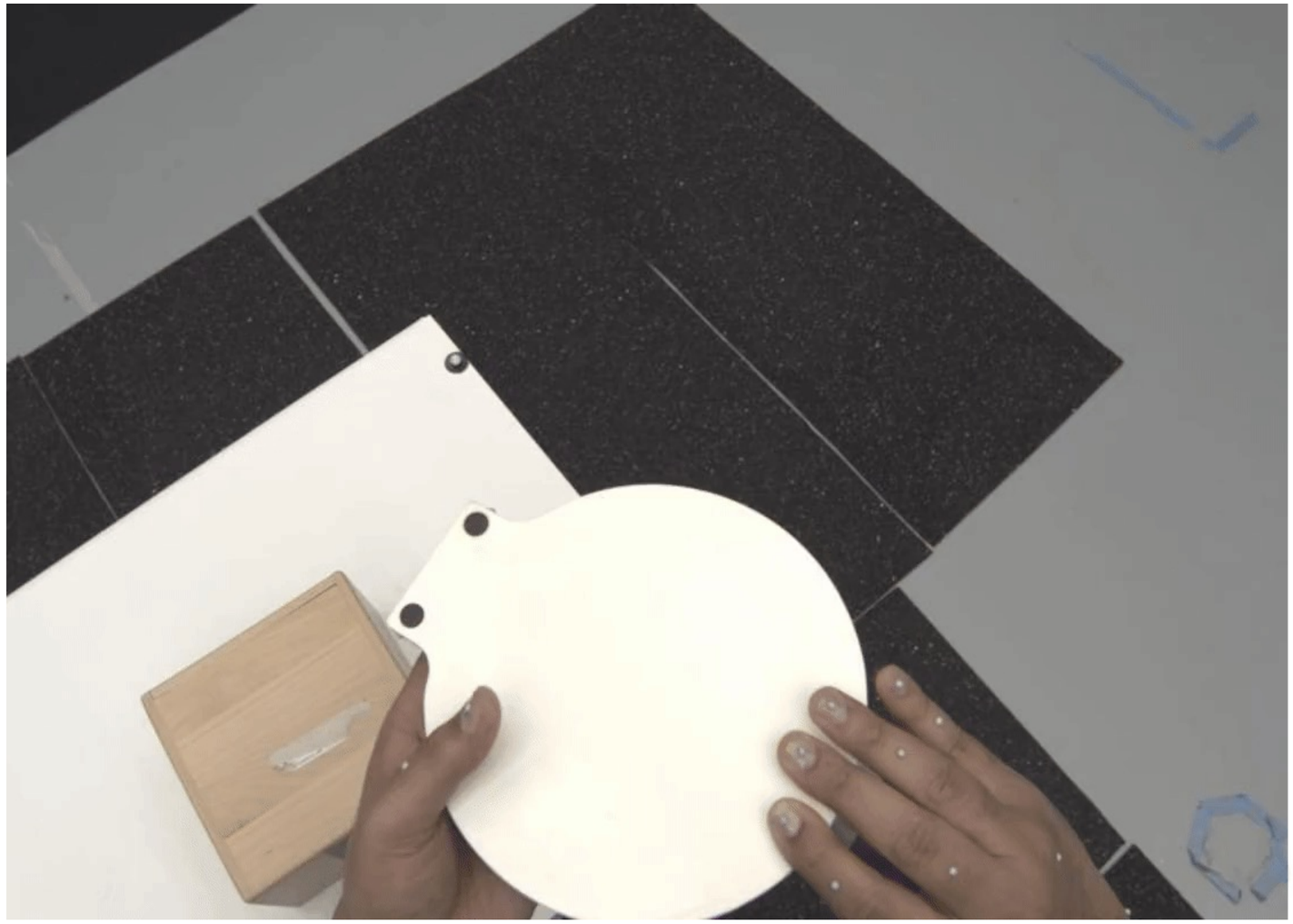
University of BRISTOL

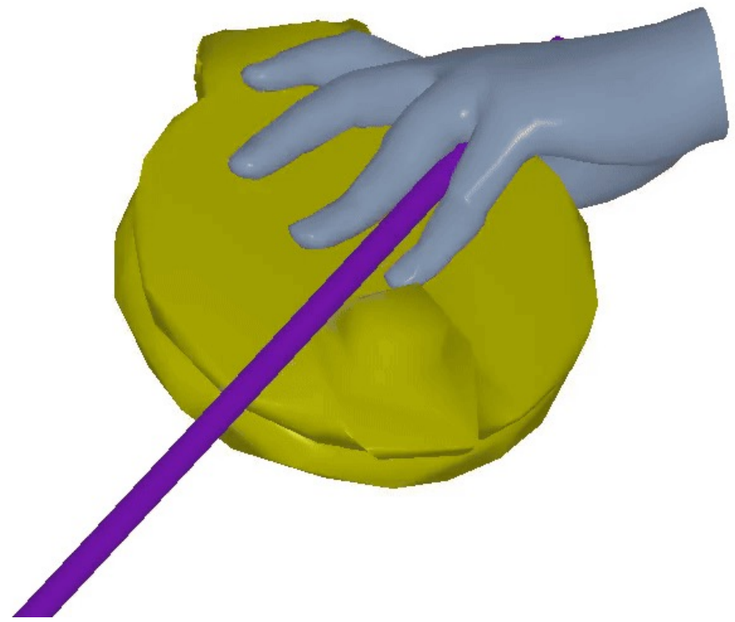
Z Zhu and D Damen (2024). Get a Grip: Reconstructing Hand-Object Stable Grasps in Egocentric Videos. *ArXiv*

Dima Damen
Rhobin W @CVPR2024

Get a Grip

with: Zhifan Zhu





University of BRISTOL

Z Zhu and D Damen (2024). Get a Grip: Reconstructing Hand-Object Stable Grasps in Egocentric Videos. *ArXiv*

Dima Damen
Rhobin W @CVPR2024

Get a Grip

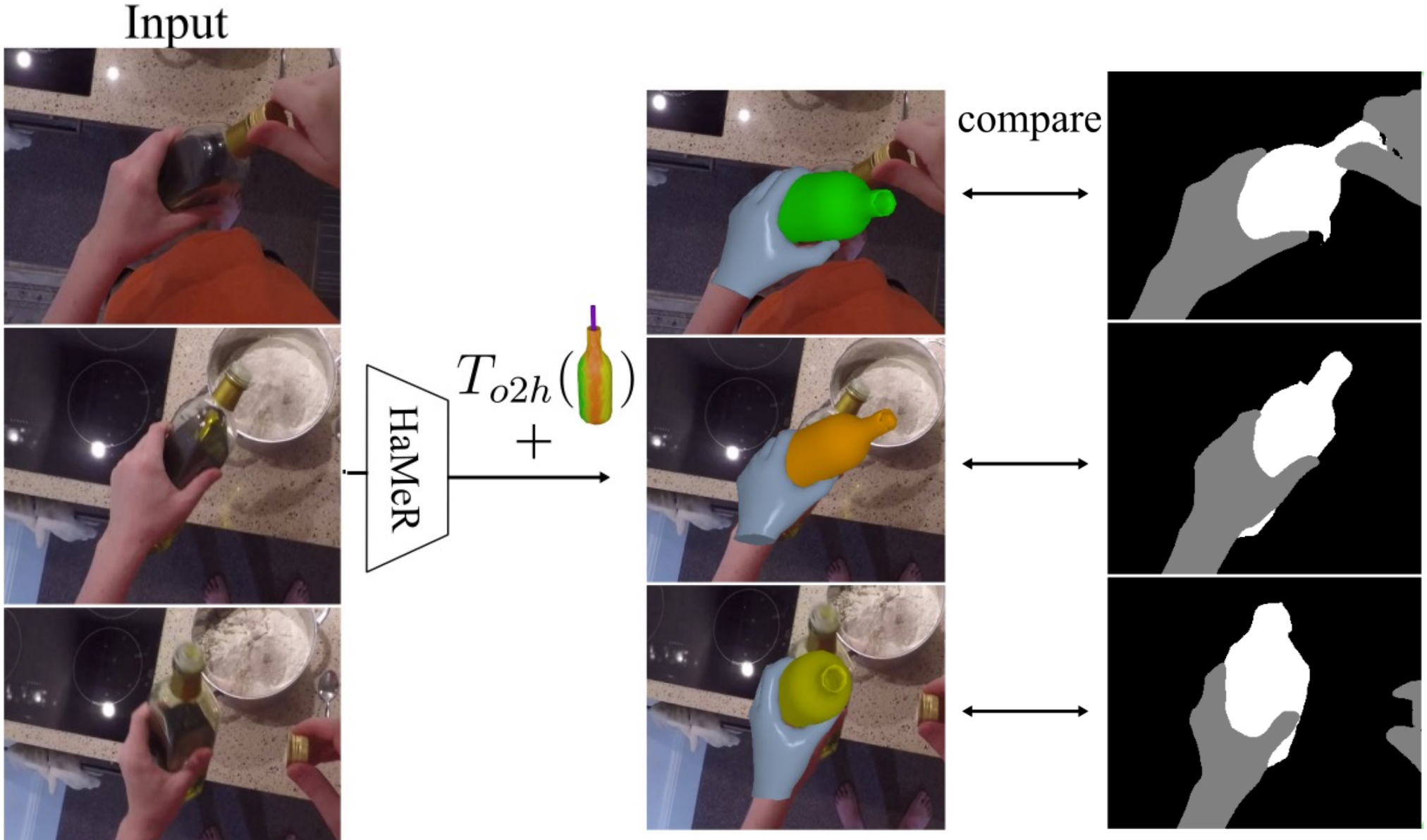
with: Zhifan Zhu

Sequences	Instances	Categories	Subjects
2431	~390	9	31

1446 left hands  , 985 right hands 

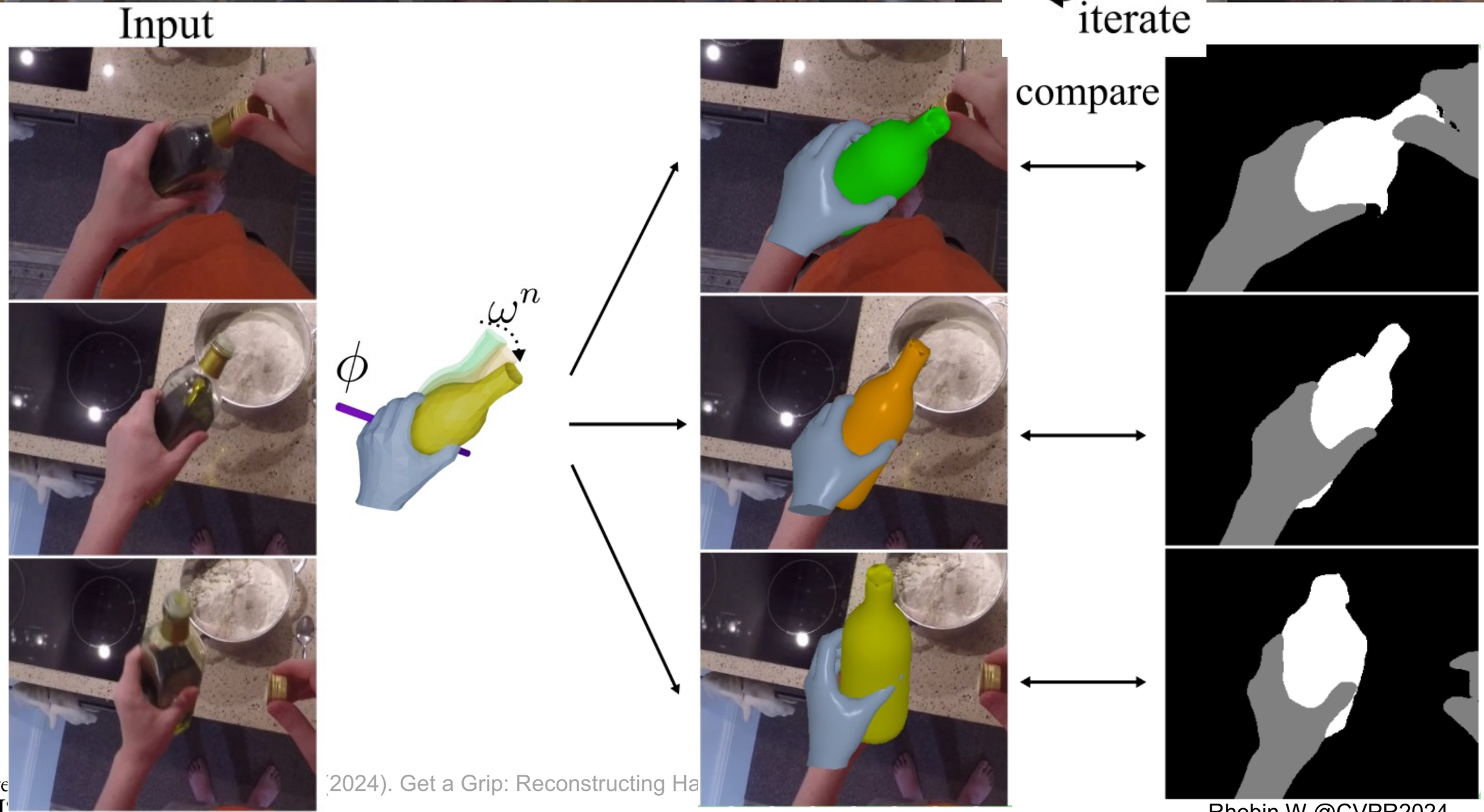
Get a Grip

with: Zhifan Zhu



Get a Grip

with: Zhifan Zhu



Get a Grip

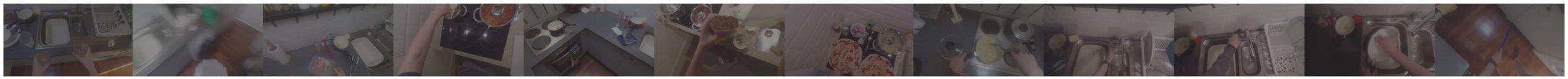
with: Zhifan Zhu

Bottle Samples

University of BRISTOL

Z Zhu and D Damen (2024). Get a Grip: Reconstructing Hand-Object Stable Grasps in Egocentric Videos. *ArXiv*

Dima Damen
Rhobin W @CVPR2024



Get a Grip

Reconstructing Hand-Object Stable Grasps in Egocentric Videos

Zhifan Zhu and Dima D

Labelled stable grasps, code and models are public

University of BRISTOL

Dima Damen
Rhobin W @CVPR2024

Ego-Exo4D

Fri Oral + Poster (Session 5 #292)

with: Kristen Grauman
+102 authors

University of BRISTOL

K Grauman et al (2023). Ego-Exo4D: Understanding Skilled Human Activity from First- and Third-Person Perspectives.. *ArXiv*

Dima Damen
Rhobin W @CVPR2024

In this talk

HOI in 2D

- VISOR (masks and hand-interactions)
- HOI-Ref

HOI 3D Reconstruction in view

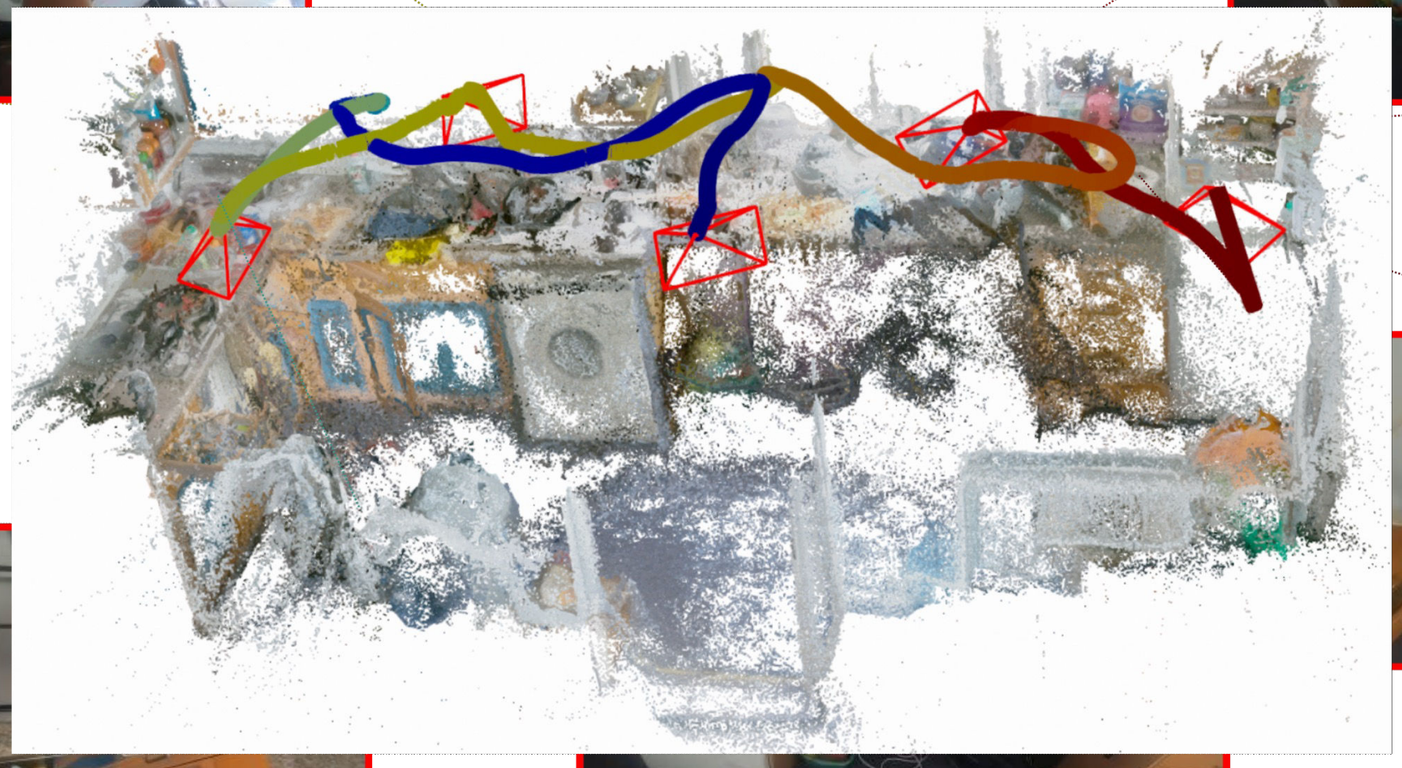
- Get a Grip

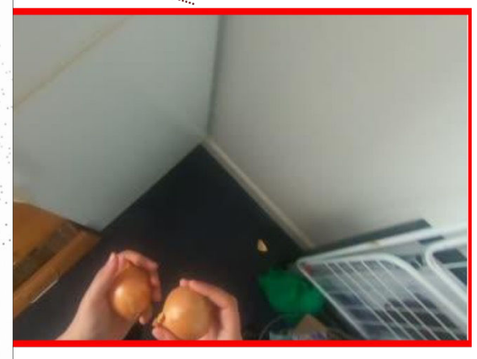
HOI 3D Reconstruction in and out of view

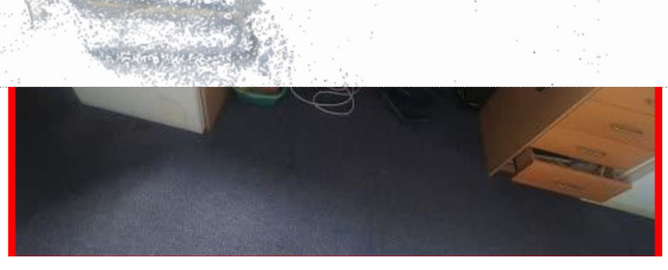
- EPIC Fields - Scene reconstruction from egocentric views
- OSNOM - 3D tracking of HOI in world coordinate frames

EPIC Fields

with: V Tschernezki*, A Darkhalil*, Z Zhu*,
D Fouhey, I Laina, D Larlus, A Vedaldi







University of BRISTOL

Dima Damen
Rhobin W @CVPR2024

EPIC-KITCHENS

EPIC Fields

with: V Tschernezki*, A Darkhalil*, Z Zhu*,
D Fouhey, I Laina, D Larlus, A Vedaldi

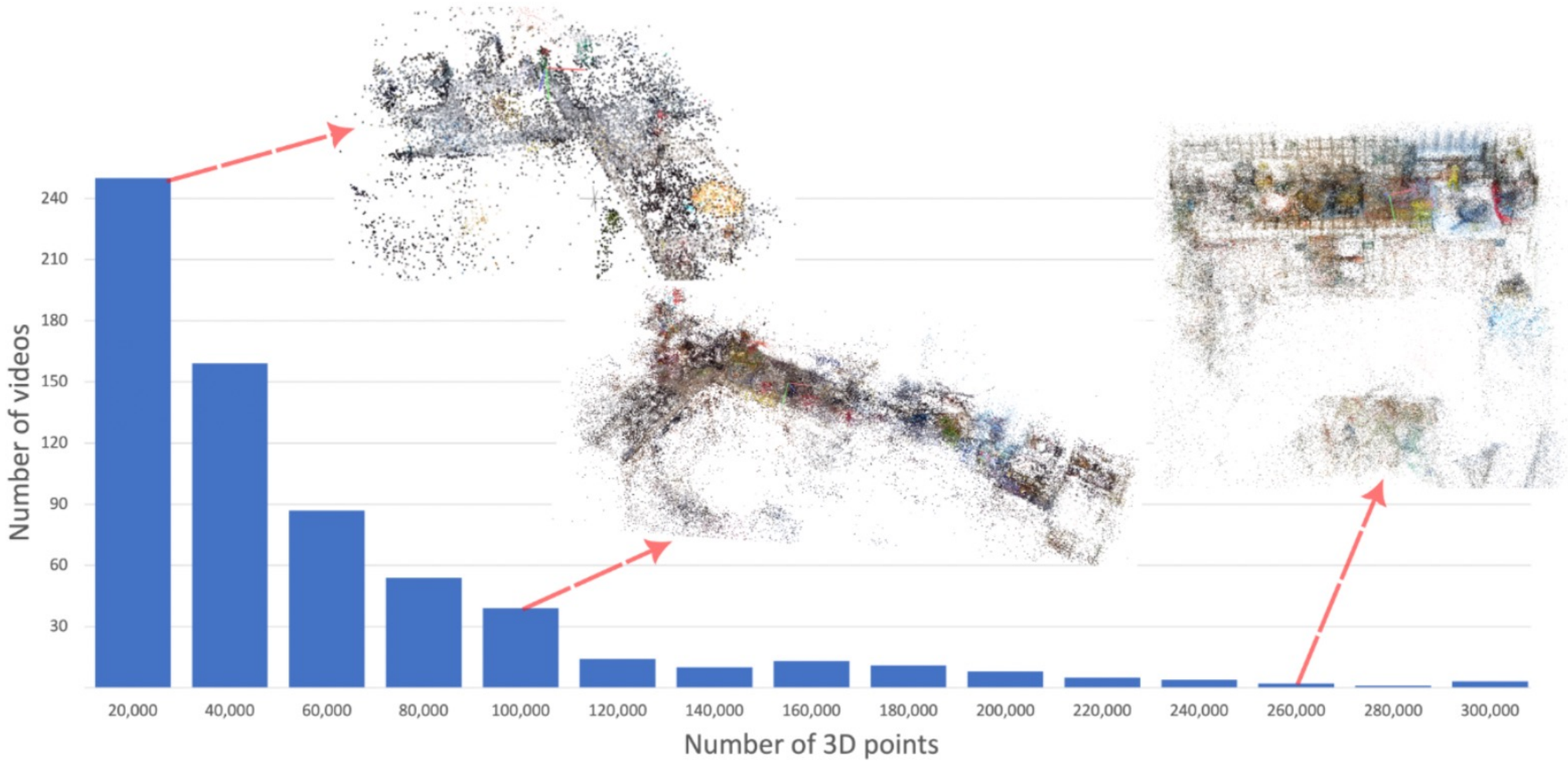


Figure 4: **Number of 3D points histogram.** The majority of our reconstructions generate less than 40,000 points that are enough to represent the kitchen. However, some reconstructions have more than 100,000, we include the point clouds for each points range showing the fine details covered by having more points

EPIC Fields

with: V Tschernezki*, A Darkhalil*, Z Zhu*,
D Fouhey, I Laina, D Larlus, A Vedaldi

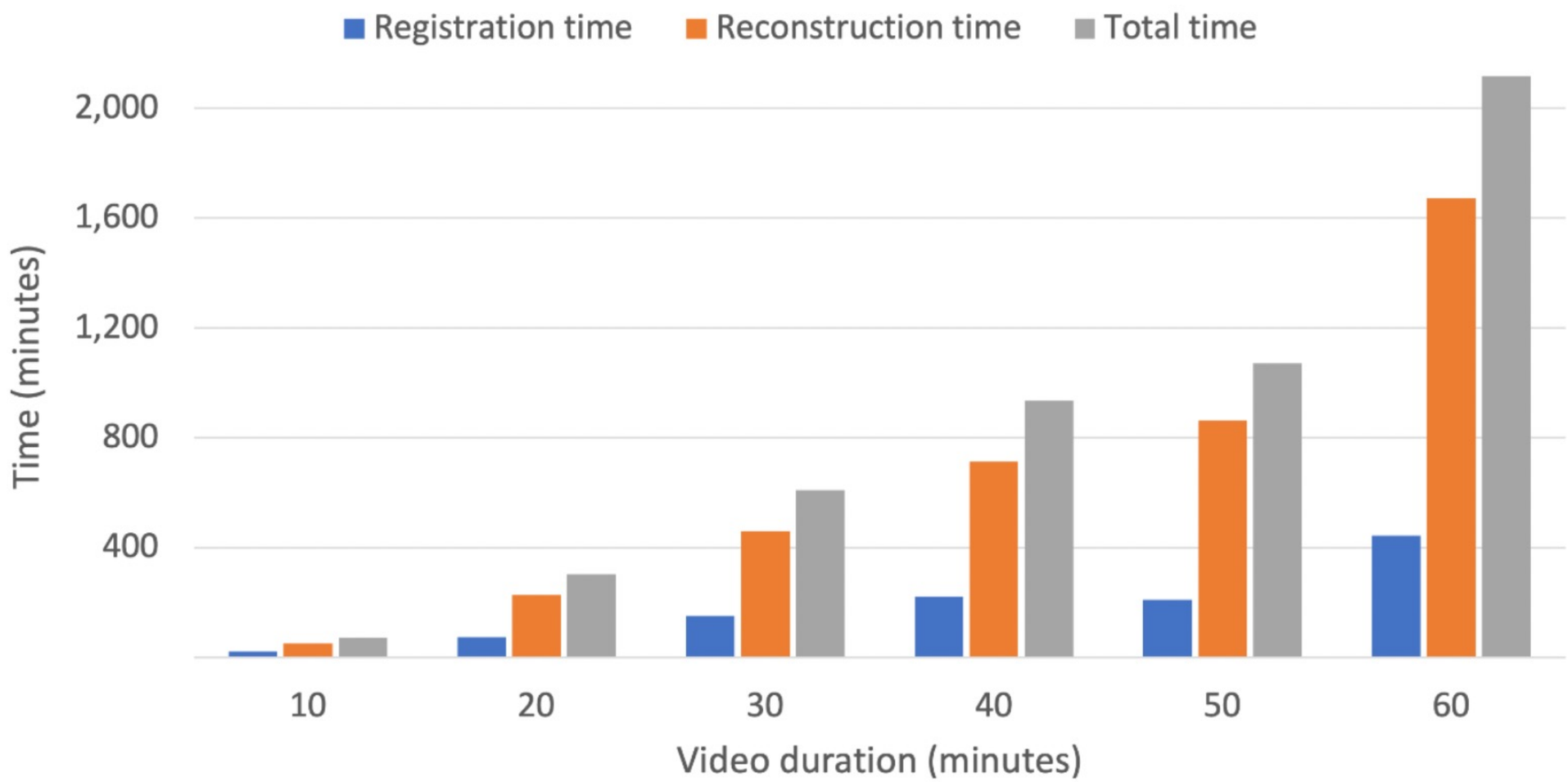


Figure 5: **Reconstruction time per video length.** We plot both times for the sparse reconstruction and the registration time to obtain the dense camera poses. While the time for registration is almost linear, the reconstruction time increase by the video length non-linearly mainly because of the bundle adjustment operation

EPIC Fields

with: V Tschernezki*, A Darkhalil*, Z Zhu*,
D Fouhey, I Laina, D Larlus, A Vedaldi

Table 1: Comparison of datasets commonly used in dynamic new-view synthesis.

Dataset	#Scenes	Seq. Length	Monocular	Semantics
Nerfies [37]	4	8–15 sec	-	-
D-NeRF [41]	8	1–3 sec	-	-
Plenoptic Video [22]	6	10-60 sec	-	-
NVIDIA Dynamic Scene Dataset [65]	12	1–5 sec	4 / 12	-
HyperNeRF [38]	16	8–15 sec	13 / 16	-
iPhone [13]	14	8–15 sec	7 / 14	-
SAFF [25]	8	1–5sec	-	✓
EPIC Fields (ours)	50	6–37 min (Avg 22)	50 / 50	✓

A Darkhalil et al (2022). EPIC-KITCHENS VISOR Benchmark: VIdeo Segmentations and Object Relations. *NeurIPS*

In this talk

HOI in 2D

- VISOR (masks and hand-interactions)
- HOI-Ref
- GenHowTo

HOI 3D Reconstruction in view

- Get a Grip

HOI 3D Reconstruction in and out of view

- EPIC Fields - Scene reconstruction from egocentric views
- OSNOM - 3D tracking of HOI in world coordinate frames

Spatial Cognition from Egocentric Video:

Out of Sight, Not Out of Mind

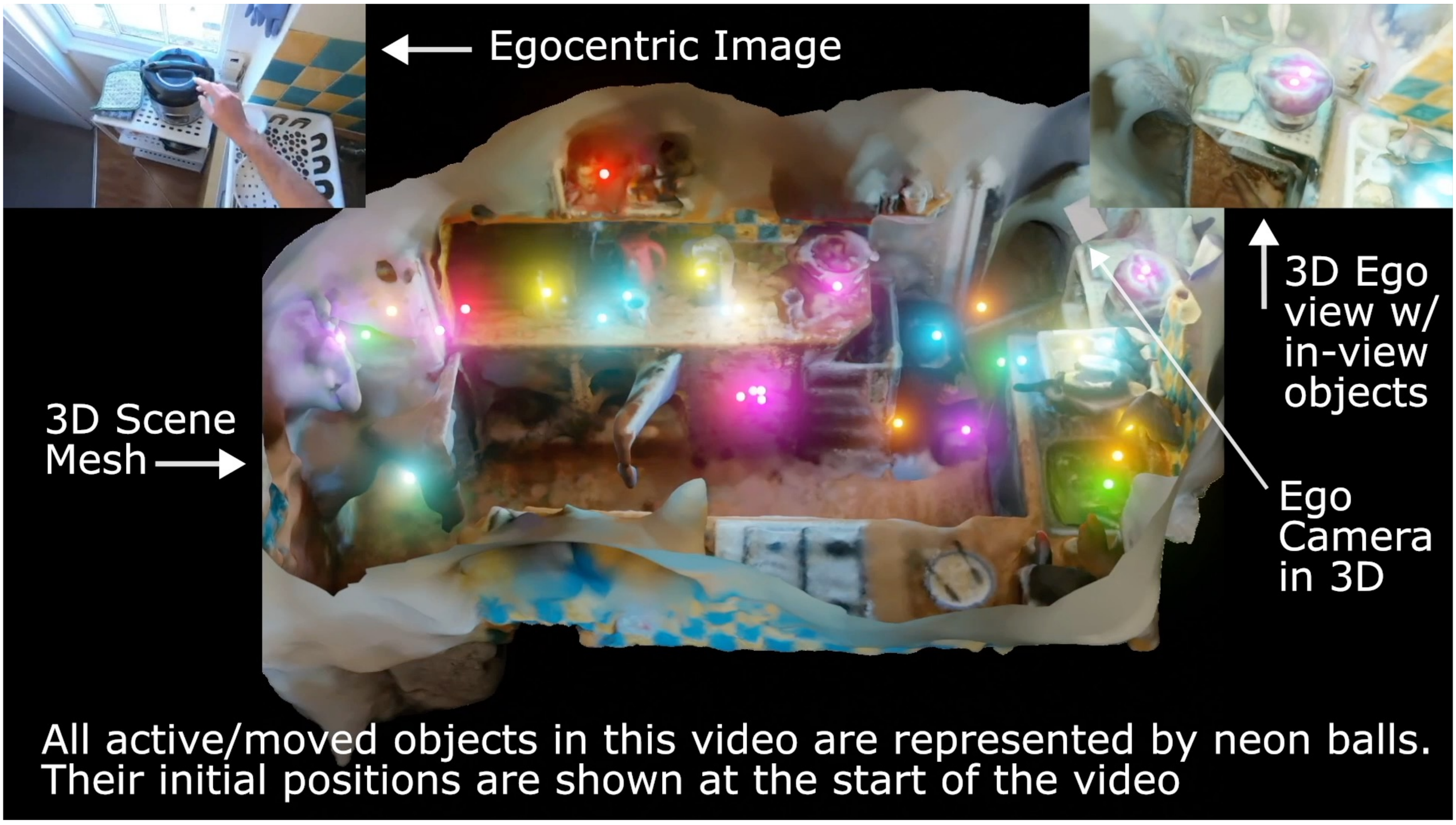
Chiara Plizzari Shubham Goel Toby Perrett Jacob Chalk

Angjoo Kanazawa Dima Damen

http://dimadamen.github.io/OSNOM

Out of Sight, not Out of Mind

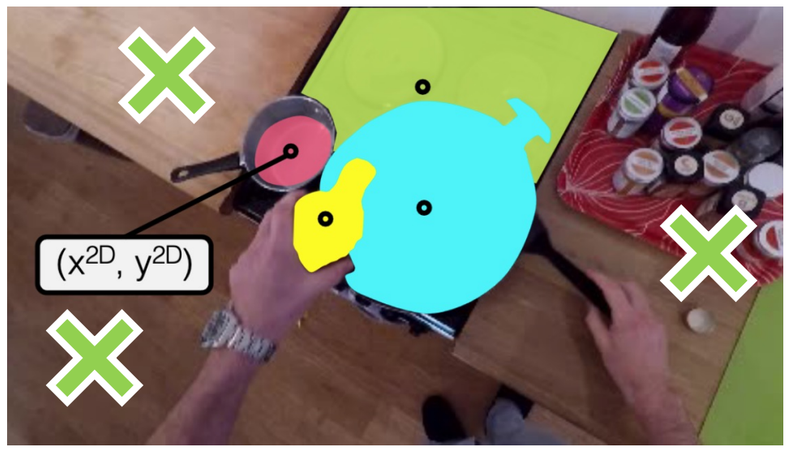
with: Chiara Plizzari, Shubham Goel, Toby Perrett, Angjoo Kanazawa

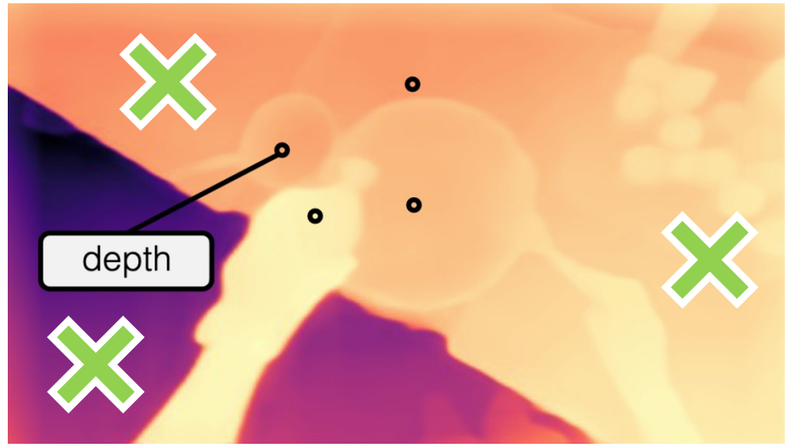

Egocentric Image
3D Scene Mesh
3D Ego view w/ in-view objects
Ego Camera in 3D
All active/moved objects in this video are represented by neon balls.
Their initial positions are shown at the start of the video

Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

Lift → Match → Keep





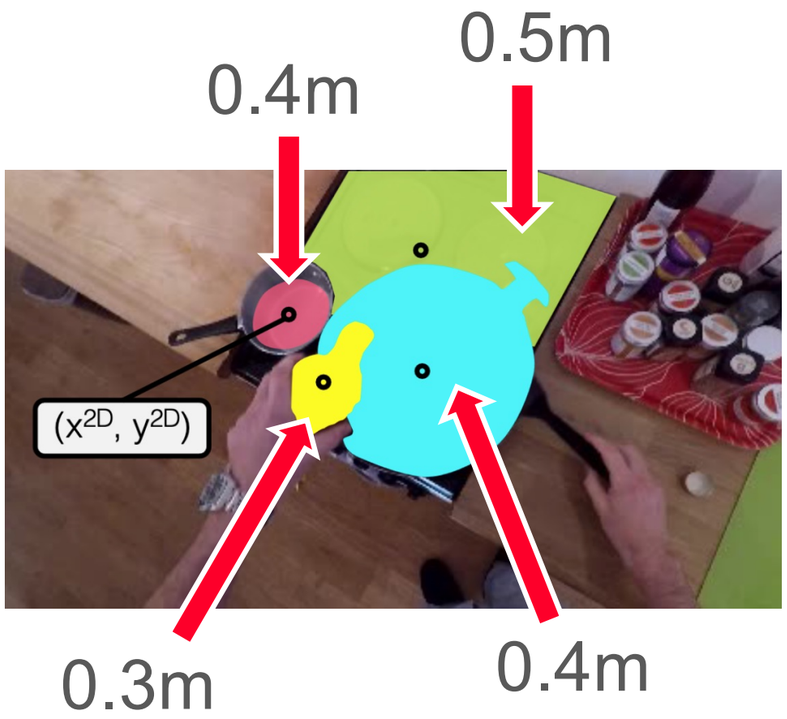
0.0 … 1.0

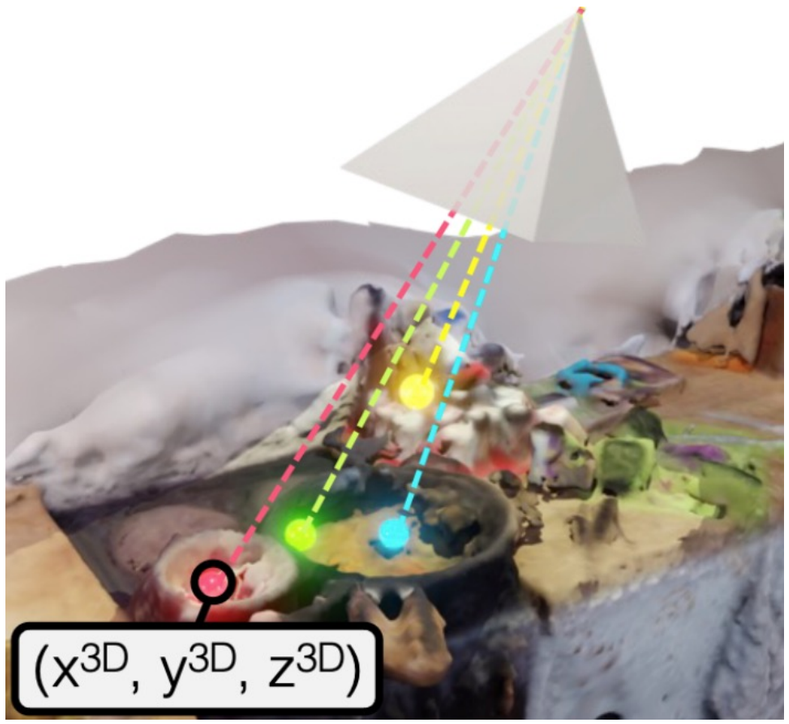
0.3m … 1.8m

Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

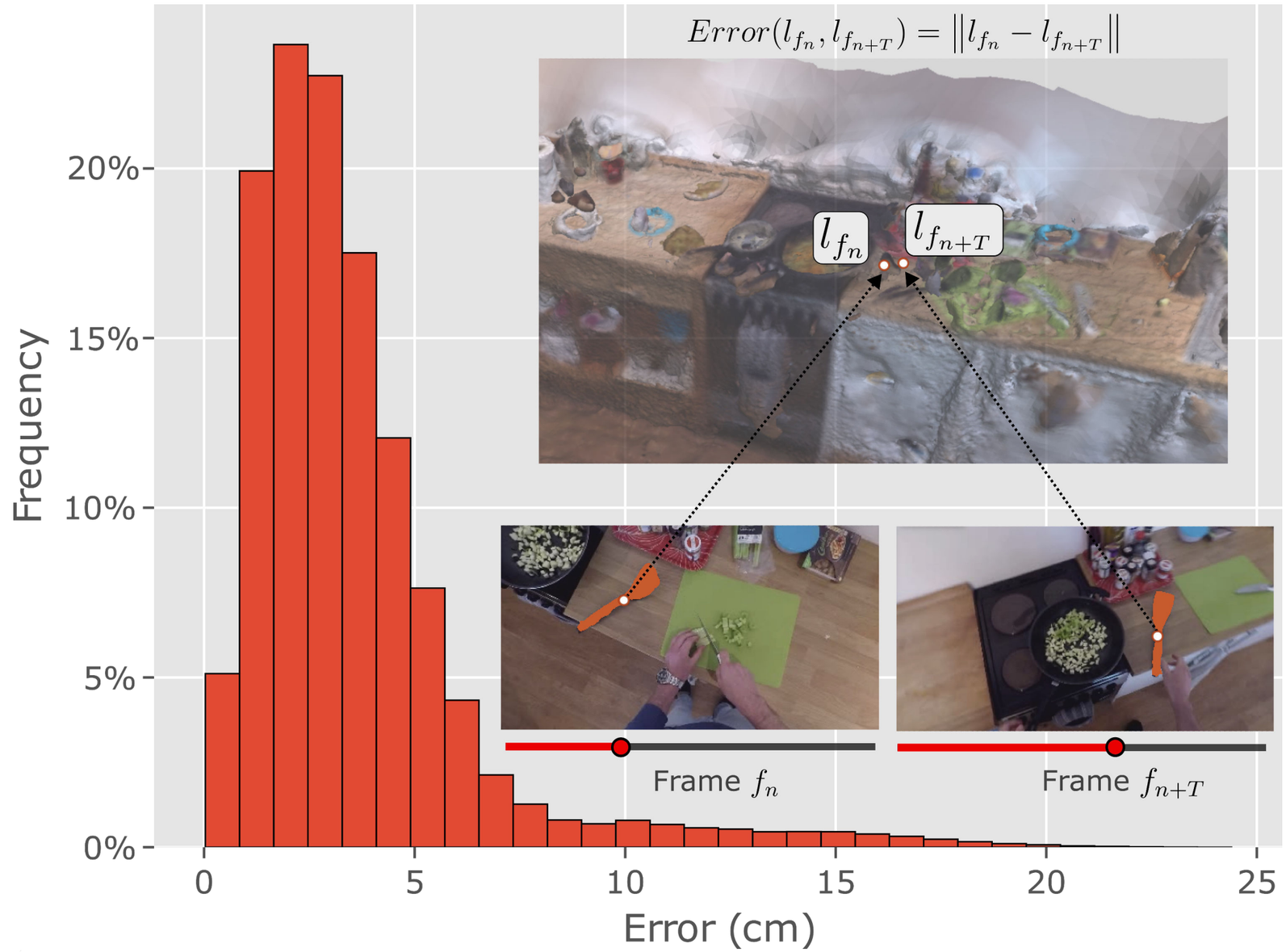
Lift → Match → Keep





Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa



Plizzari et al (2024). Spatial Cognition from Egocentric Video: Out of Sight, Not Out of Mind. ArXiv

Dima Damen
Rhobin W @CVPR2024

Out of Sight, not Out of Mind

with: Chiara Plizzari
Toby Perrett
Shubham Goel
Angjoo Kanazawa

Lift → Match → Keep

Instead of tracking in 2D, we track in 3D, using combination of appearance and location distances

Out of Sight, not Out of Mind

with: Chiara Plizzari Shubham Goel
Toby Perrett Angjoo Kanazawa

After we Lift, Match and Keep (LMK), we can reason about an object's visibility and position

- In-View vs Out-of-View
- In-Sight vs Out-of-Sight (Occluded)
- Within-Reach vs Out-of-Reach (defining the camera wearer's near space)

Out of Sight, not Out of Mind

with: Chiara Plizzari Shubham Goel
Toby Perrett Angjoo Kanazawa

After we Lift, Match and Keep (LMK), we can reason about an object's visibility and position

- In-View vs Out-of-View
- In-Sight vs Out-of-Sight (Occluded)
- Within-Reach vs Out-of-Reach (defining the camera wearer's near space)

Out of Sight, not Out of Mind

with: Chiara Plizzari Shubham Goel
Toby Perrett Angjoo Kanazawa

After we Lift, Match and Keep (LMK), we can reason about an object's visibility and position

- In-View vs Out-of-View
- In-Sight vs Out-of-Sight (Occluded)
- Within-Reach vs Out-of-Reach (defining the camera wearer's near space)

In this talk

HOI in 2D

- VISOR (masks and hand-interactions)
- HOI-Ref

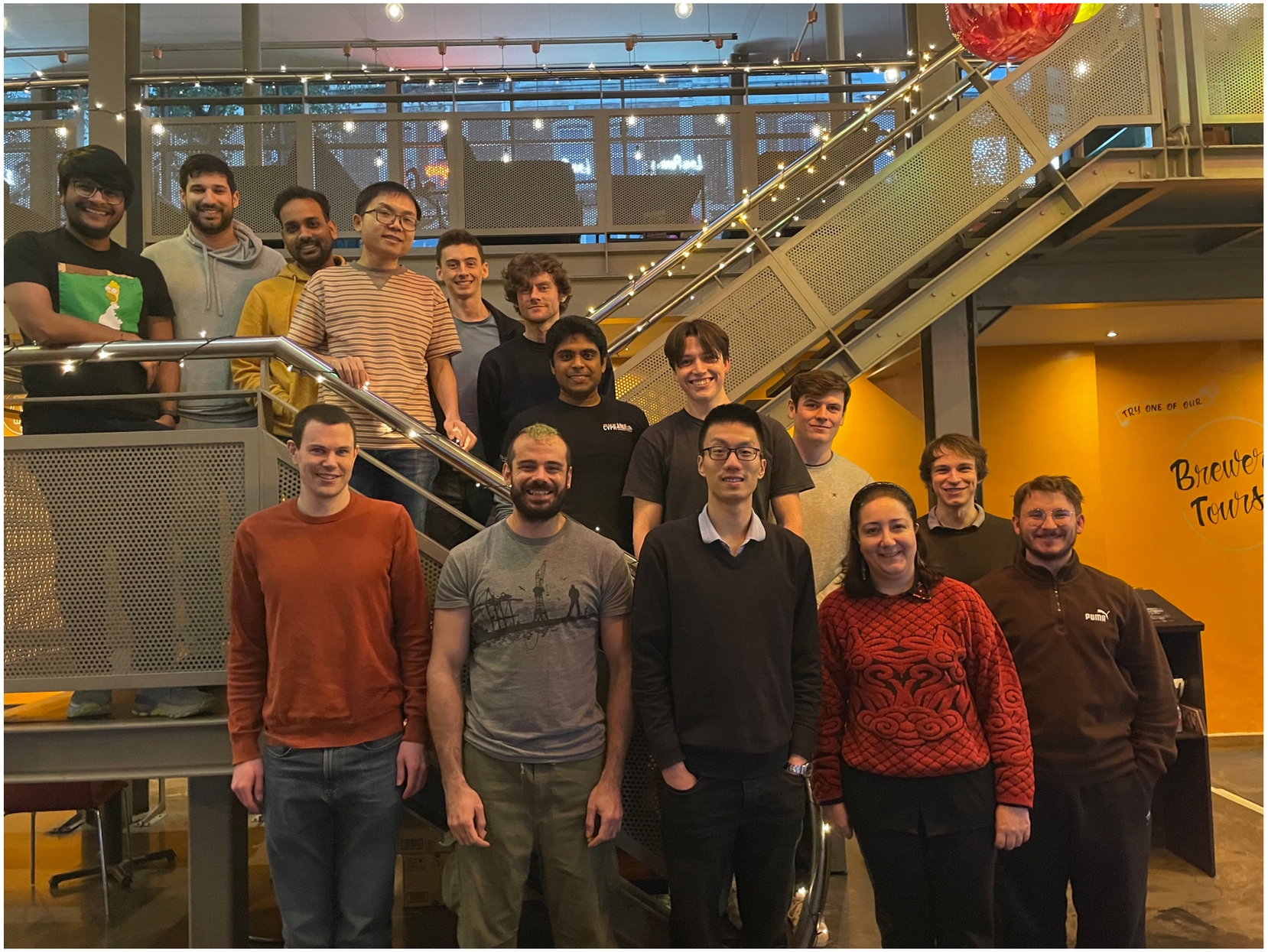
HOI 3D Reconstruction in view

- Get a Grip

HOI 3D Reconstruction in and out of view

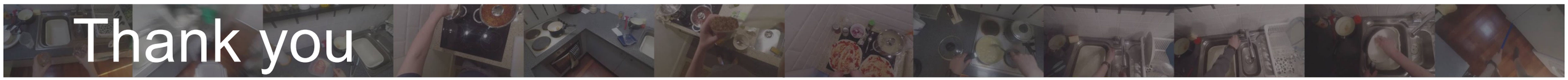
- EPIC Fields - Scene reconstruction from egocentric views
- OSNOM - 3D tracking of HOI in world coordinate frames

The Team



University of BRISTOL

Dima Damen
Rhobin W @CVPR2024



Thank you

For further info, datasets, code, publications…

http://dimadamen.github.io

@dimadamen

http://www.linkedin.com/in/dimadamen

Q&A