Object-Centric Perception for Robot Manipulation

Yu Xiang

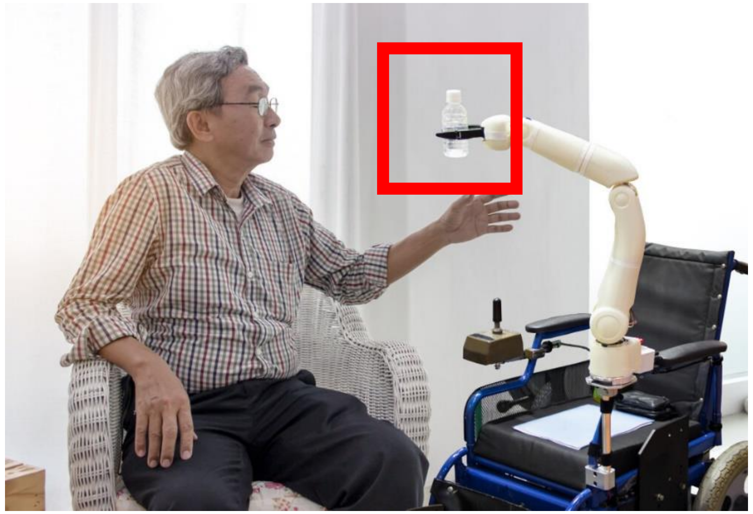
Assistant Professor

Computer Science

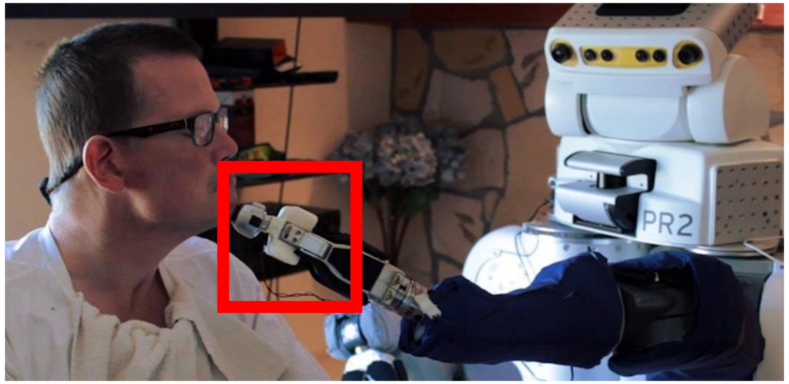
The University of Texas at Dallas

Future Intelligent Robots in Human Environments

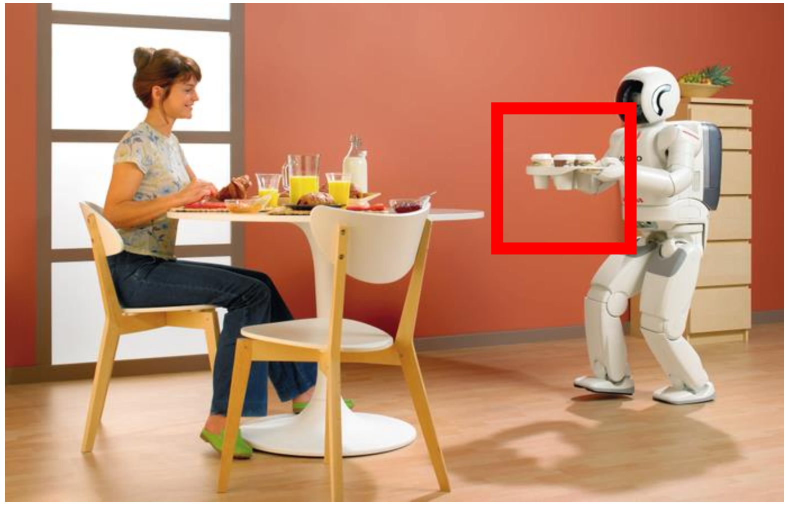
Manipulation



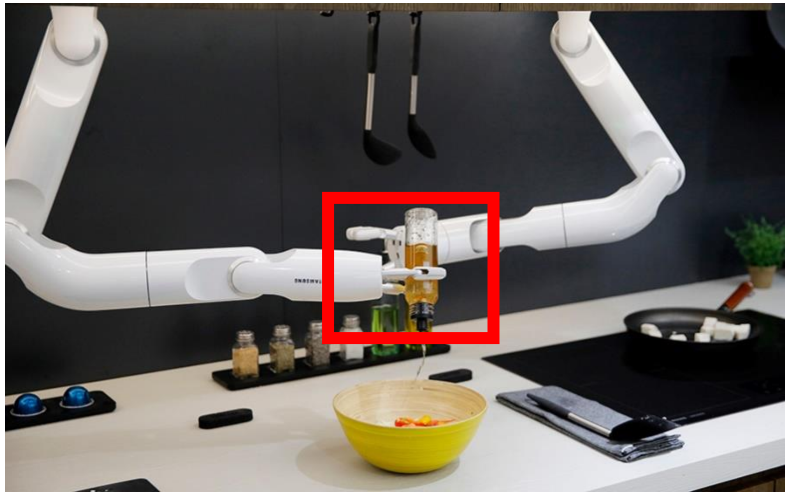
Senior Care



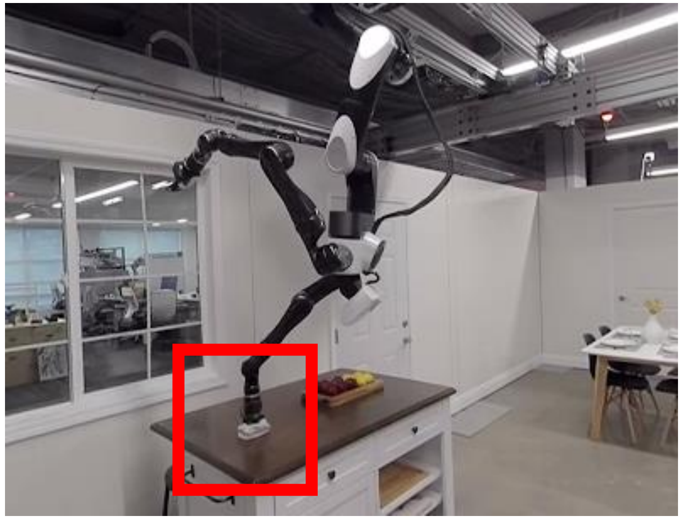
Assisting



Serving



Cooking



Cleaning

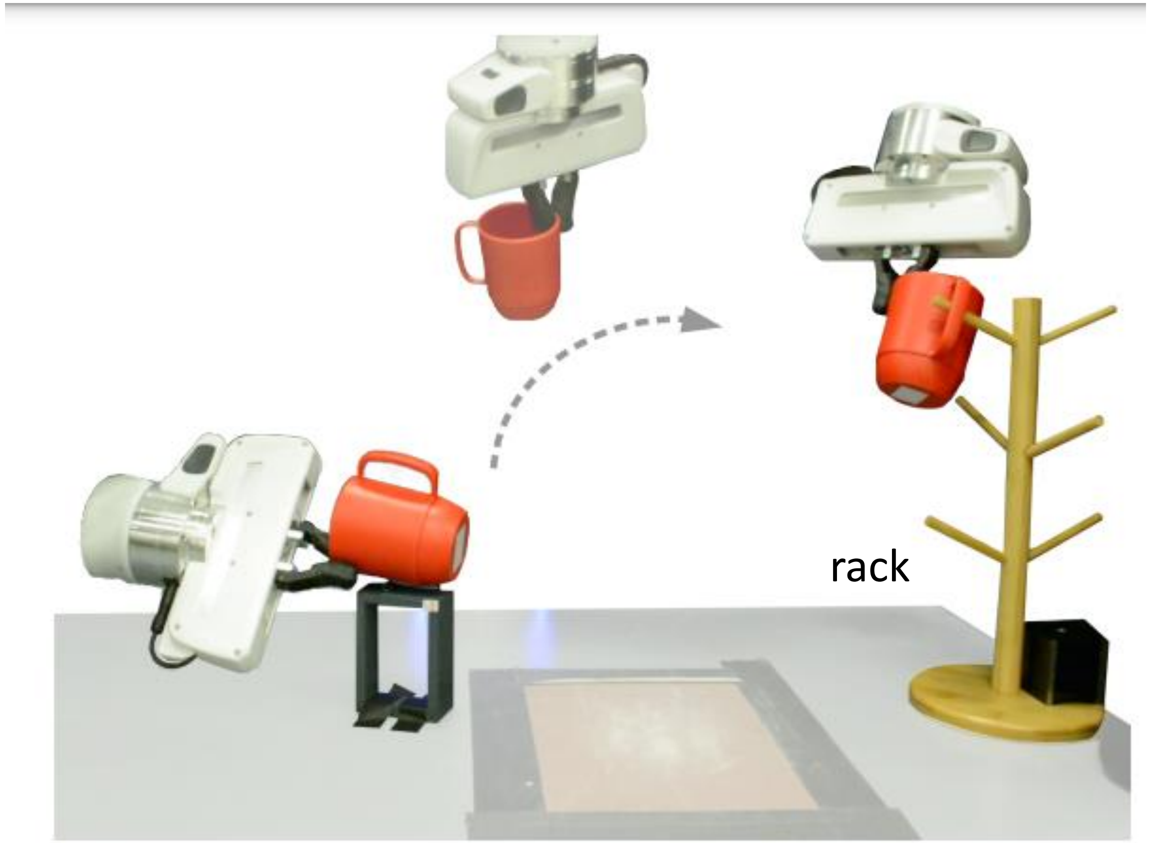
Dish washing

Object-Centric Manipulation vs. Robot-Centric Manipulation

- Object-centric
  - How the object should be controlled
  - Not specific to any robot
  - Require object perception

Generalization

- Robot-centric
  - How the robot should be controlled
  - Difficult to generalize to different robot
  - Can be end-to-end (RL)

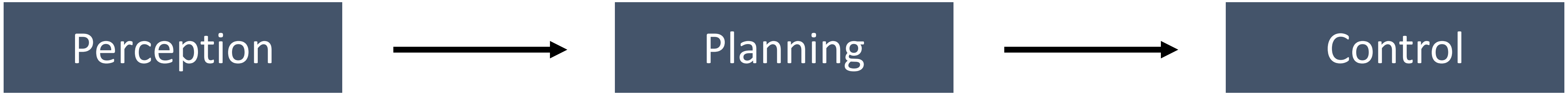


Neural Descriptor Fields. Simeonov, et al. ICRA, 2022.

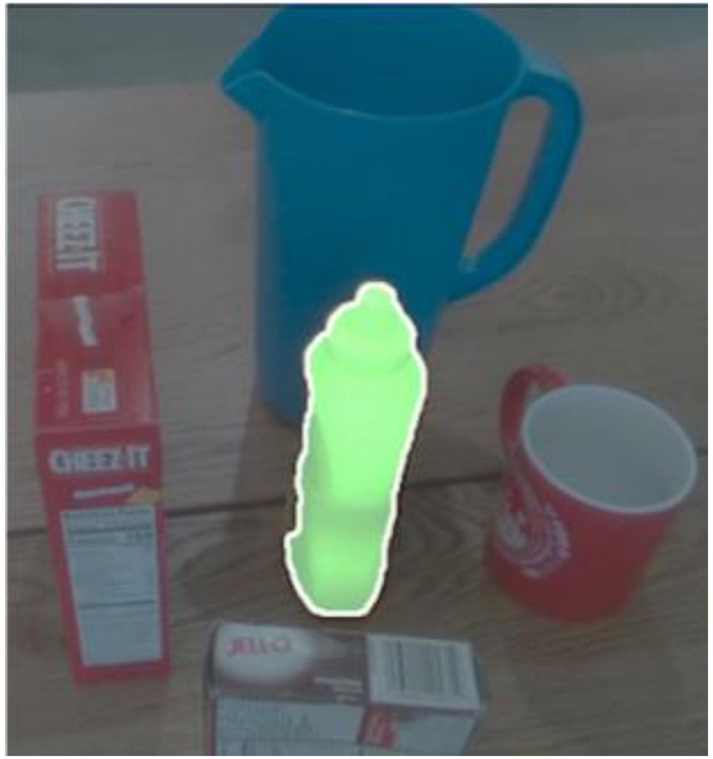
Robots in Unstructured Environments

How can a robot manipulate objects in this cluttered kitchen?

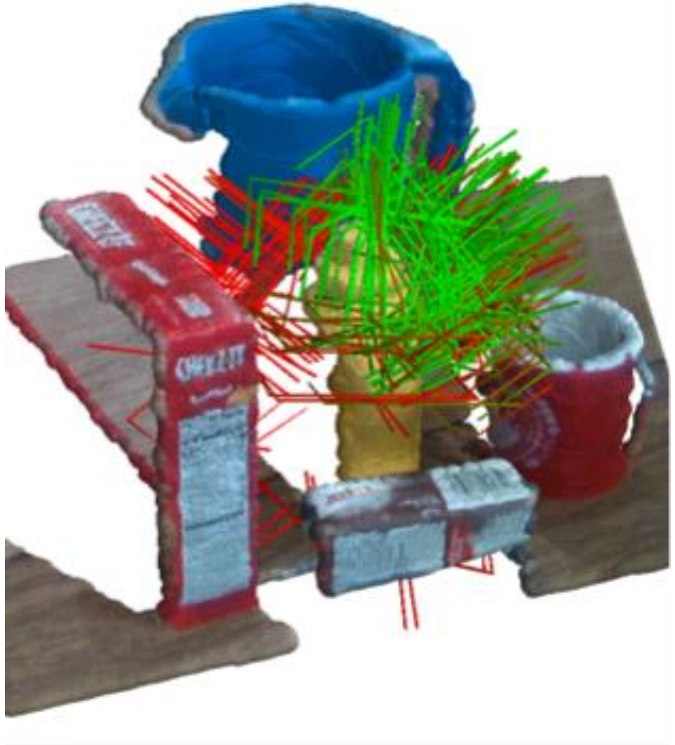
Object Model-free Robotic Grasping



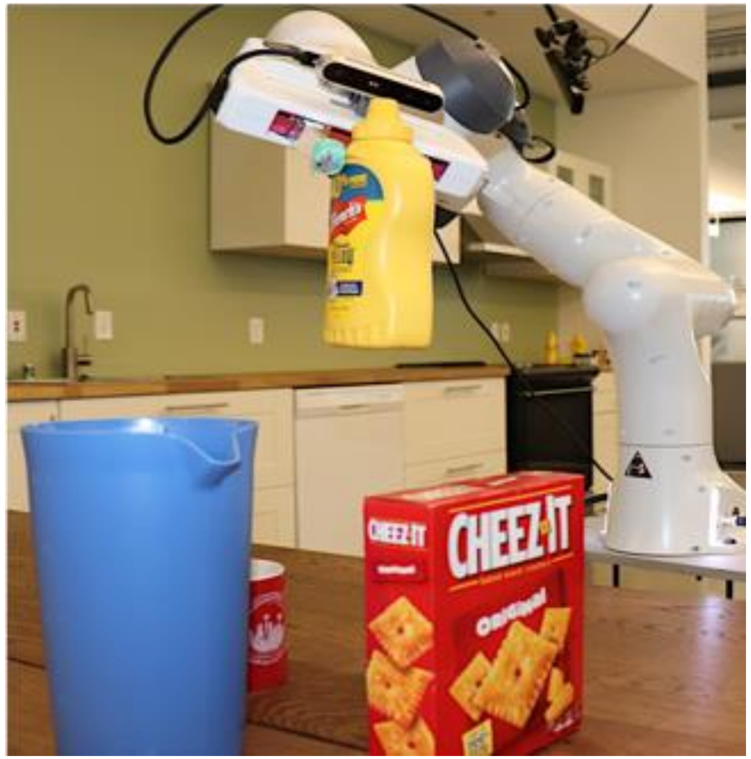
Perception → Planning → Control



Unseen object instance segmentation



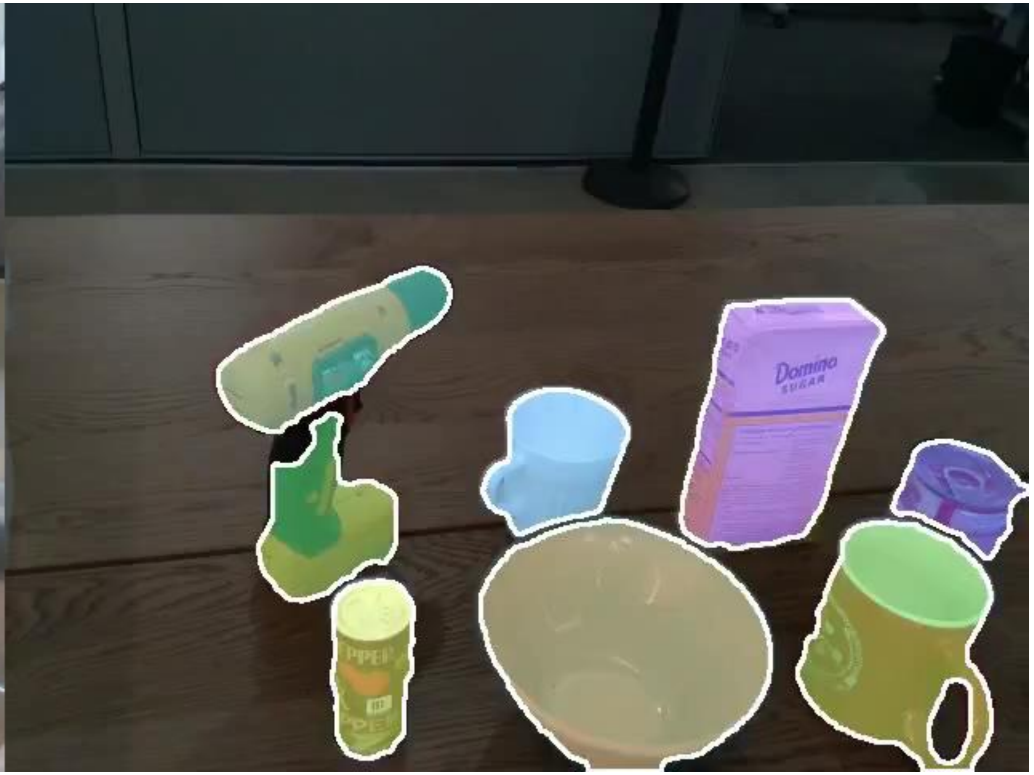
Grasp planning from point clouds



Position control to reach grasp

Figure Credit: Murali-Mousavian-Eppner-Paxton-Fox, ICRA'20

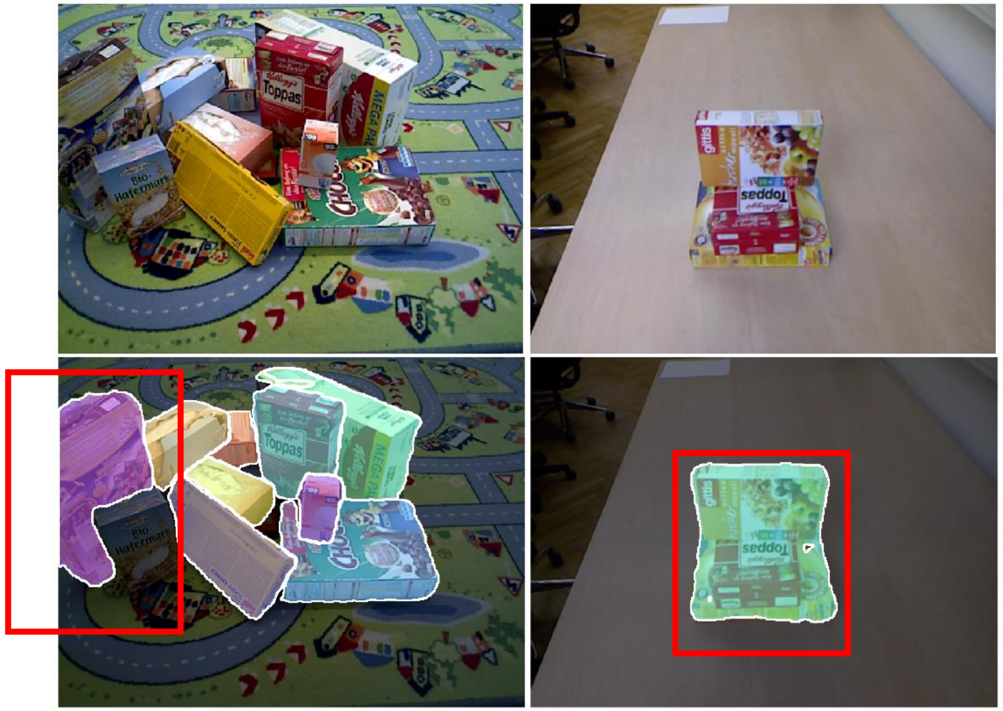
Object Model-free Robotic Grasping



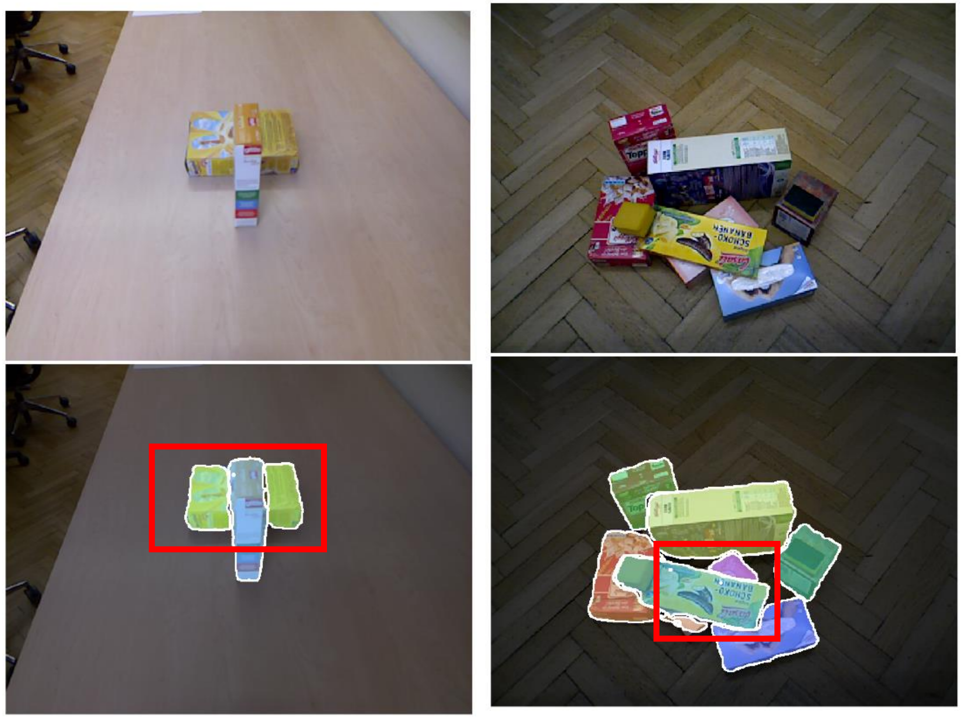
Unseen Object Instance Segmentation:
Xie-Xiang-Mousavian-Fox, CoRL’19, T-RO’21
Xiang-Xie-Mousavian-Fox, CoRL’20

6-DOF GraspNet:
Mousavian-Eppner-Fox, ICCV’19

Segmentation Failure Cases



Under-segmentation



Over-segmentation

How Can We Fix These Failures?

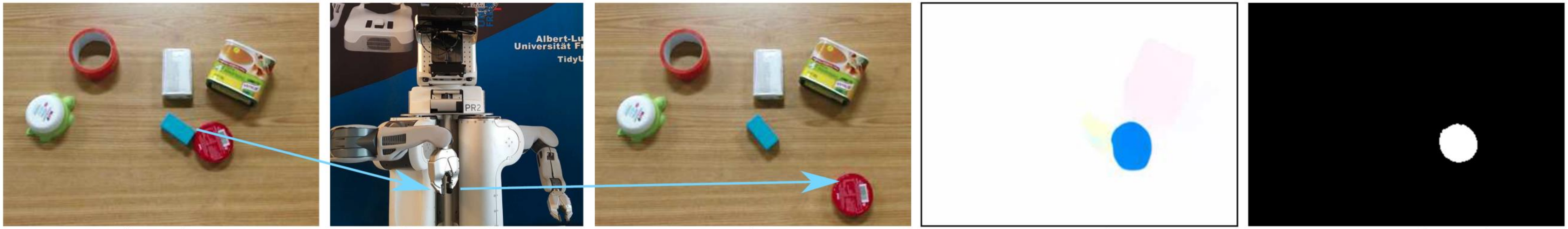
- Better models
  - Swin Transformers
  - OpenAI CLIP
  - ?

- Better training data
  - Photo-realistic synthetic data

UOAIS-Net (Back et al. ICRA’22)

  - Real-world data
    (How can we obtain real-world data for training?)

Self-supervised Segmentation

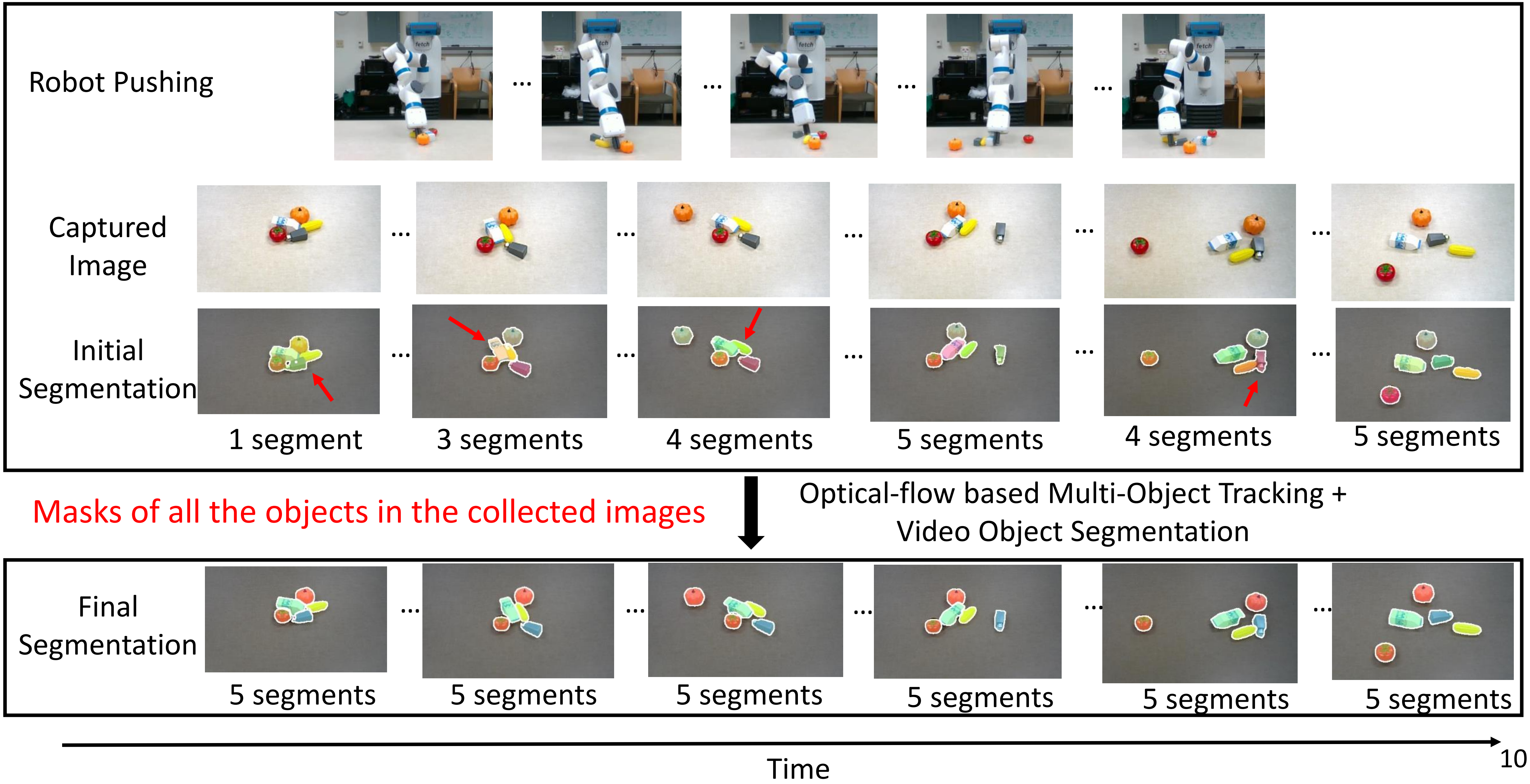


previous image robot pushing next image optical flow generated mask

- One push cannot separate objects sometimes
- These approaches can only obtain one mask in an image

[1] Andreas Eitel, Nico Hauff, and Wolfram Burgard. Self-supervised transfer learning for instance segmentation through physical interaction. IROS, 2019.
[2] Houjian Yu and Changhyun Choi. Self-supervised interactive object segmentation through a singulation-and grasping approach. ECCV, 2022.

Leveraging Long-term Robot Interaction



Robot Pushing ... ...

Captured Image ... ...

Initial Segmentation ... ...

1 segment | 3 segments | 4 segments | 5 segments | 4 segments | 5 segments

Masks of all the objects in the collected images

Optical-flow based Multi-Object Tracking + Video Object Segmentation

Final Segmentation ... ...

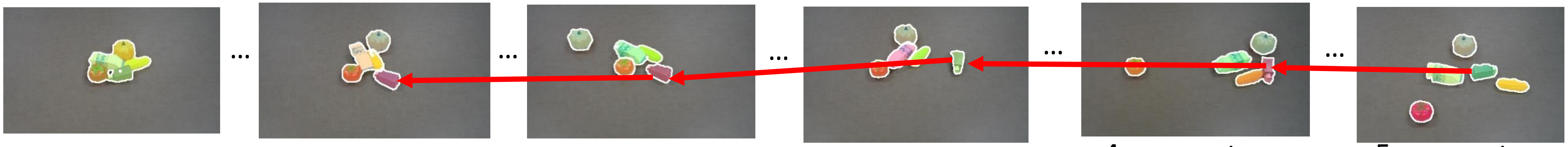
5 segments | 5 segments | 5 segments | 5 segments | 5 segments | 5 segments

Time

Leveraging Long-term Robot Interaction

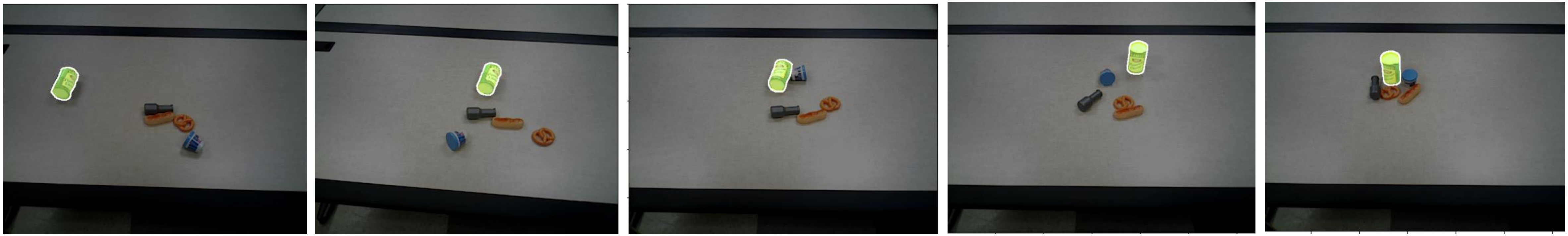
Tracking by Segmentation and Video Object Segmentation

Initial Segmentation



1 segment | 3 segments | 4 segments | 5 segments | 4 segments | 5 segments

Tracklet



Initial mask: frame 20 → frame 10 → frame 7 → frame 4 → frame 0

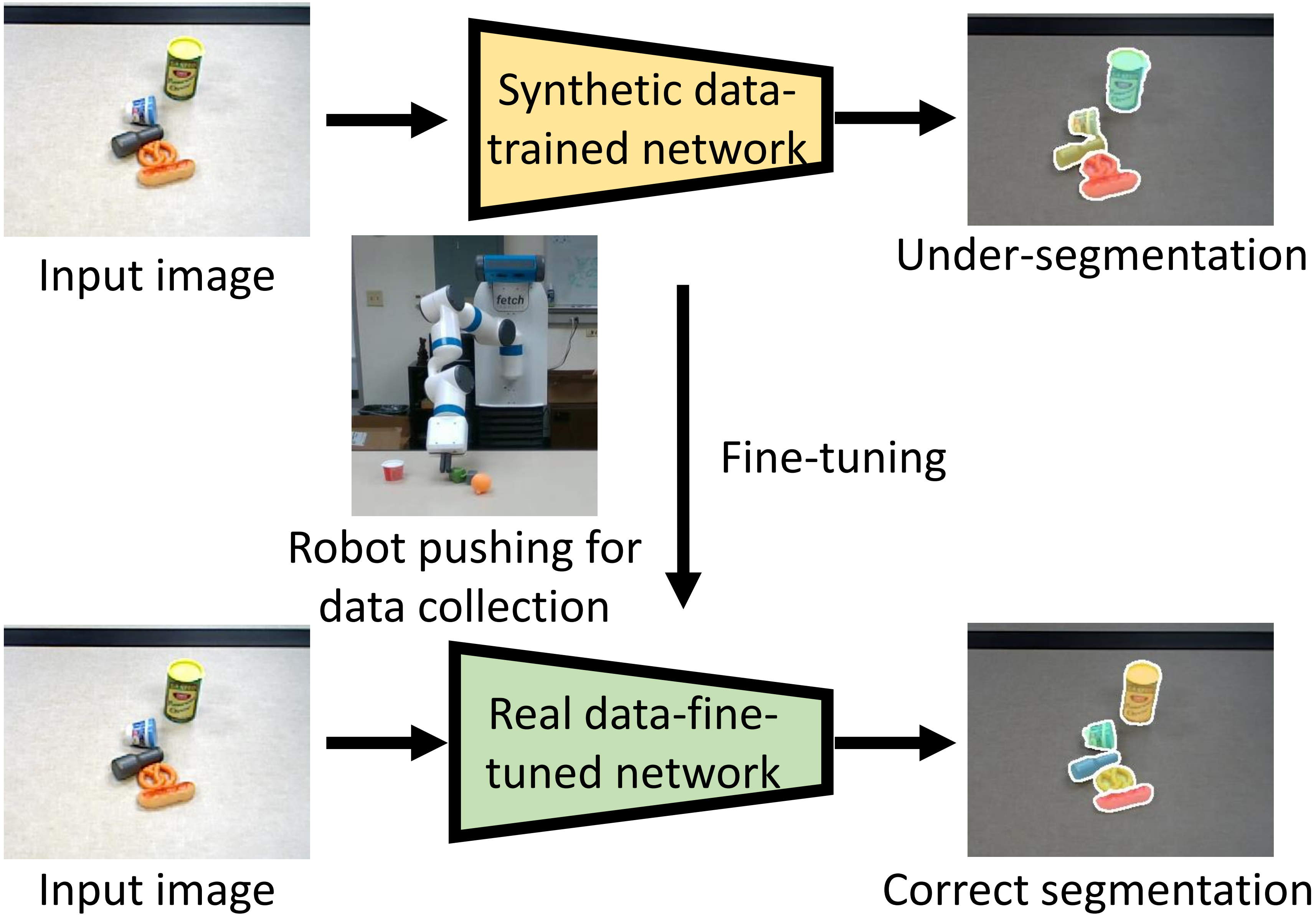
Select the highest score mask in a tracklet

Propagation to other frames

Long-Term Video Object Segmentation with an Atkinson-Shiffrin Memory Model.
Ho Kei Cheng, Alexander Schwing, ECCV, 2022. https://github.com/hkchengrex/XMem

Data Collected by the Robot

Self-supervised Segmentation with Robot Interaction



Synthetic data-trained network

Input image

Under-segmentation

Robot pushing for data collection

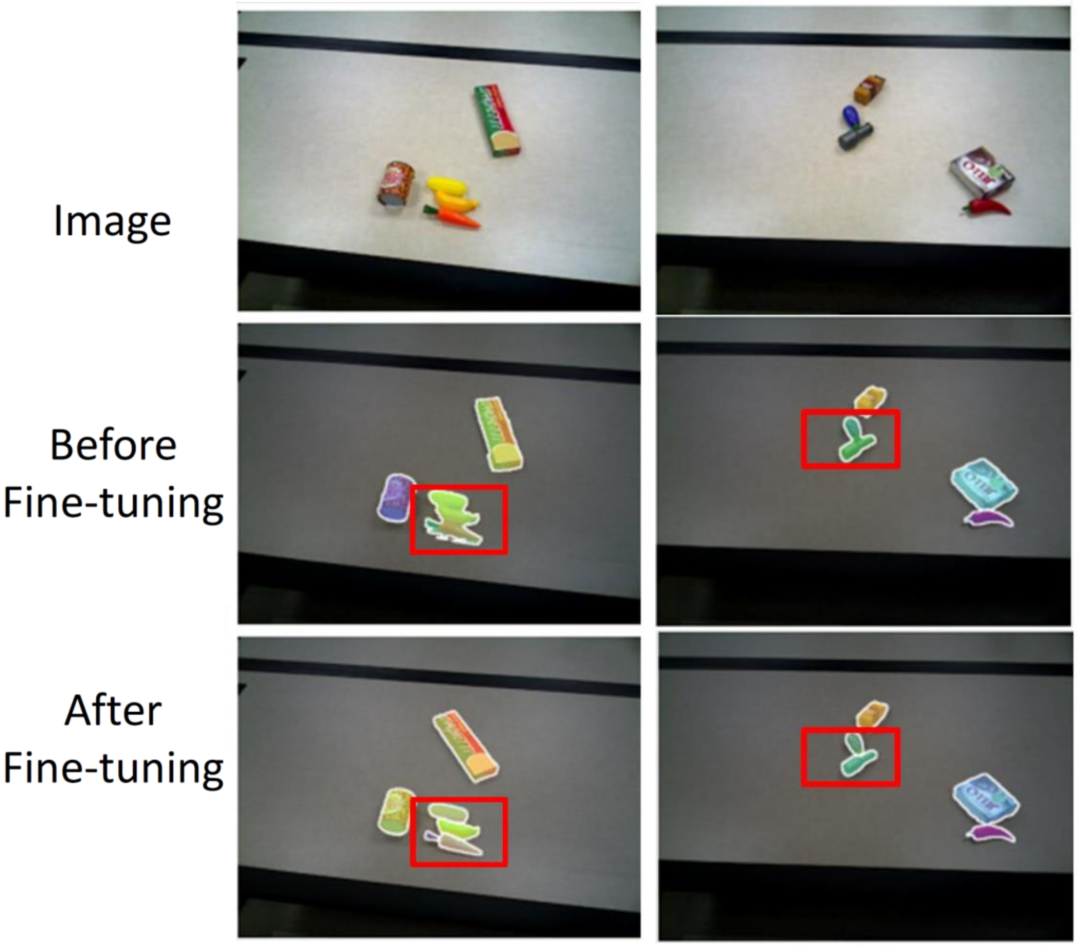
Fine-tuning

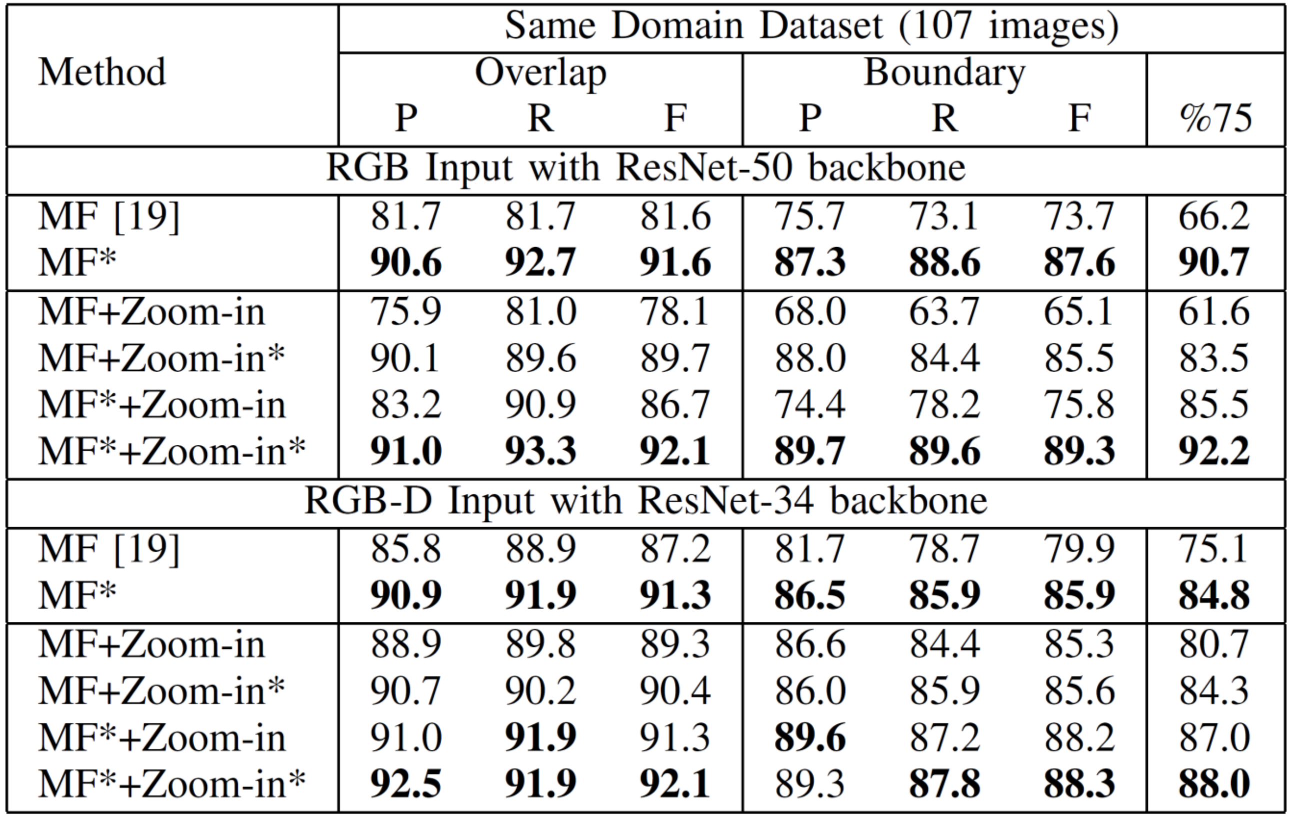
Real data-fine-tuned network

Input image

Correct segmentation

Fine-tuning MSMFormer for Unseen Object Segmentation



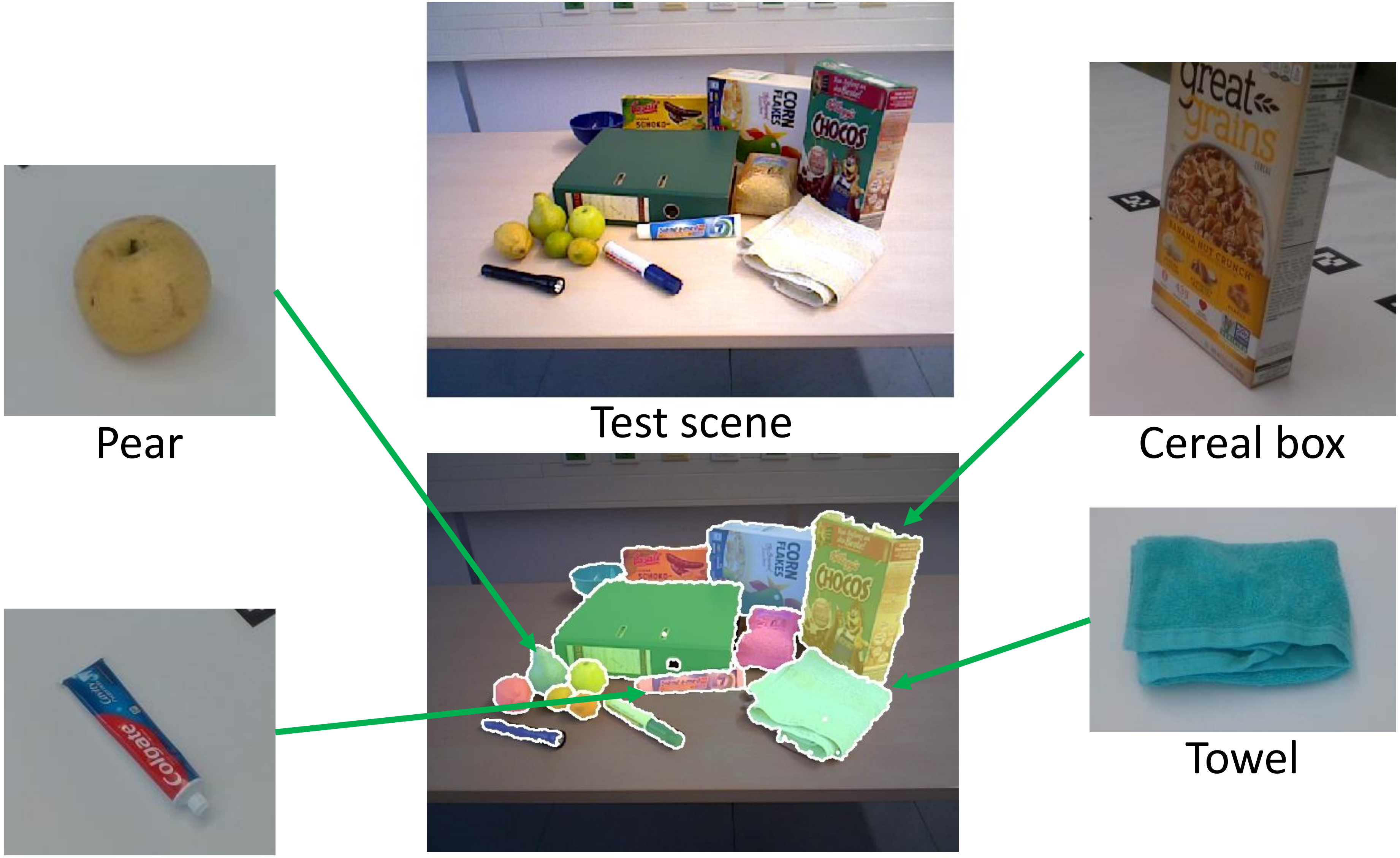


Method	Same Domain Dataset (107 images)						
	Overlap			Boundary			
	P	R	F	P	R	F	%75
RGB Input with ResNet-50 backbone							
MF [19]	81.7	81.7	81.6	75.7	73.1	73.7	66.2
MF*	**90.6**	**92.7**	**91.6**	**87.3**	**88.6**	**87.6**	**90.7**
MF+Zoom-in	75.9	81.0	78.1	68.0	63.7	65.1	61.6
MF+Zoom-in*	90.1	89.6	89.7	88.0	84.4	85.5	83.5
MF*+Zoom-in	83.2	90.9	86.7	74.4	78.2	75.8	85.5
MF*+Zoom-in*	**91.0**	**93.3**	**92.1**	**89.7**	**89.6**	**89.3**	**92.2**
RGB-D Input with ResNet-34 backbone							
MF [19]	85.8	88.9	87.2	81.7	78.7	79.9	75.1
MF*	**90.9**	**91.9**	**91.3**	**86.5**	**85.9**	**85.9**	**84.8**
MF+Zoom-in	88.9	89.8	89.3	86.6	84.4	85.3	80.7
MF+Zoom-in*	90.7	90.2	90.4	86.0	85.9	85.6	84.3
MF*+Zoom-in	91.0	**91.9**	91.3	**89.6**	87.2	88.2	87.0
MF*+Zoom-in*	**92.5**	**91.9**	**92.1**	89.3	**87.8**	**88.3**	**88.0**

*: model after fine-tuning

Top-Down Grasping

Few-Shot Object Recognition



Pear

Test scene

Cereal box

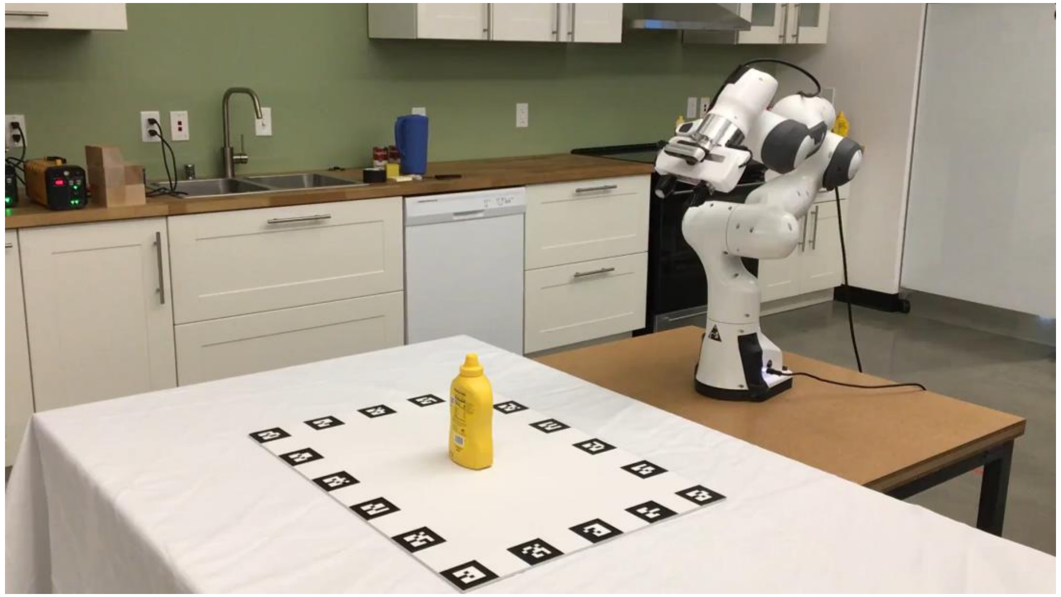
Towel

Toothpaste

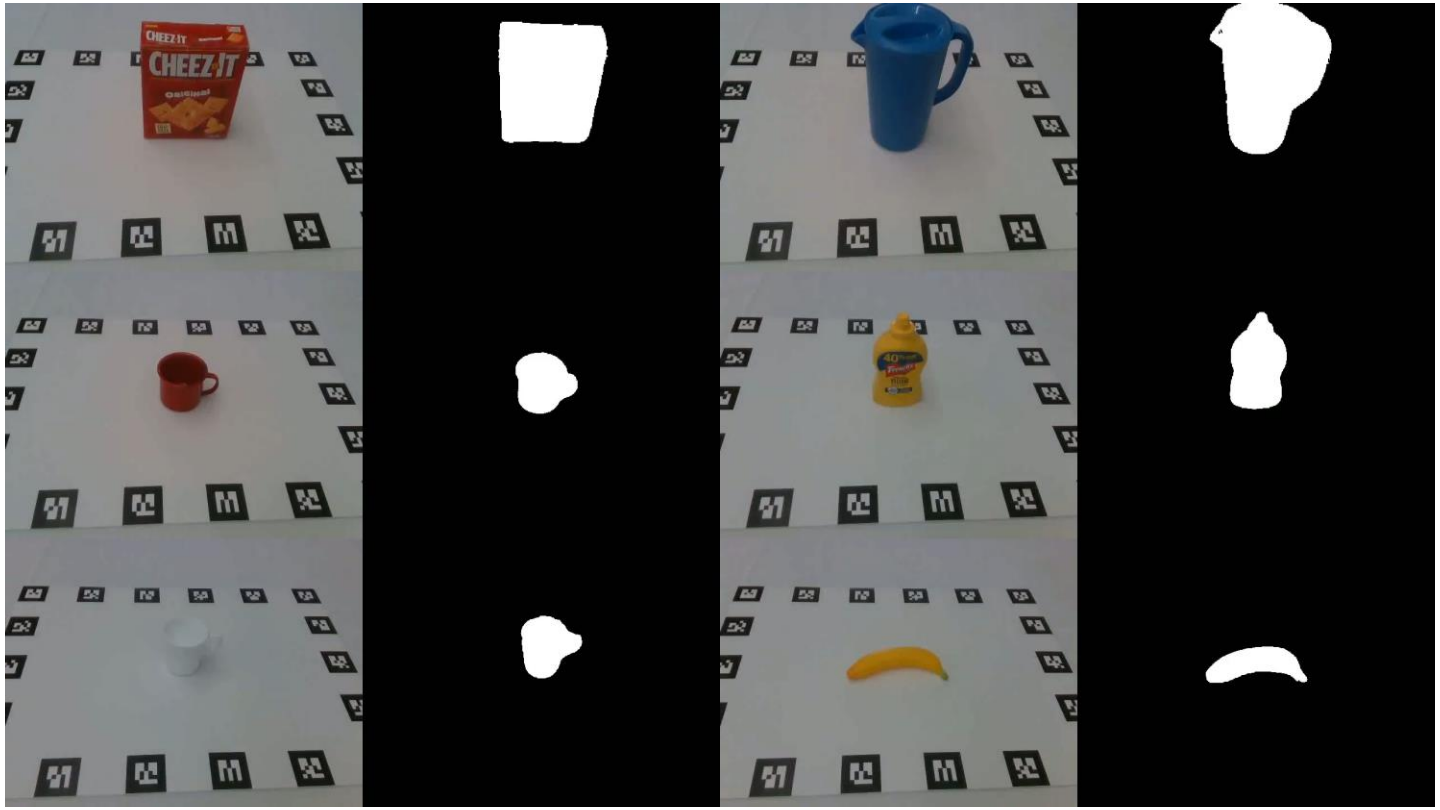
Unseen Object Instance Segmentation

Few-Shot Object Recognition

- A large-scale dataset for few-shot object recognition



Training data collected by a robot

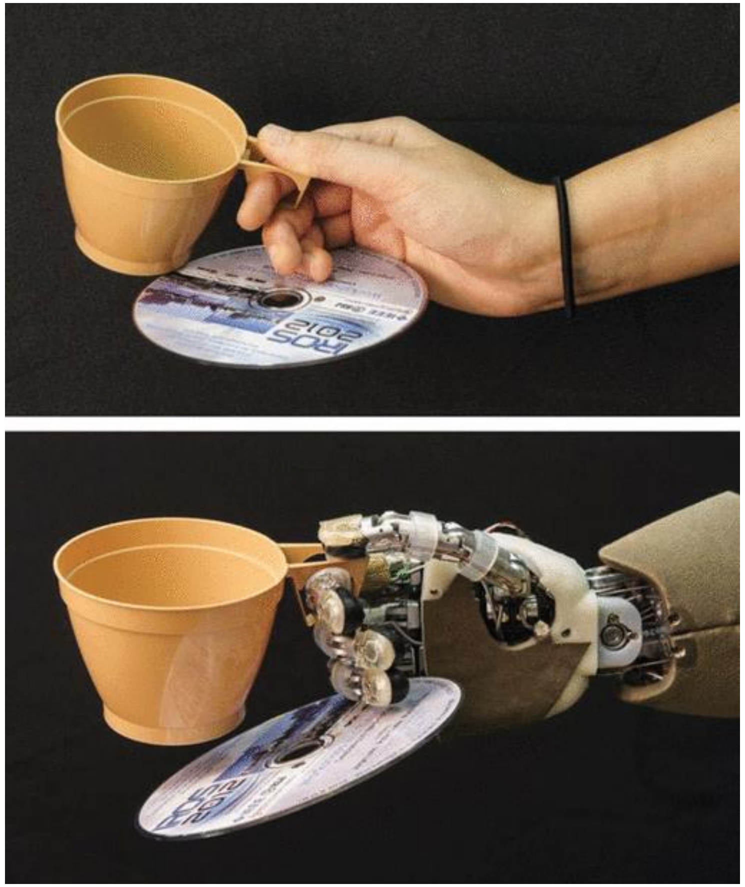


- 336 objects
- 198 object categories
- 9 images per object
- RGB-D images with segmentation masks and camera poses

FewSOL: A Dataset for Few-Shot Object Learning in Robotic Environments
Jishnu Jaykumar P, Yu-Wei Chao, Yu Xiang. ICRA, 2023.

Object-Centric Grasp Transfer

Grasp Transfer



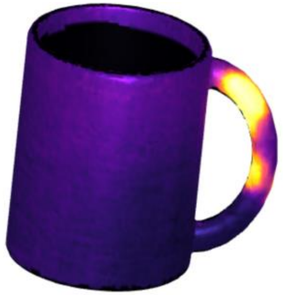
Barrett

Allegro

Human Hand

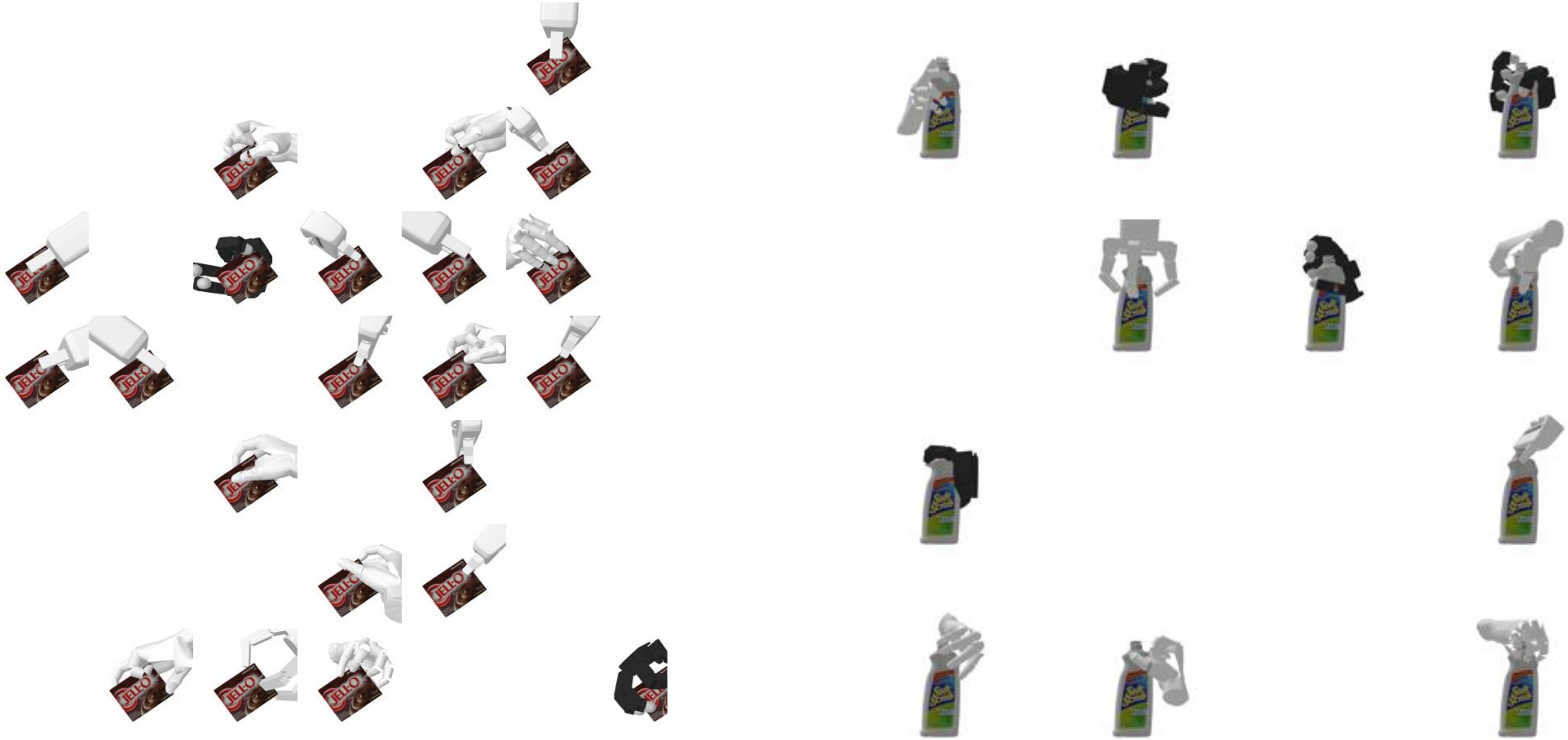
Franka Panda

Fetch Gripper



Object-centric contact regions

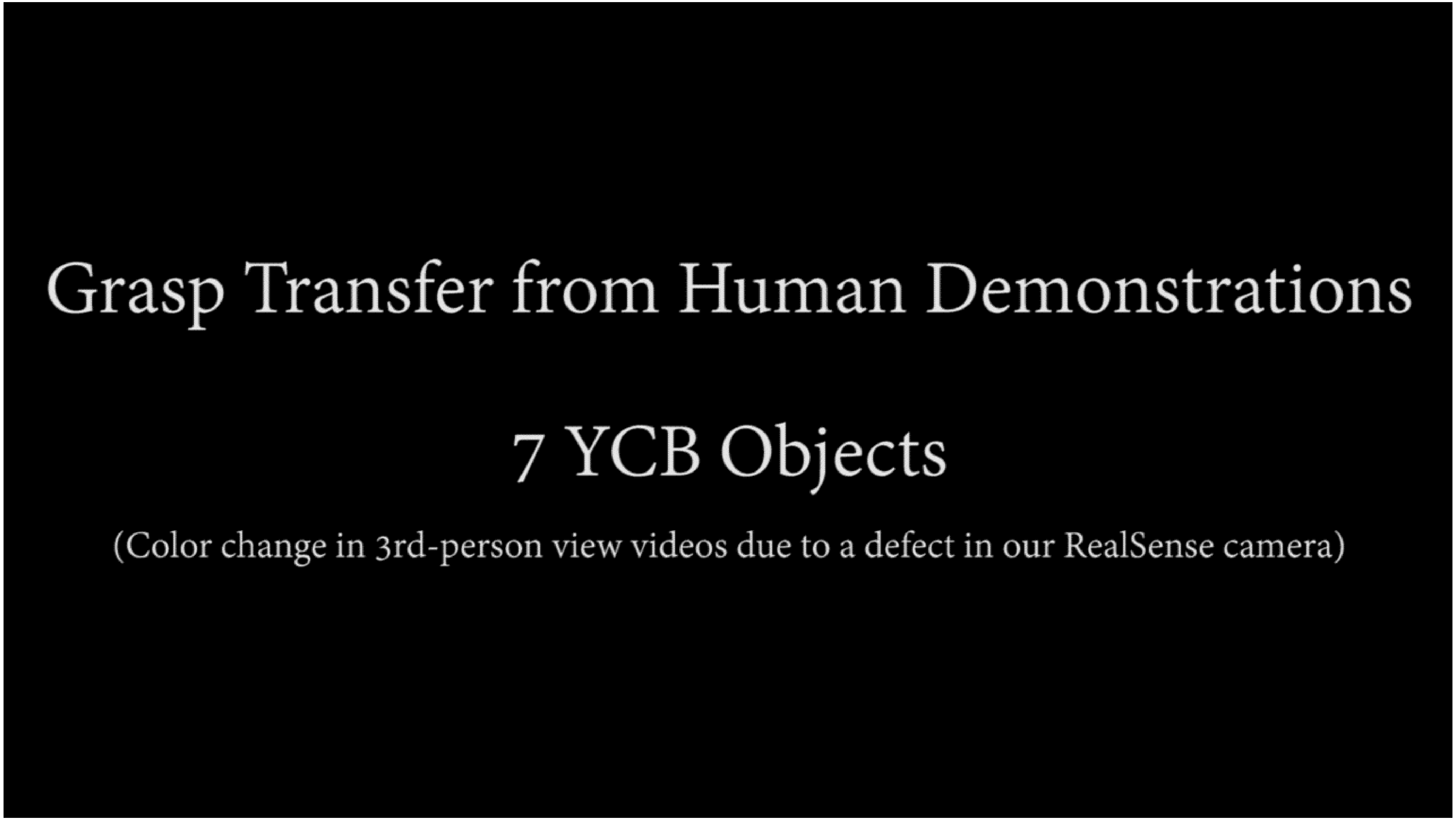
NeuralGrasps



t-SNE visualization of learned latent space

NeuralGrasps: Learning Implicit Representations for Grasps of Multiple Robotic Hands
Ninad Khargonkar, Neil Song, Zesheng Xu, Balakrishnan Prabhakaran, Yu Xiang. CoRL, 2022.

Object-Centric Grasp Transfer



Grasp Transfer from Human Demonstrations

7 YCB Objects

(Color change in 3rd-person view videos due to a defect in our RealSense camera)

Conclusion

- Object-centric perception for manipulation
  - Segmenting unseen objects → Grasping of unseen objects
  - Few-shot object recognition → object grounding in cluttered scenes
  - Grasp transfer among multiple grippers → sharing grasping skills among robots
- End-goal: robots use objects to perform tasks

yu.xiang@utdallas.edu

Thank you!