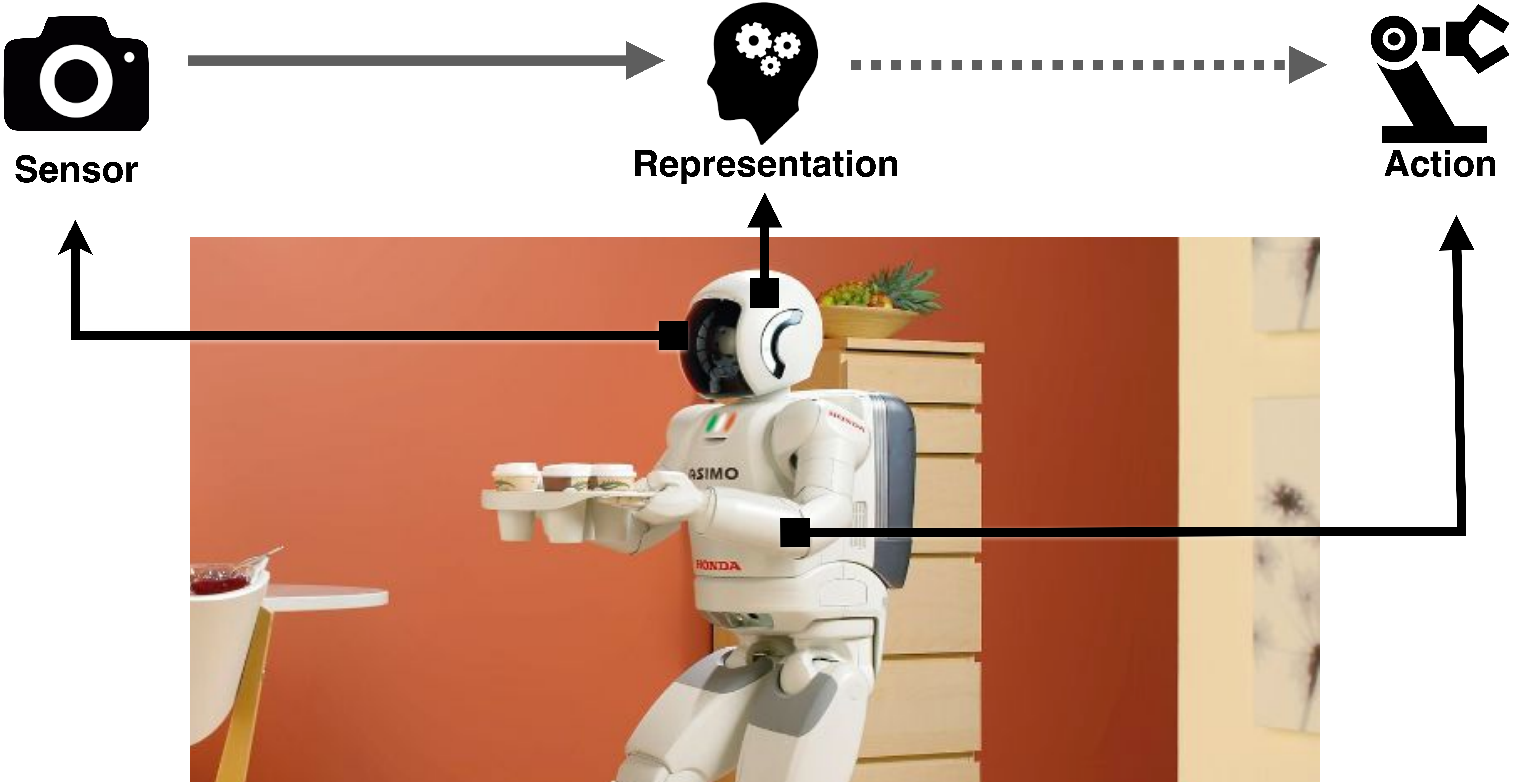
Learning Visual Representations for Generalizable Manipulation

Shuran Song

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

What’s Visual Representations?

What's Visual Representations?



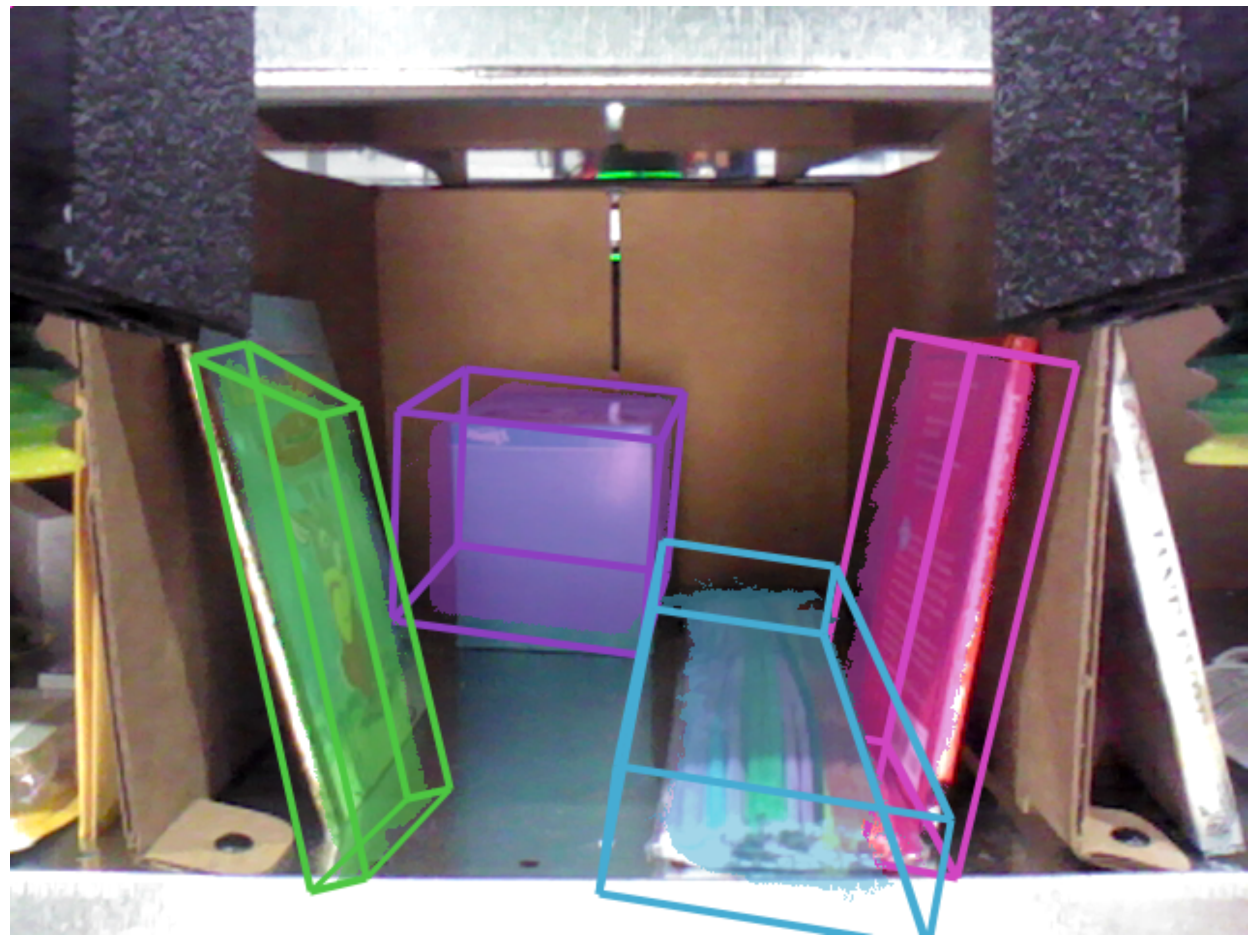
What's Visual Representations?

What's Visual Representations?

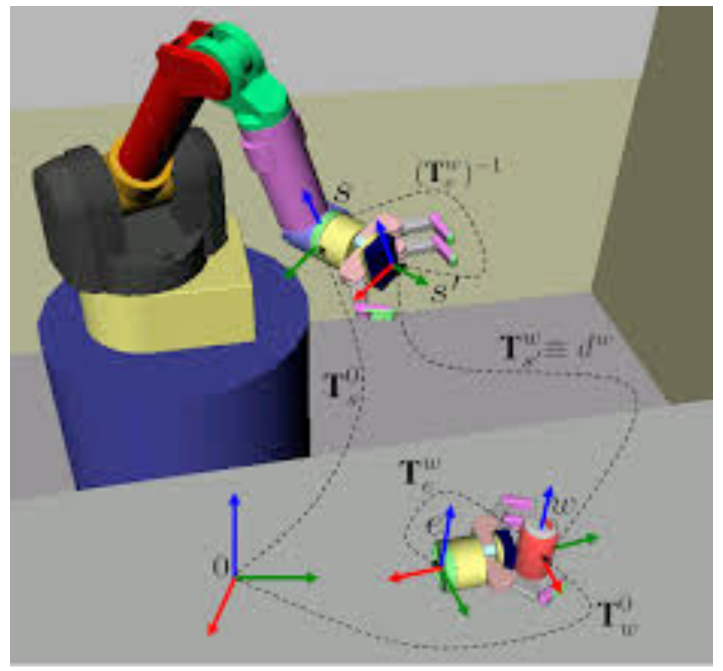
Sensor

Representation

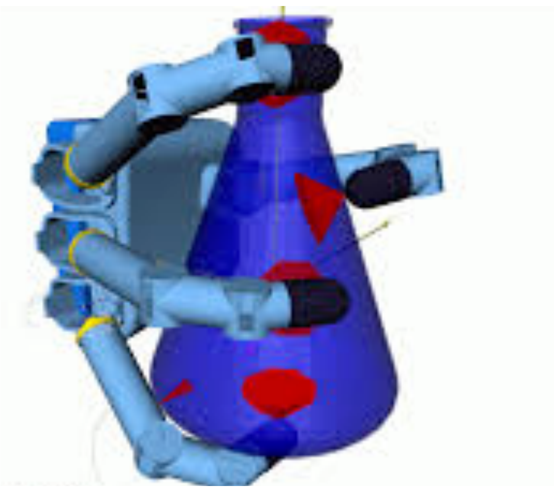
Action



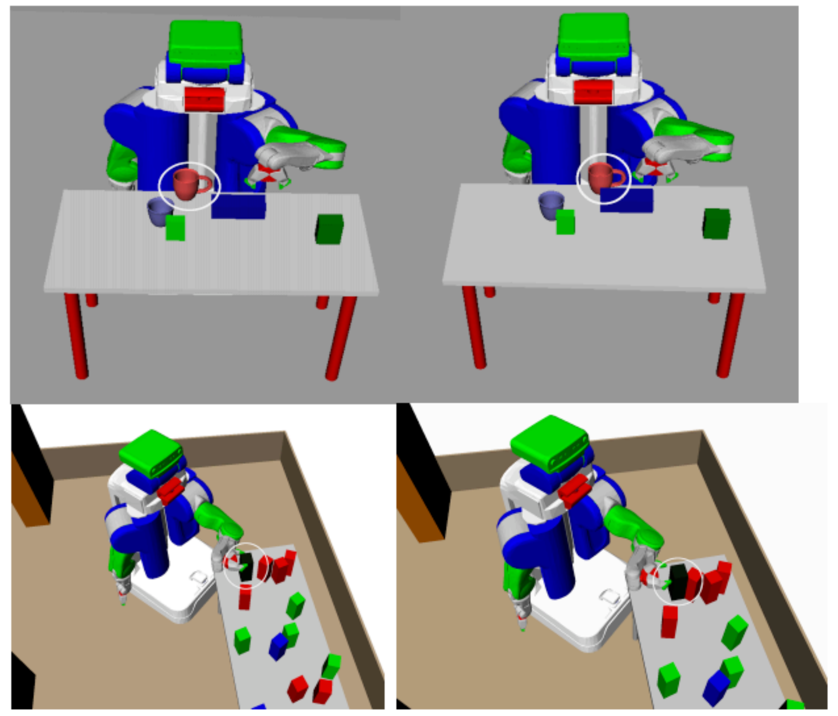
3D Object Detection
6D Poses Estimation



Berenson et al., 2009a
Berenson and Srinivasa, 2010

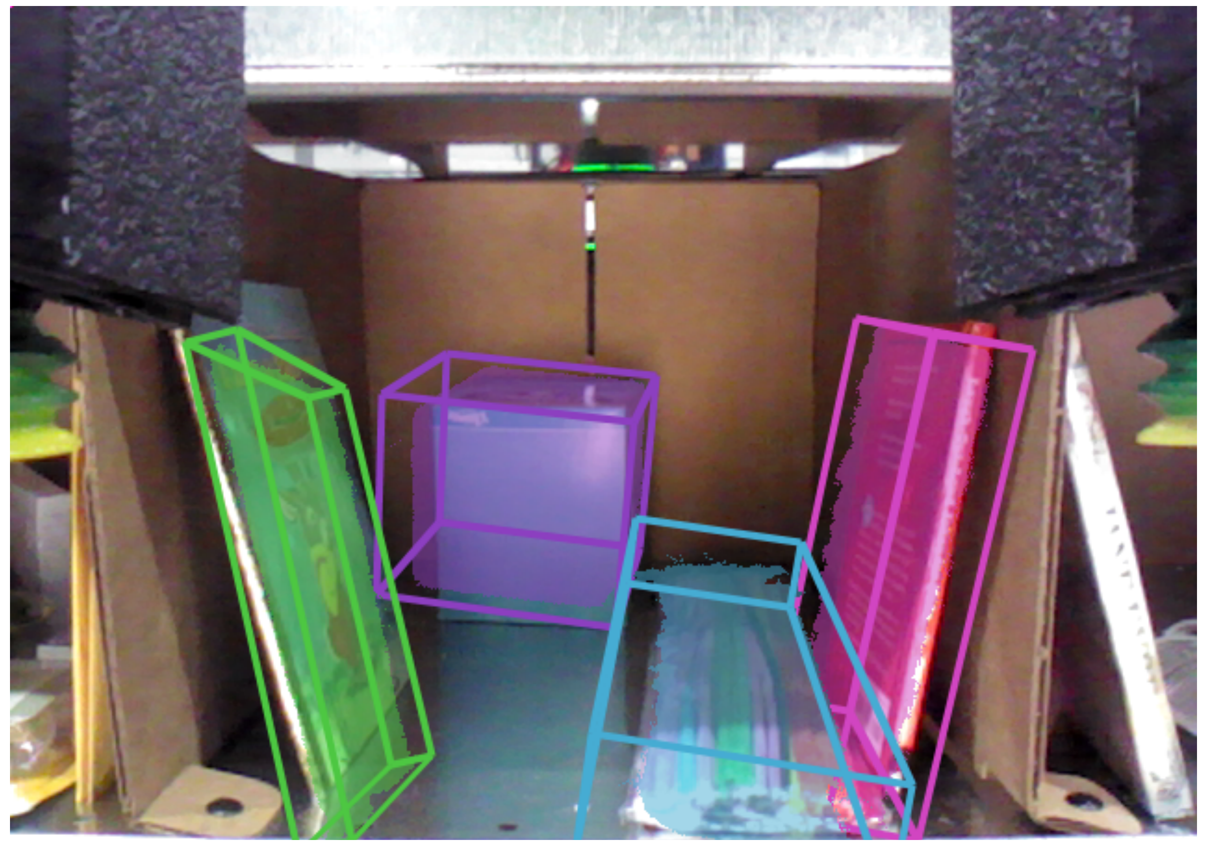


Miller and Allen 2009



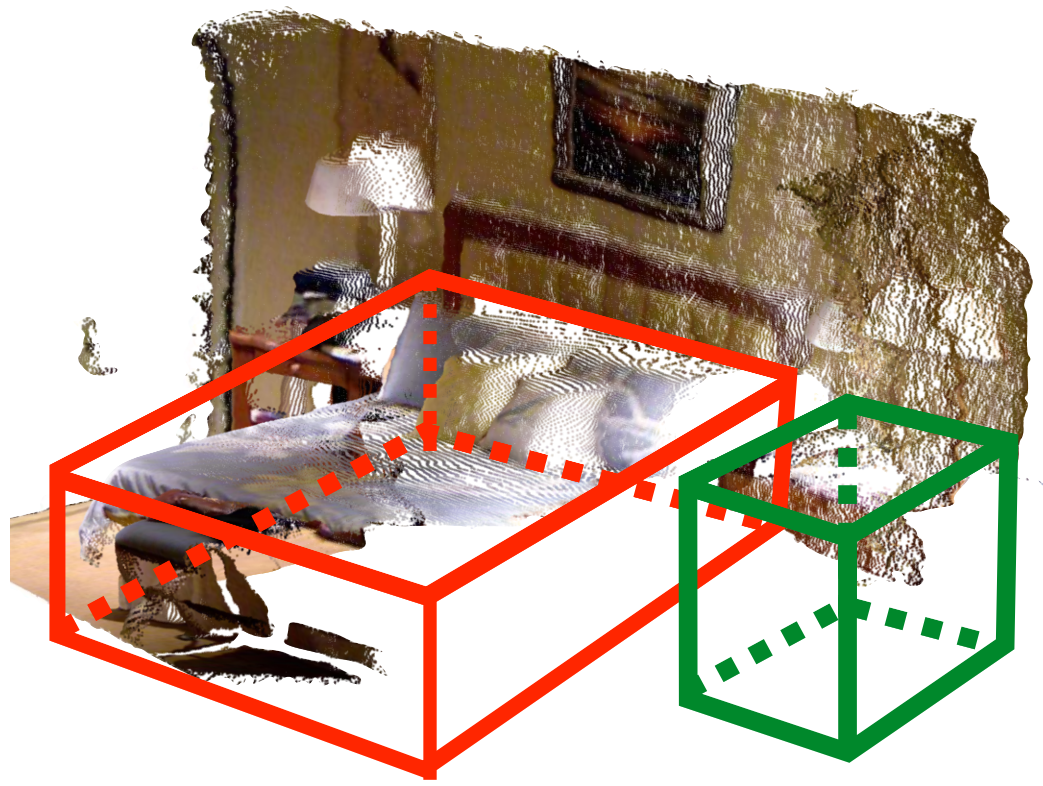
Kimm et al 2019

Object Detection + Pose Estimation

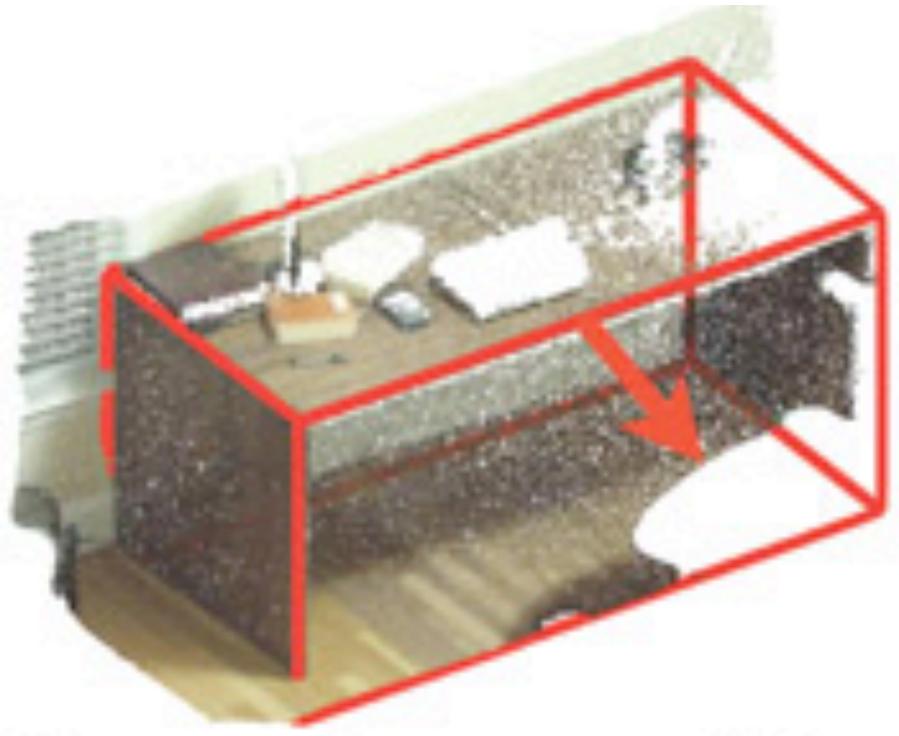


3D Object Detection
6D Poses Estimation

SlidingShapes
Song and Xiao
ECCV'14

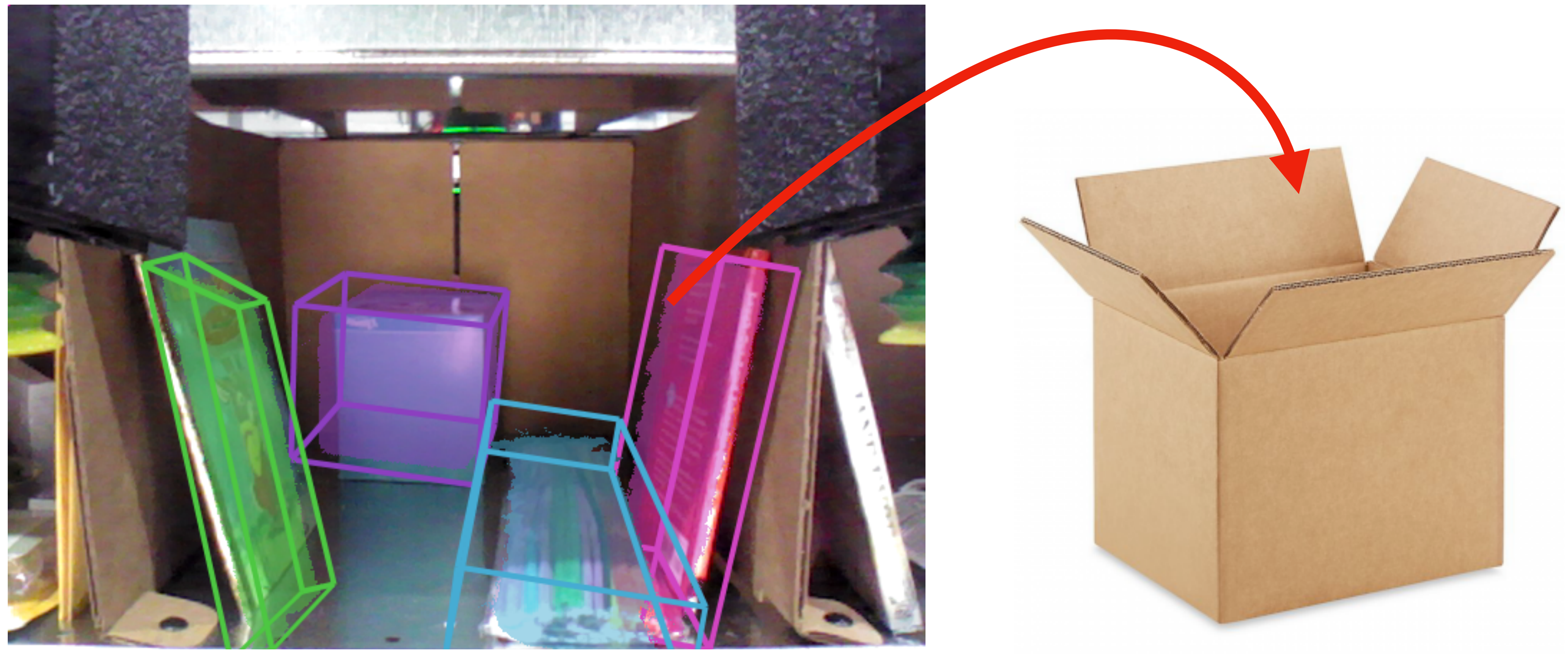


DeepSlidingShapes
Song and Xiao
CVPR'16



SUNRGB-D
Song et al.
CVPR'15

Object Detection + Pose Estimation



3D Object Detection
6D Poses Estimation

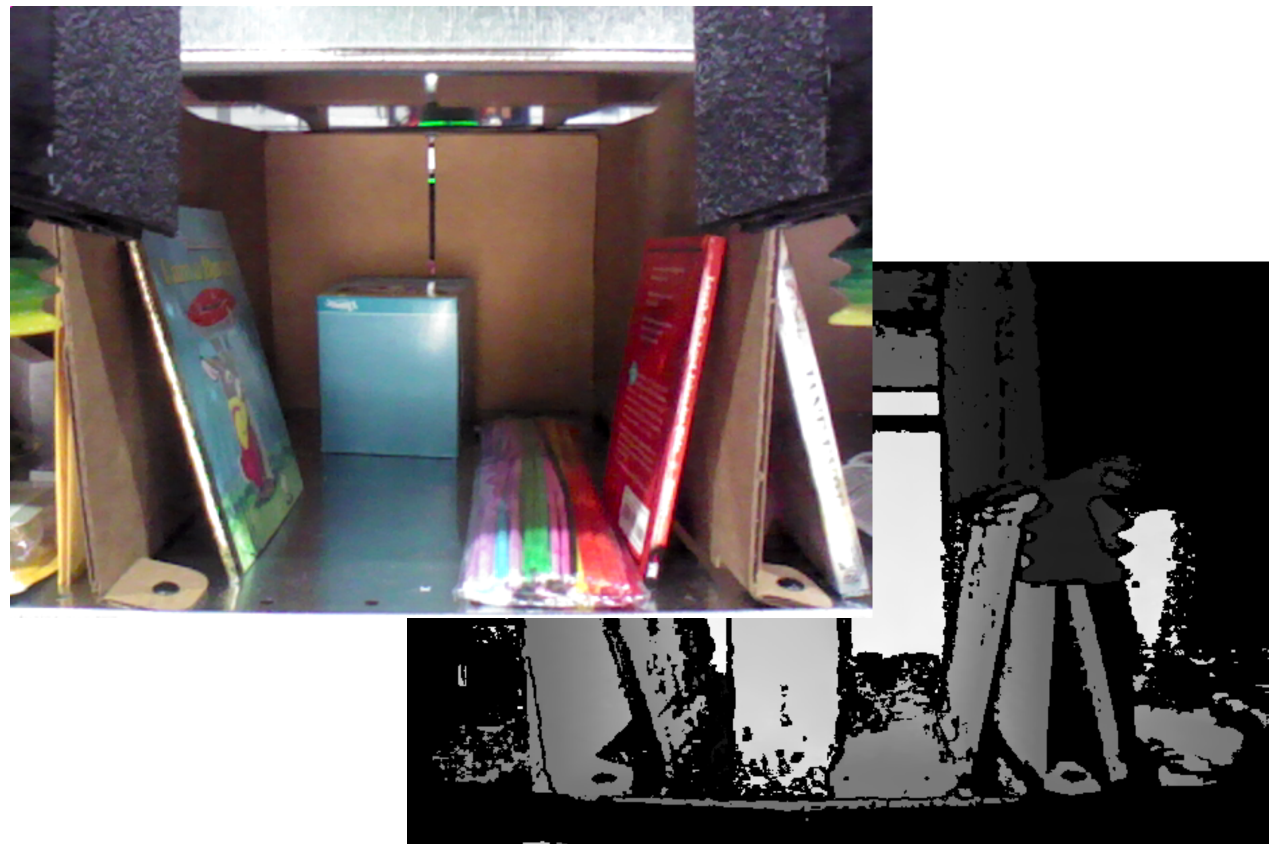
Team MIT-Princeton at Amazon Picking Challenge 2017

Amazon Picking Challenge 2016

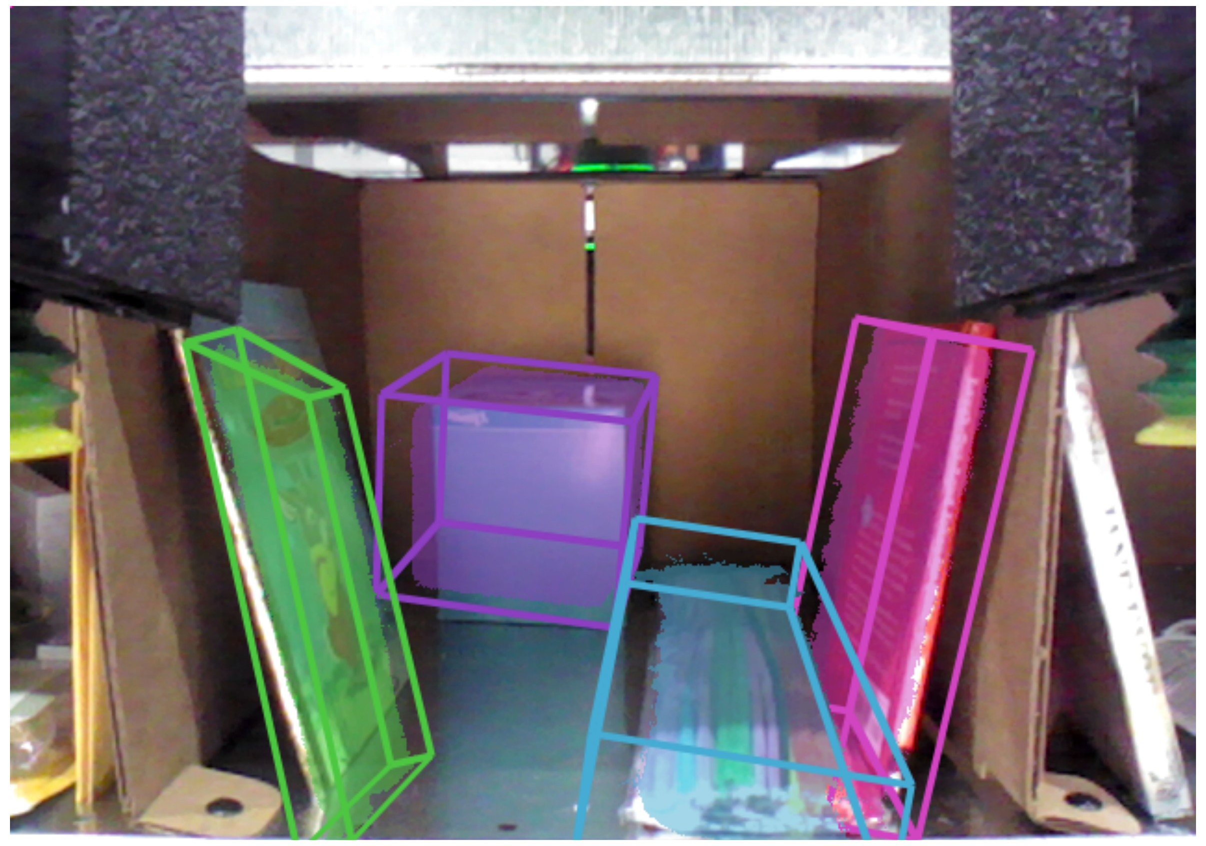
Sensor

Representation

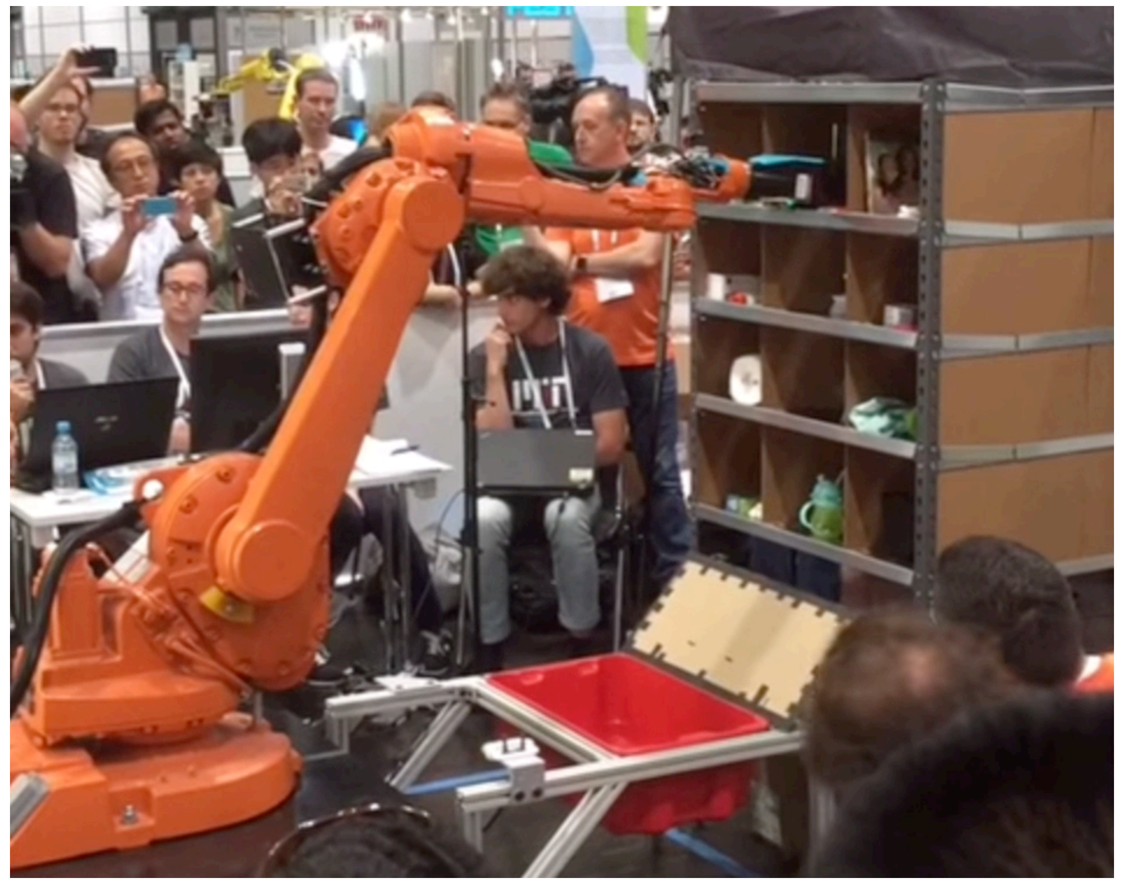
Action



RGB-D image



3D Object Detection
6D Poses Estimation



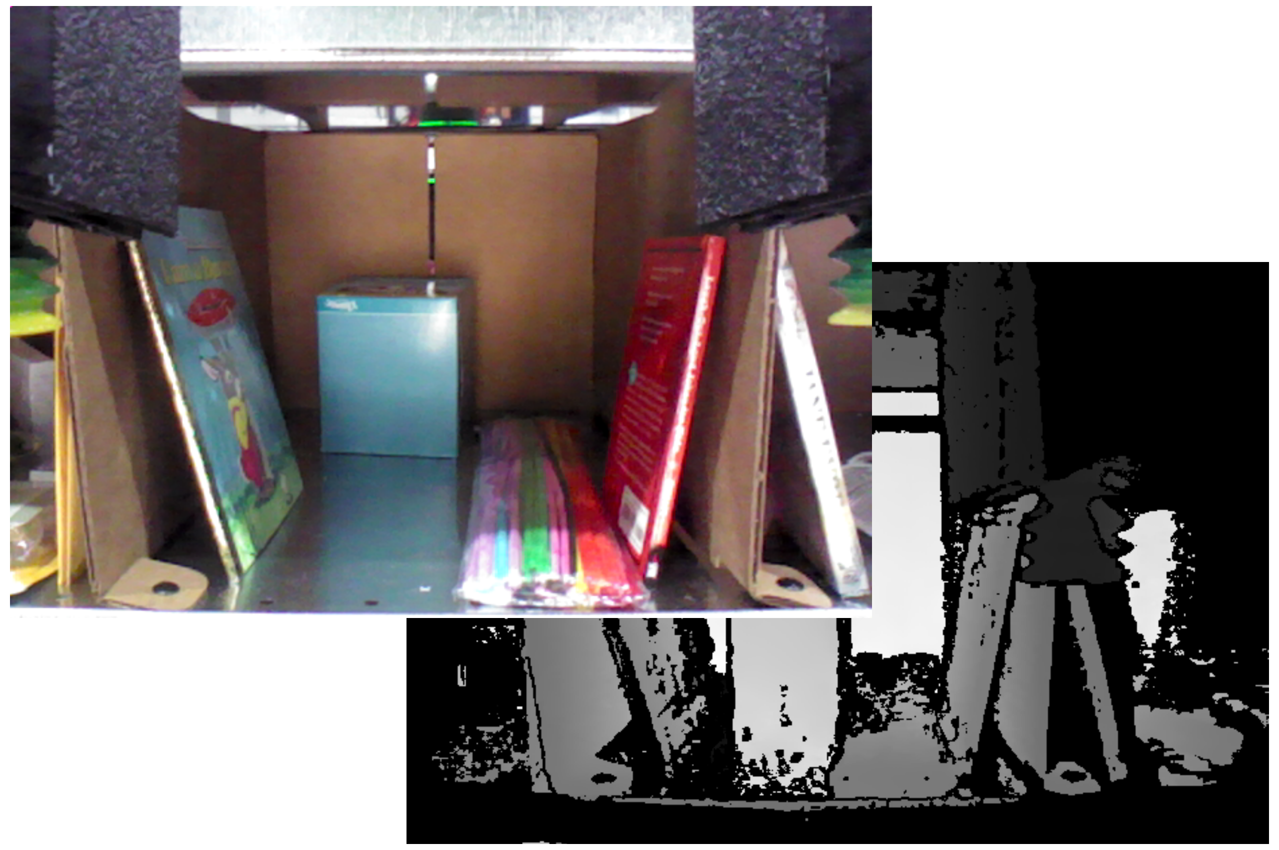
Motion Planing

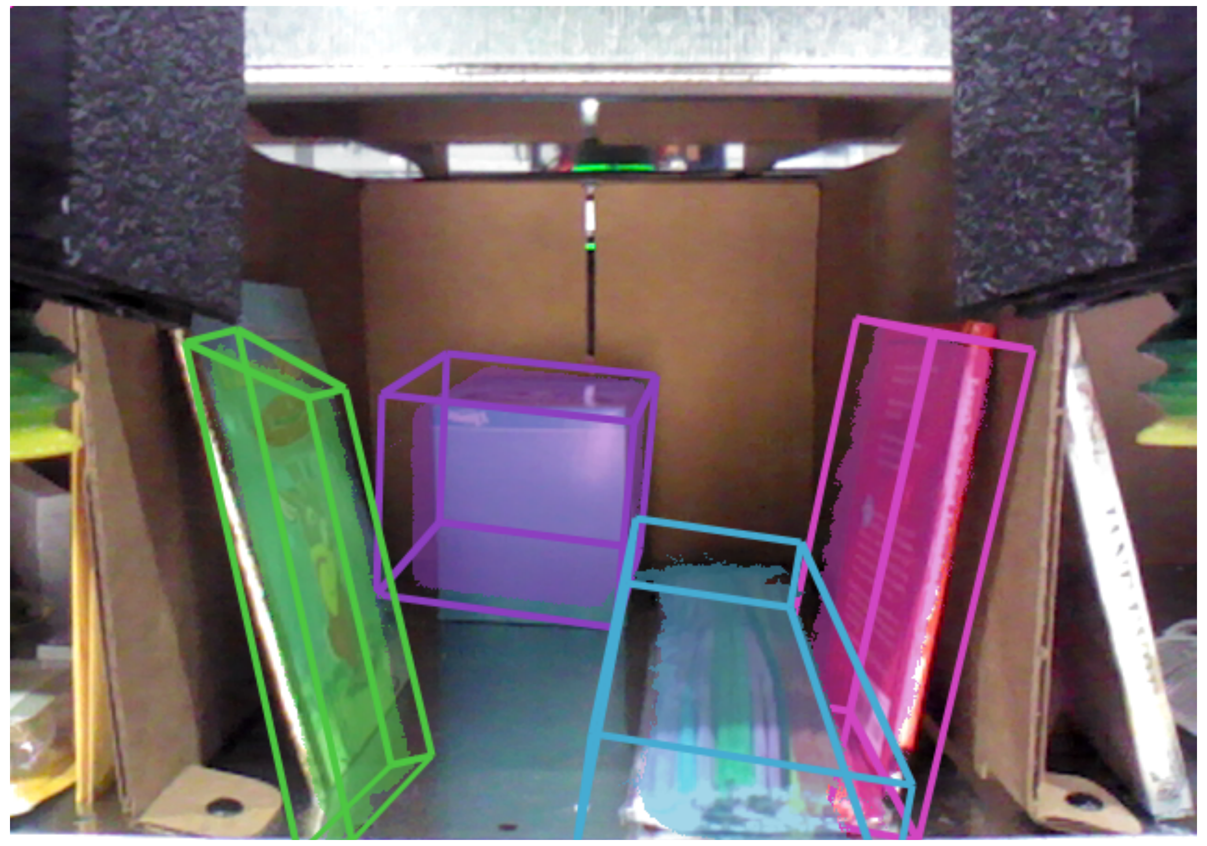
Princeton Vision Group

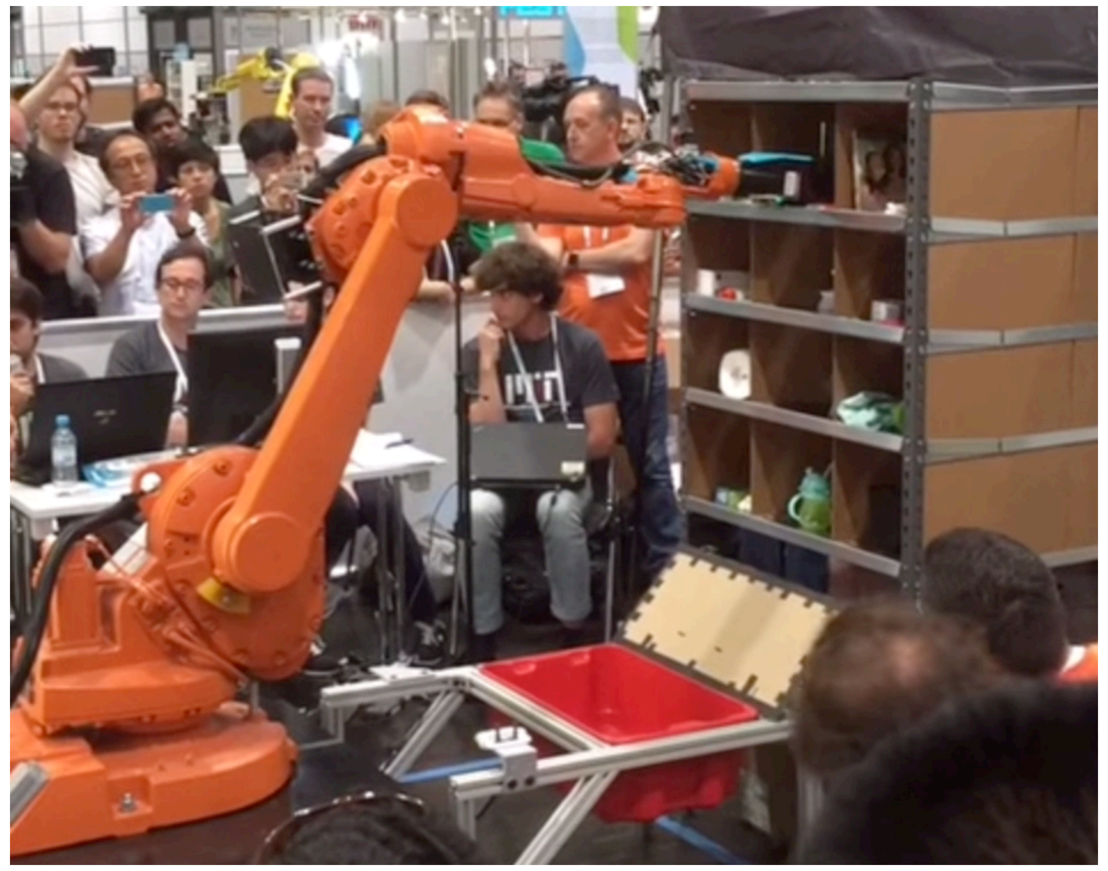
MIT MCube Lab

Robotic Pick-and-Place of Novel Objects in Clutter with Multi-Affordance Grasping and Cross-Domain Image Matching
A. Zeng, **S. Song**, K. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazeli, F. Alet, N. C. Dafle, R. Holladay, I. Morona, P. Q. Nair, D. Green, I. Taylor, W. Liu, T. Funkhouser, A. Rodriguez (**ICRA2018**)

Amazon Picking Challenge 2016







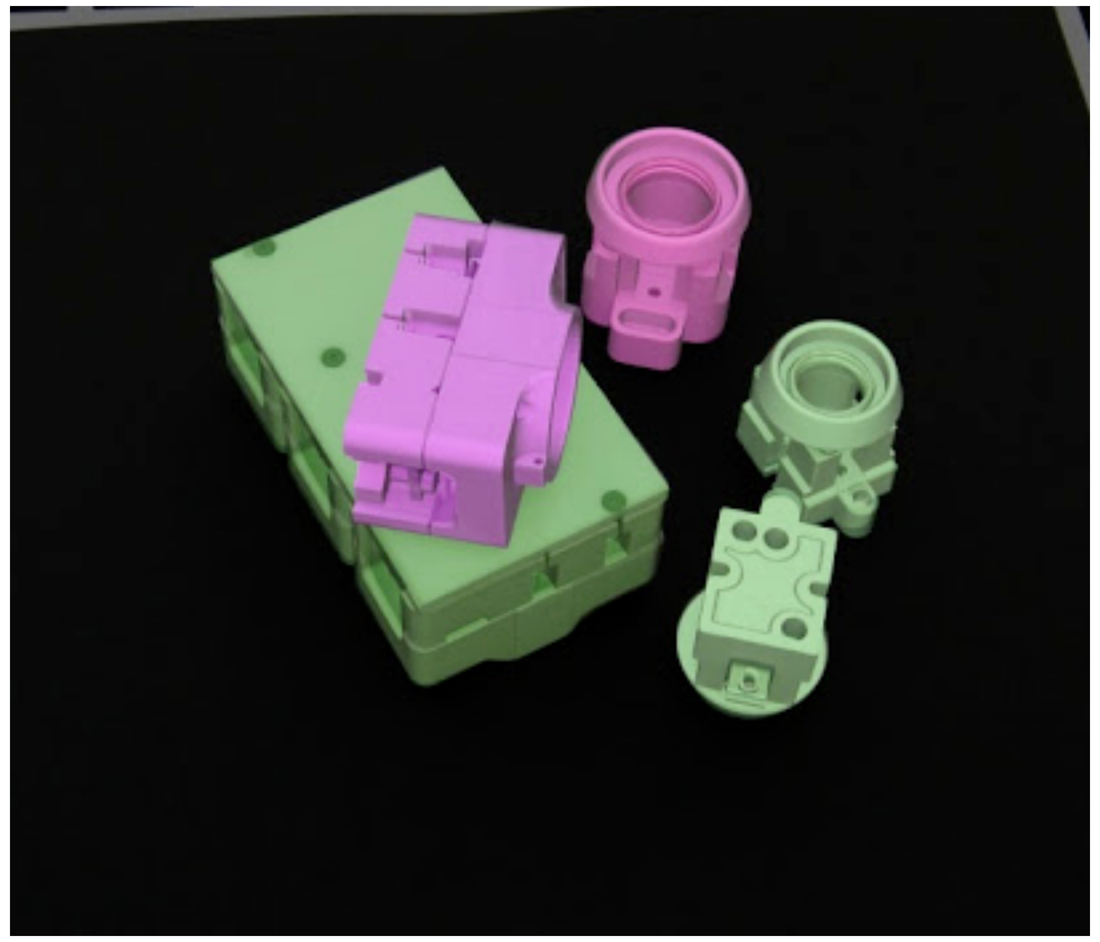
Limitations of this Approach

- **Error propagation:** object pose estimation under heavy cluster is still hard! Vision error will propagate to planning and result in failure execution.

- **Bad generalization:** Need 3D models of the objects during training, therefore hard to generalize to unseen objects.

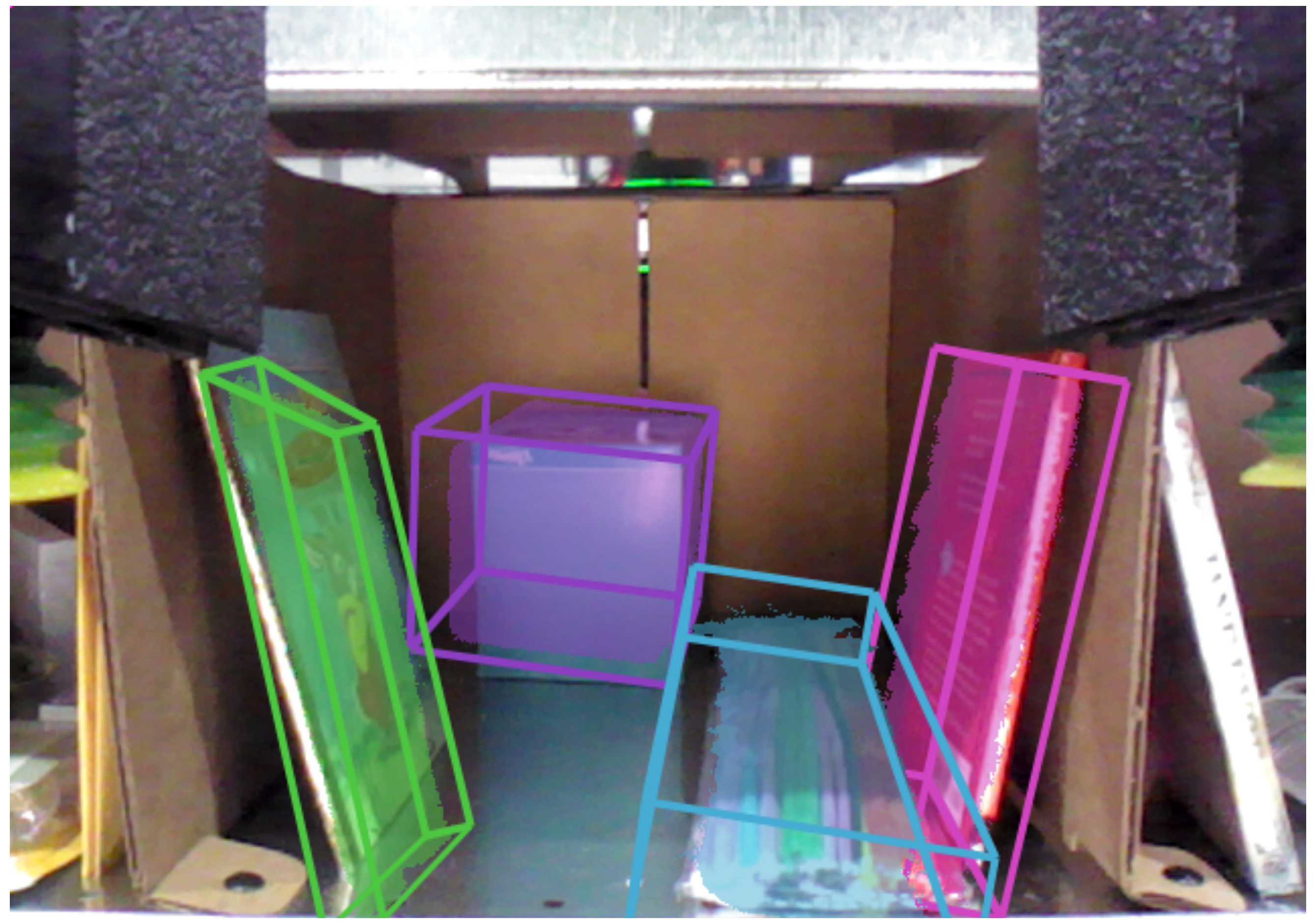
Generalizable Manipulation

Goal: Manipulation algorithm is able to generalize to new objects without the need of strong prior knowledge about the object, such as their 3D CAD model, predefined category, and poses.



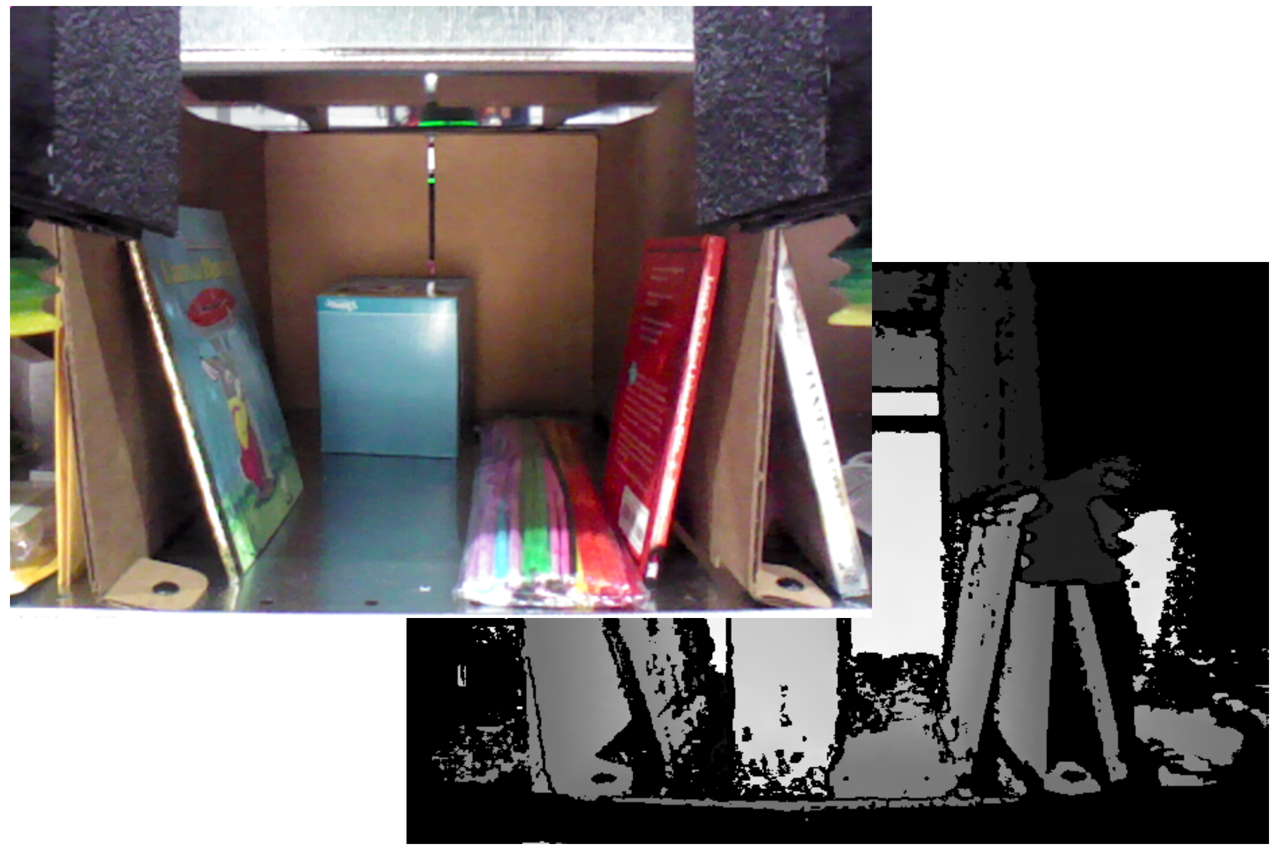
CAD Model

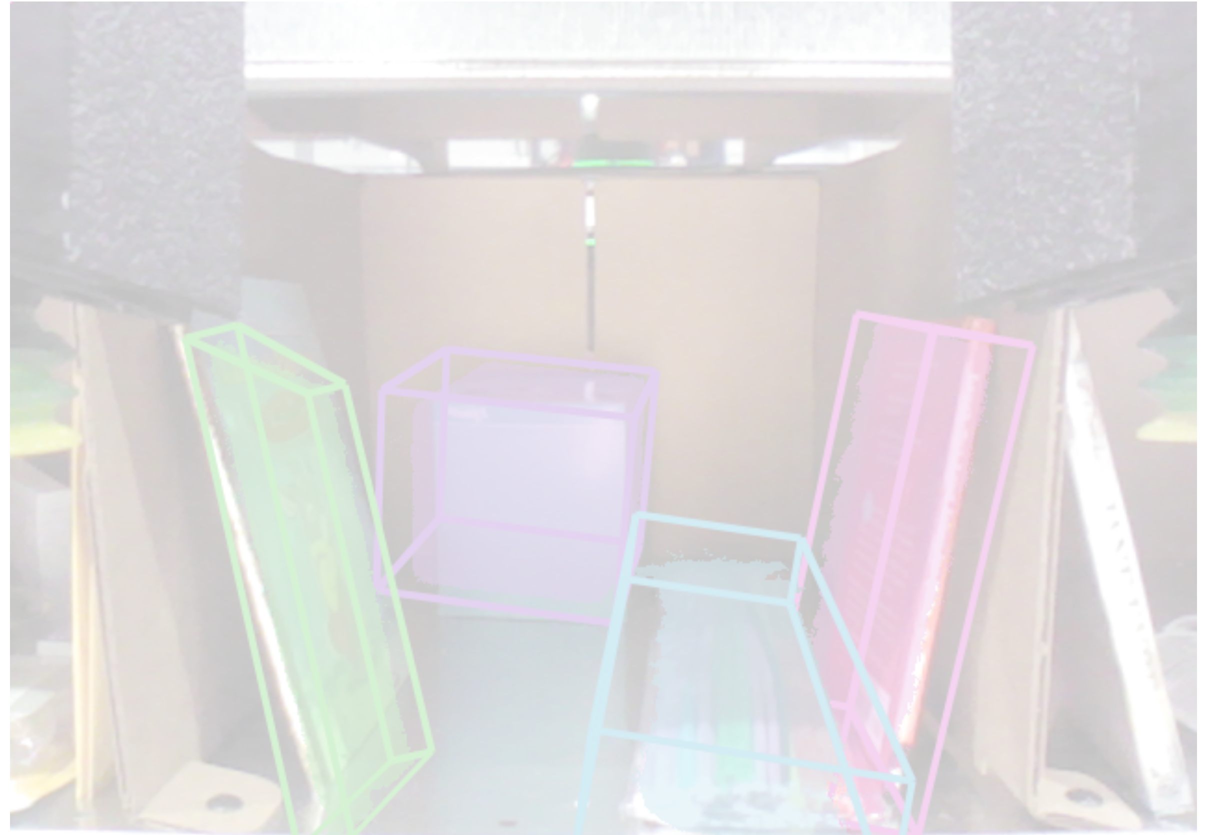
Object Category



Object Poses

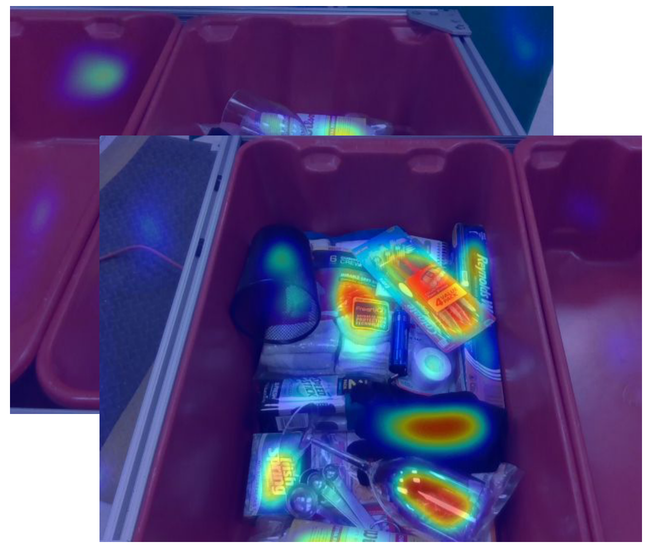
Generalizable Grasp Planning

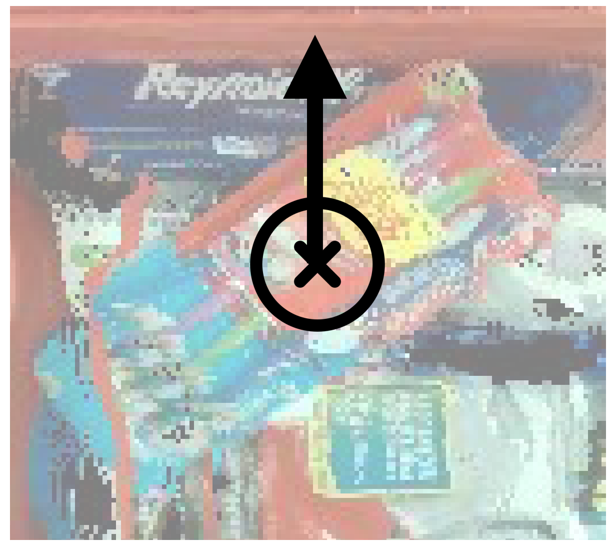


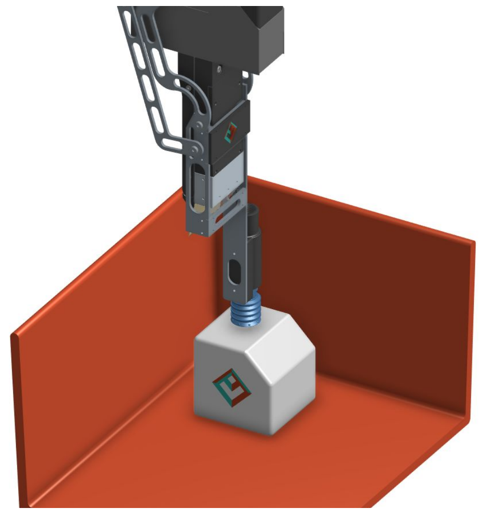


Action Affordance

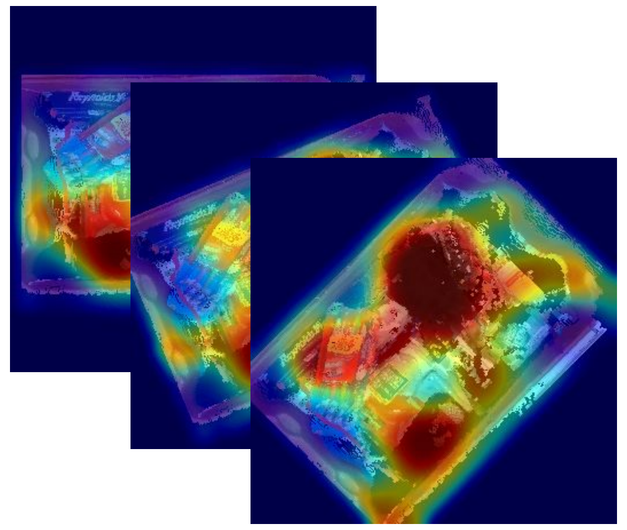
Suction

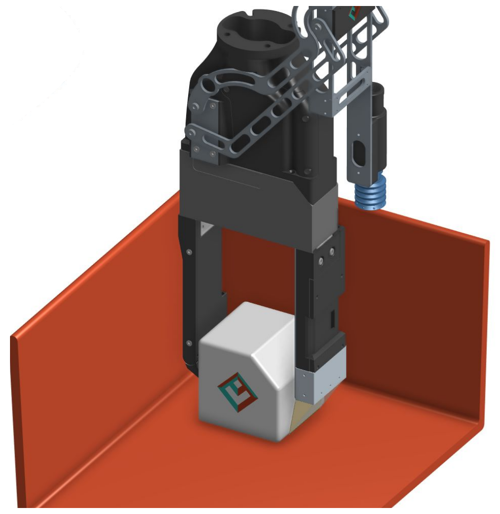






Grasping





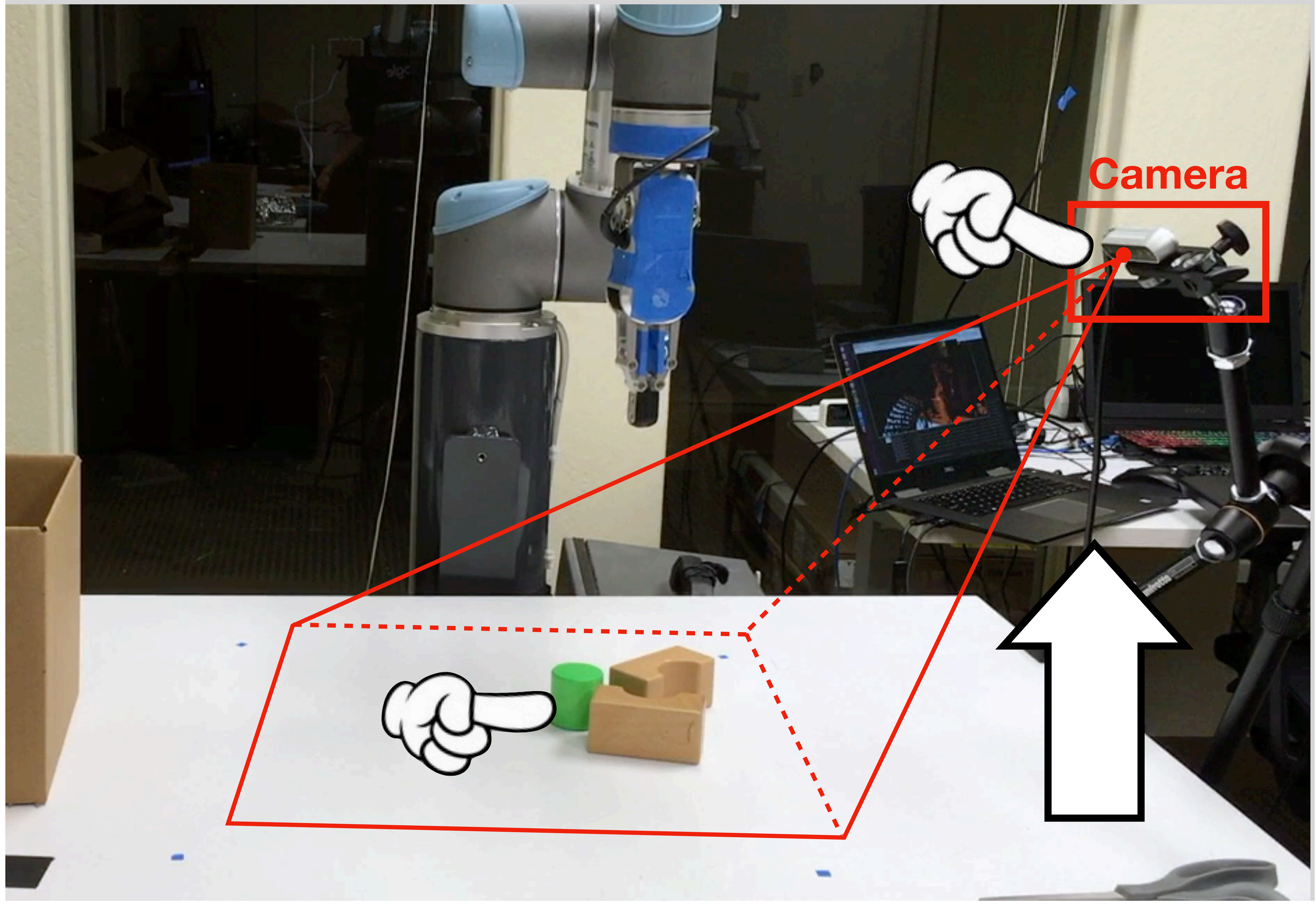
Recognize
Isolated the Object

Amazon Robotics Challenge 2017

Amazon Robotics Challenge 2017

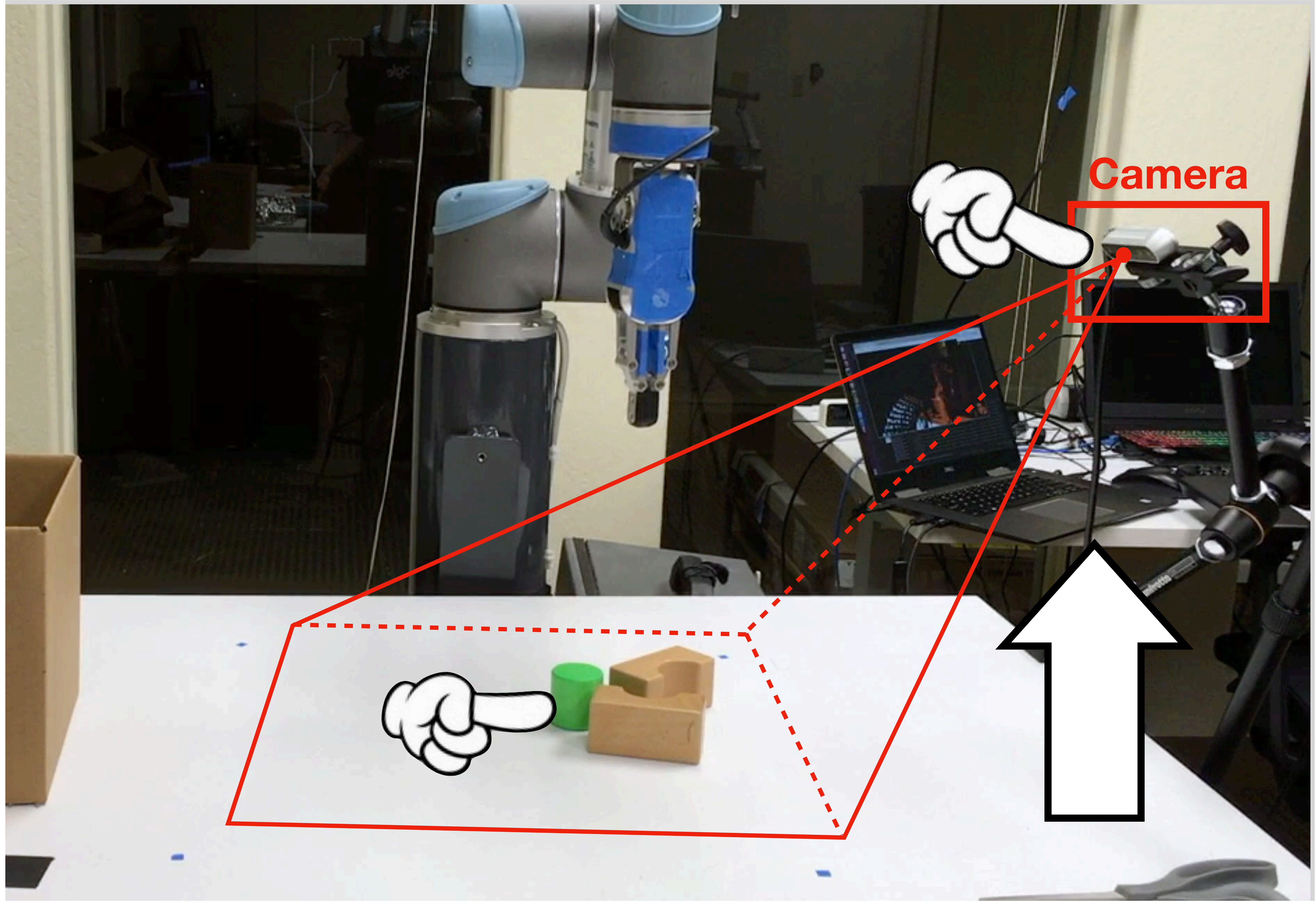
Is Grasping Problem Solved?

Is Grasping Problem Solved?

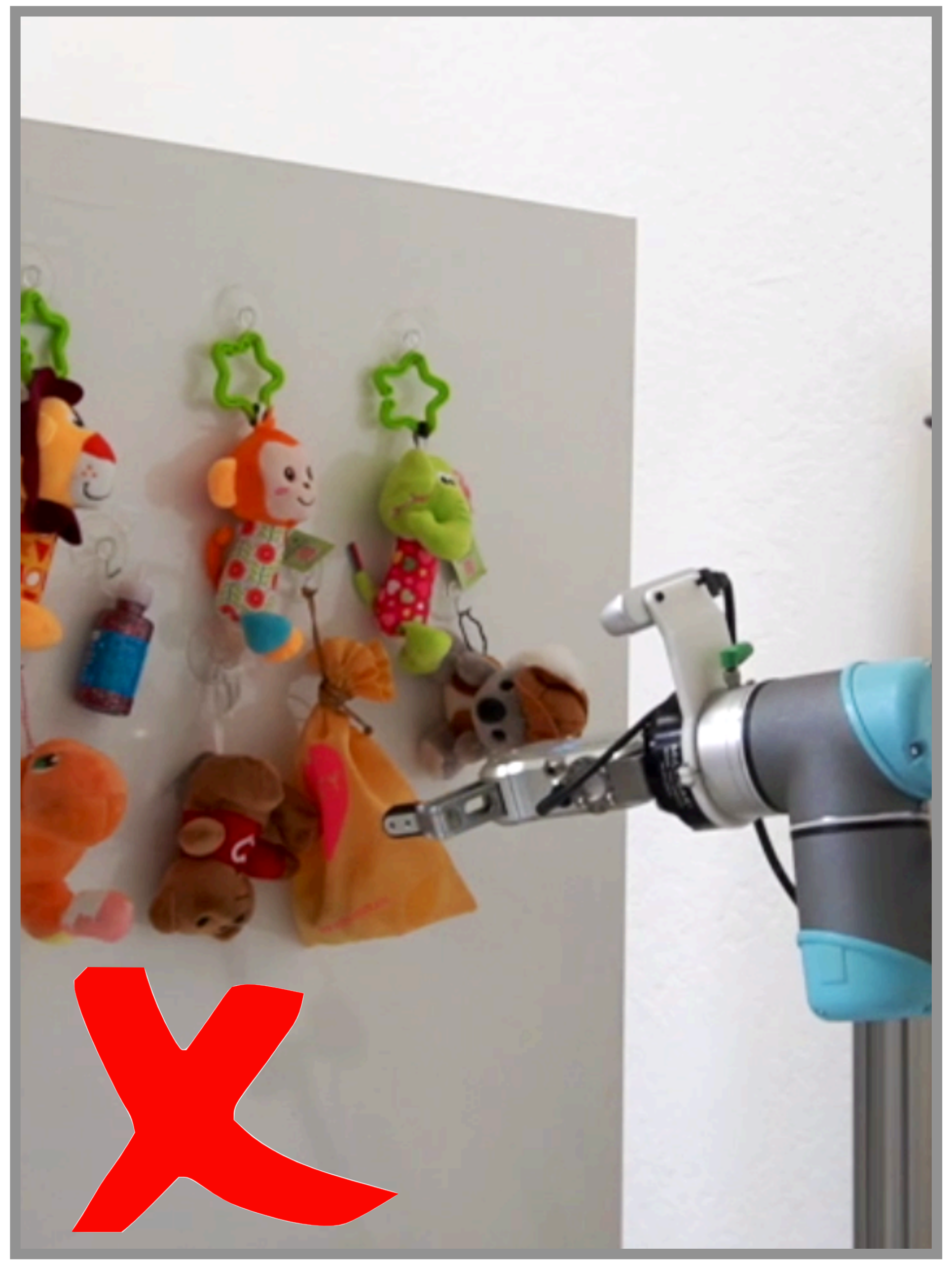


5 DIFFERENT WAYS TO BREAK IT!

Is Grasping Problem Solved?



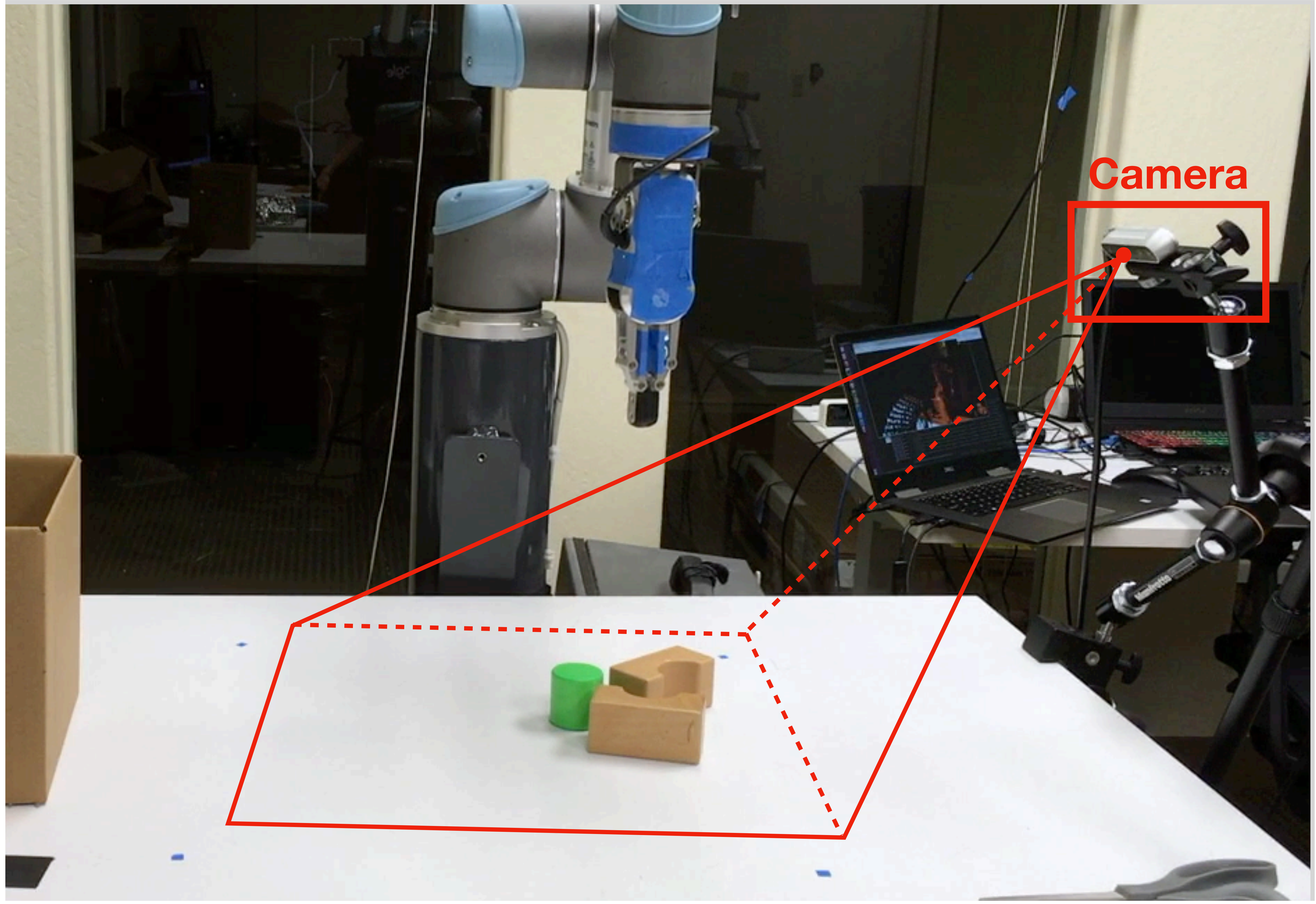
5 DIFFERENT WAYS TO BREAK IT!



✗ Reactive
Open-loop execution
(static scene only)

✗ Flexible
4-DoF grasps
(Top-down grasp only)

Is Grasping Problem Solved?



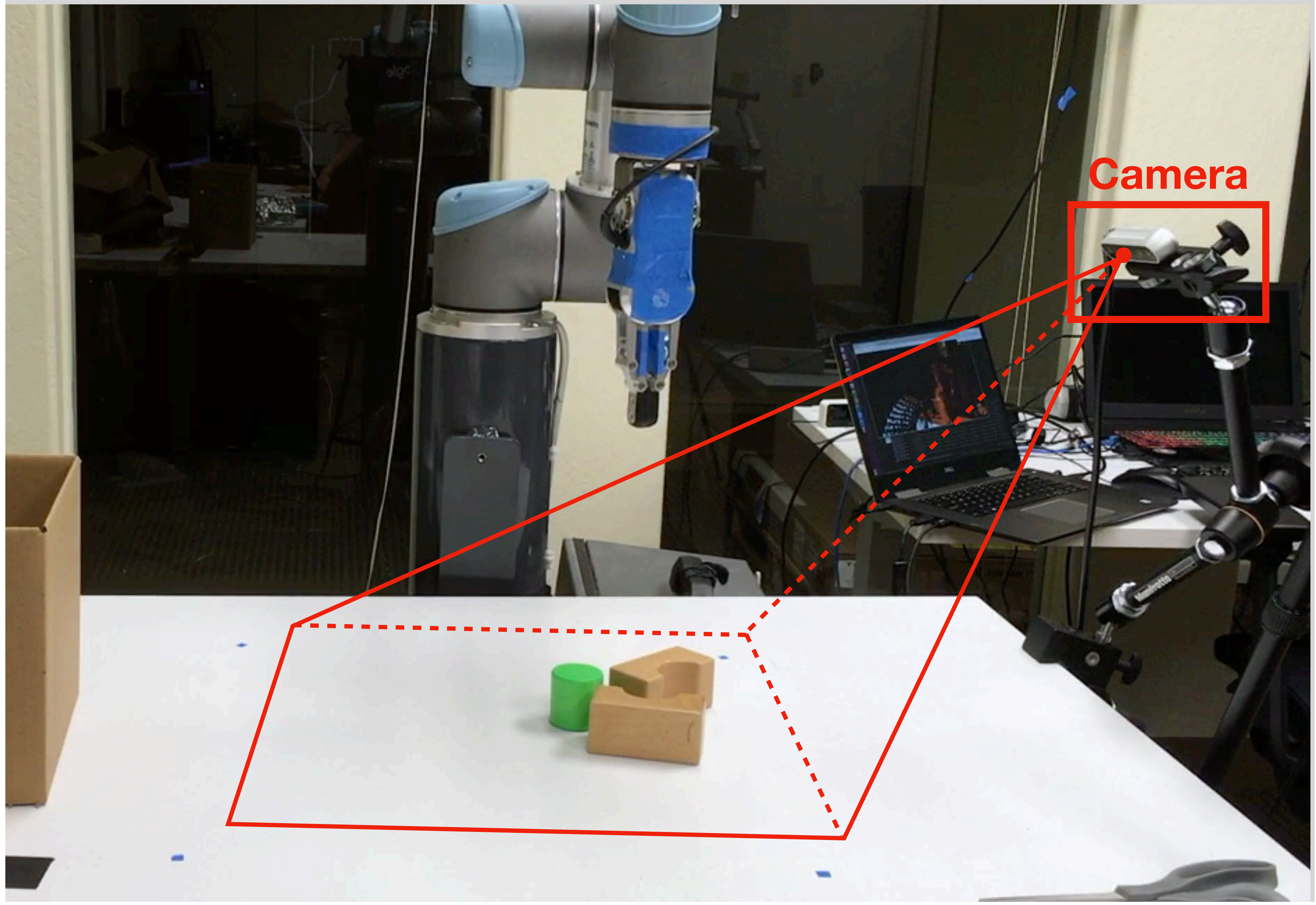
Learning-based Grasping

	Open-loop	Closed-loop
4DoF	Mahler et al. ICRA. 2016 Mahler et al. RSS,2017 Pinto and Gupta, ICRA 2016 Zeng et al. IROS 2018 Zeng et al, ICRA 2018 Zeng et al ICRA 2017	Morrison et al. RSS, 2018. Kalashnikov et al, CORL 2018 Redmon and Angelova ICRA, 2015 Viereck et al CORL 2017 Levine et al IJRR 2018
6DoF	Gualtieri et al IROS 2016 Goldfeder et al 2009 Gualtieri and Platt ICRA 2018 Weisz and Allen ICRA 2012	??

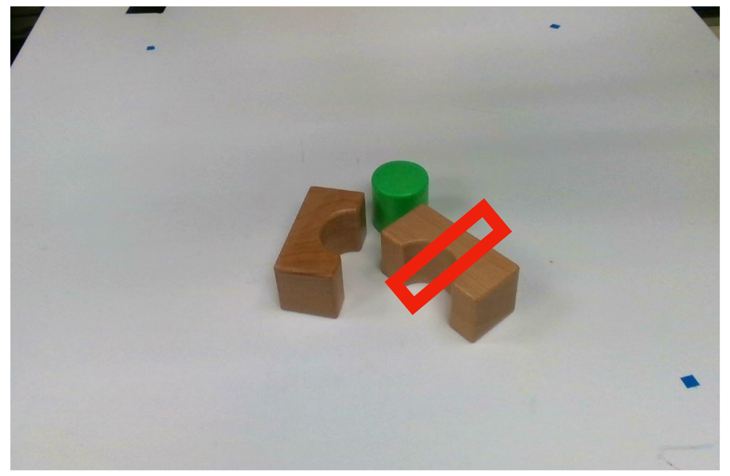
✗ Flexible
4-DoF grasps
(Top-down grasp only)

✗ Reactive
Open-loop execution
(static scene only)

Open-loop Topdown Grasp



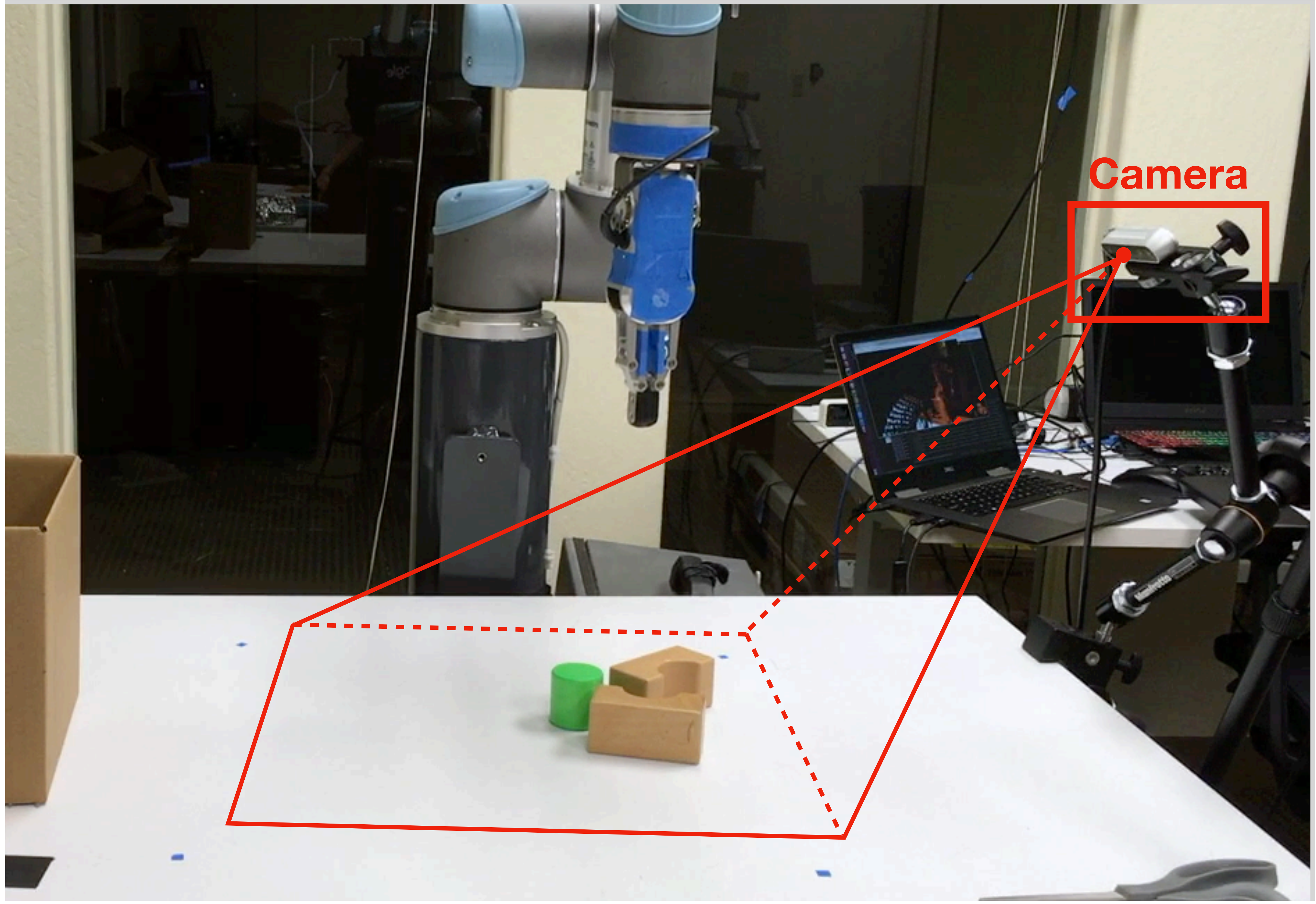
$$f(\text{state}) \rightarrow \text{action}$$



Where + How to grasp?

Open-loop Topdown Grasping

Closed-loop 6DoF Grasp



$f(\text{state}) \rightarrow \text{action}$

Closed-loop:
large state space

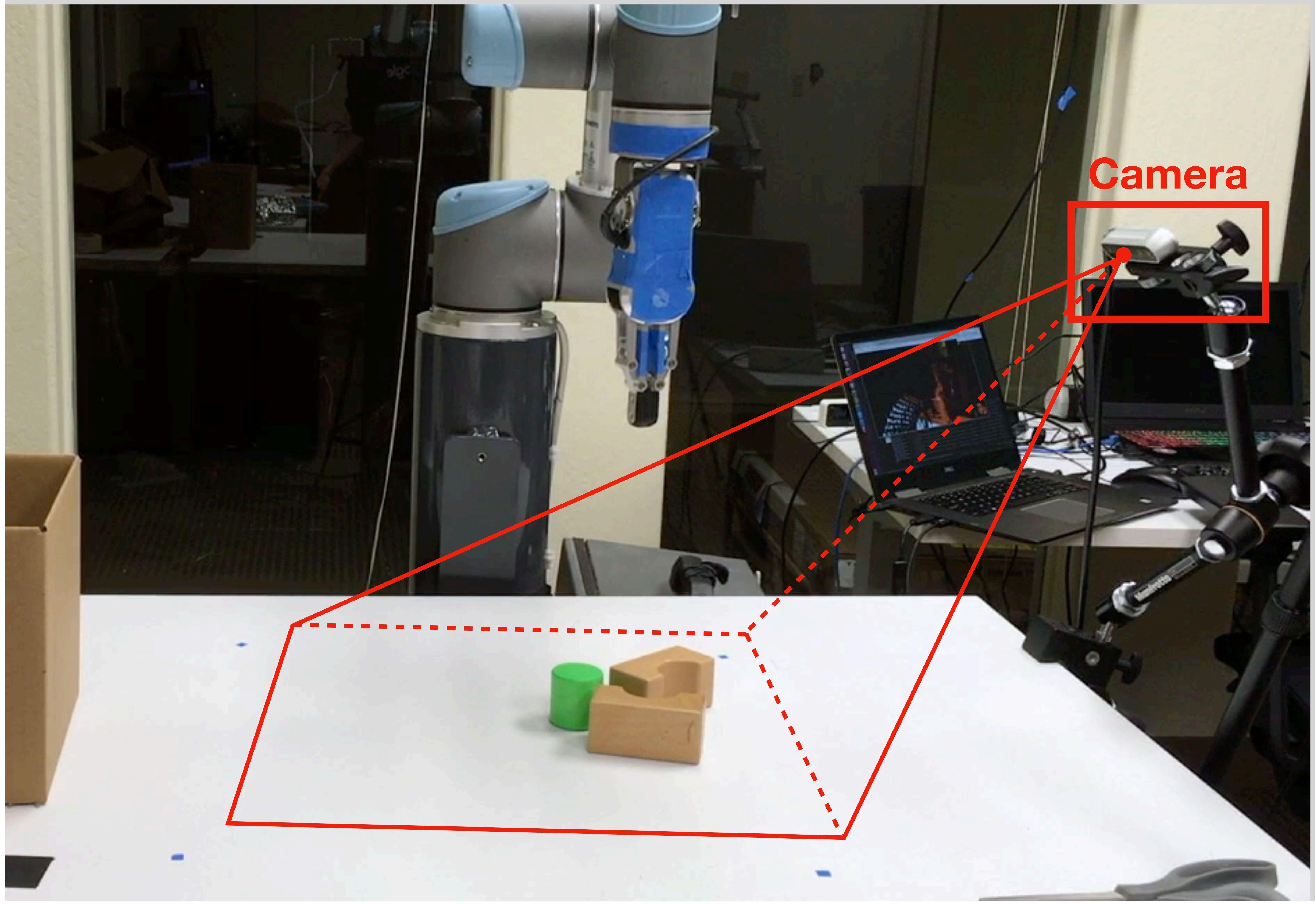
6DoF:
large action space

Qt-Opt: Scalable deep reinforcement learning for vision-based robotic manipulation.
Kalashnikov et al CORL, 2018.

Learning based Closed-loop (Topdown)

580,000 off policy + 28,000 on-policy robot grasping trials

Closed-loop 6DoF Grasp



$f(\text{state}) \rightarrow \text{action}$

Closed-loop:
large state space

6DoF:
large action space

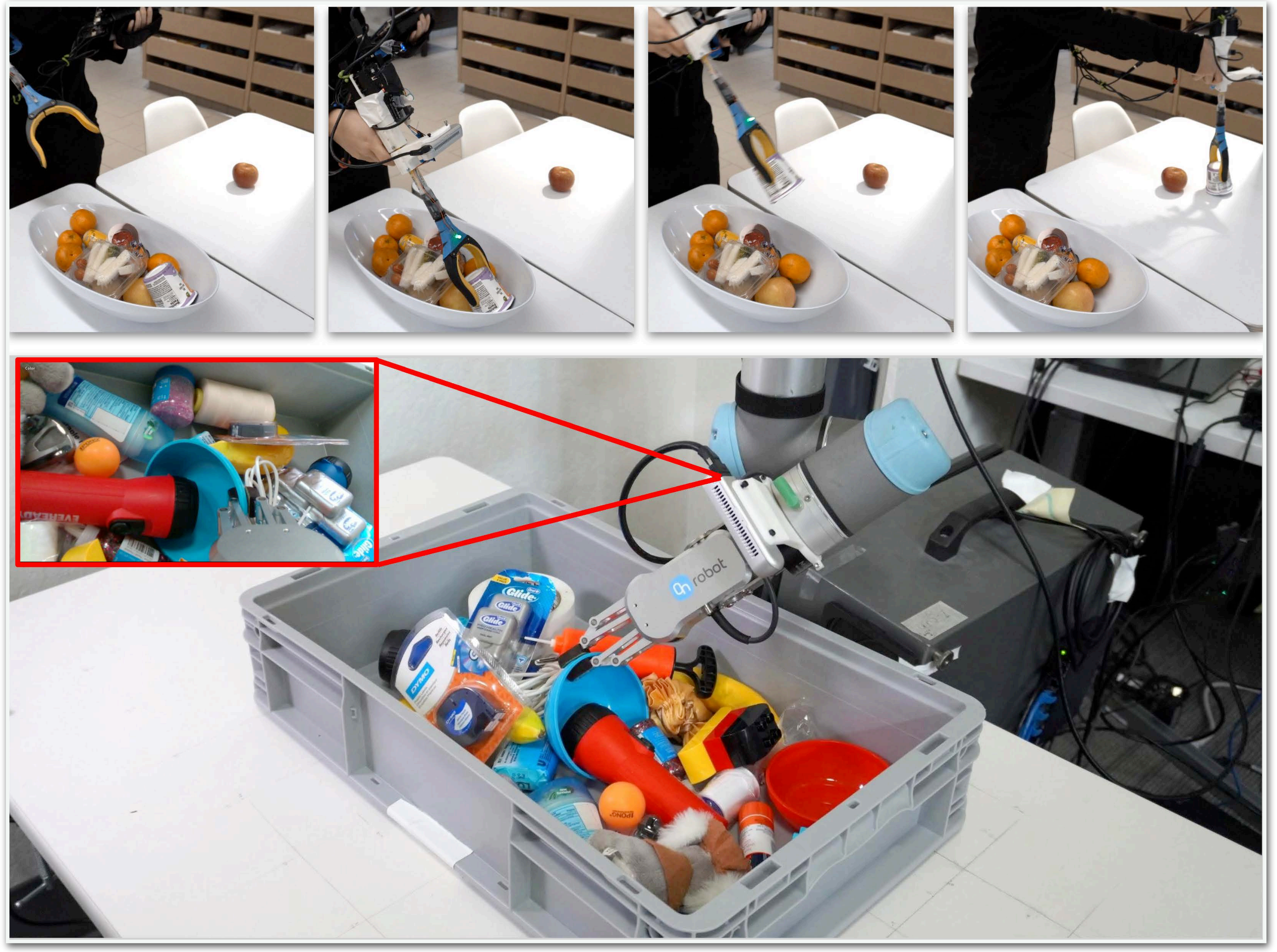
How to get the training data?

How to enable efficient learning?

Closed-loop 6DoF Grasp

Grasping In the Wild:
Learning Flexible Grasping Policy
with Low-cost Demonstration

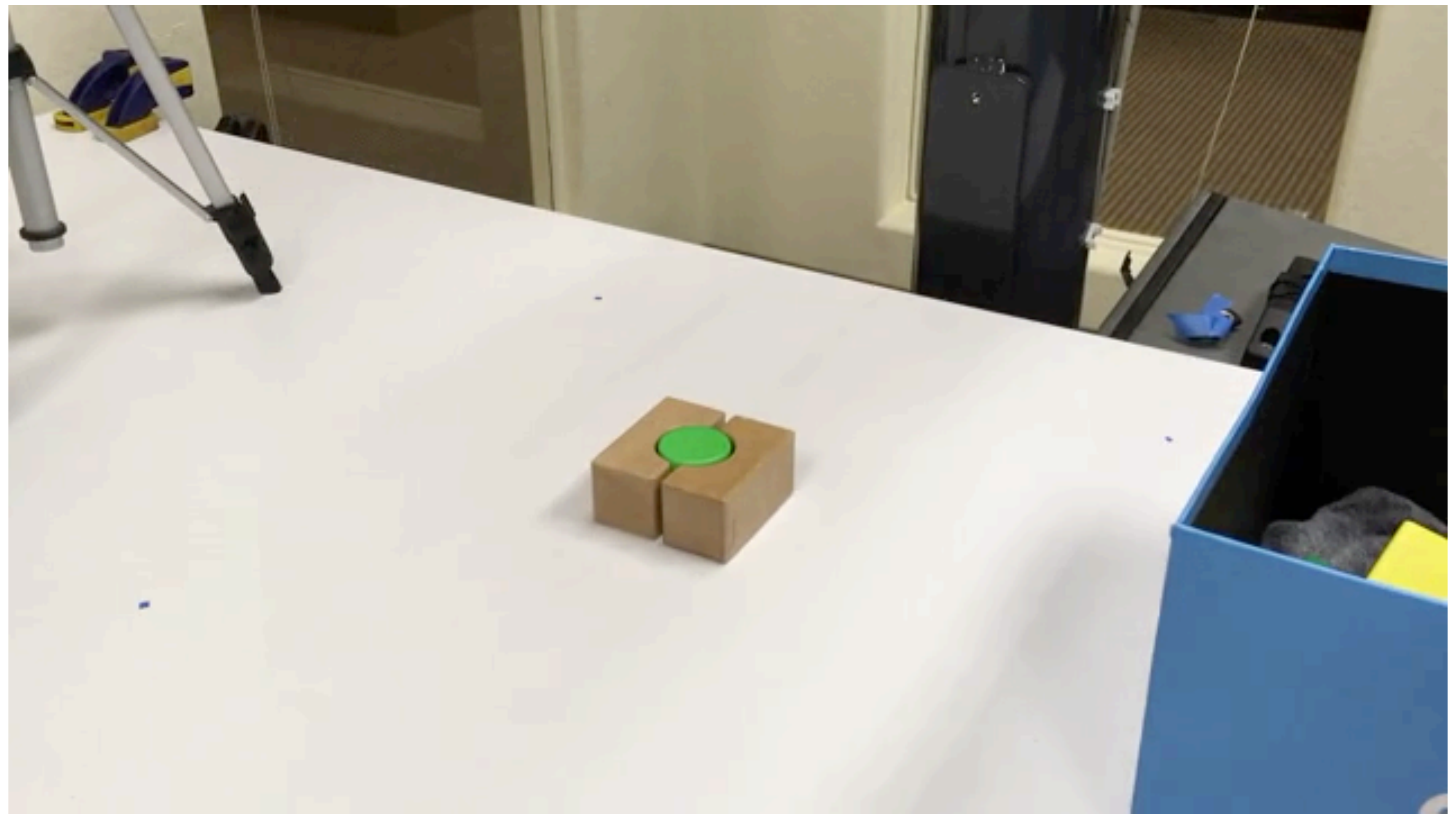
Shuran Song, Andy Zeng, Johnny Lee, Thomas Funkhouser
RA-L, IROS 2020



The Data Problem

The Data Problem

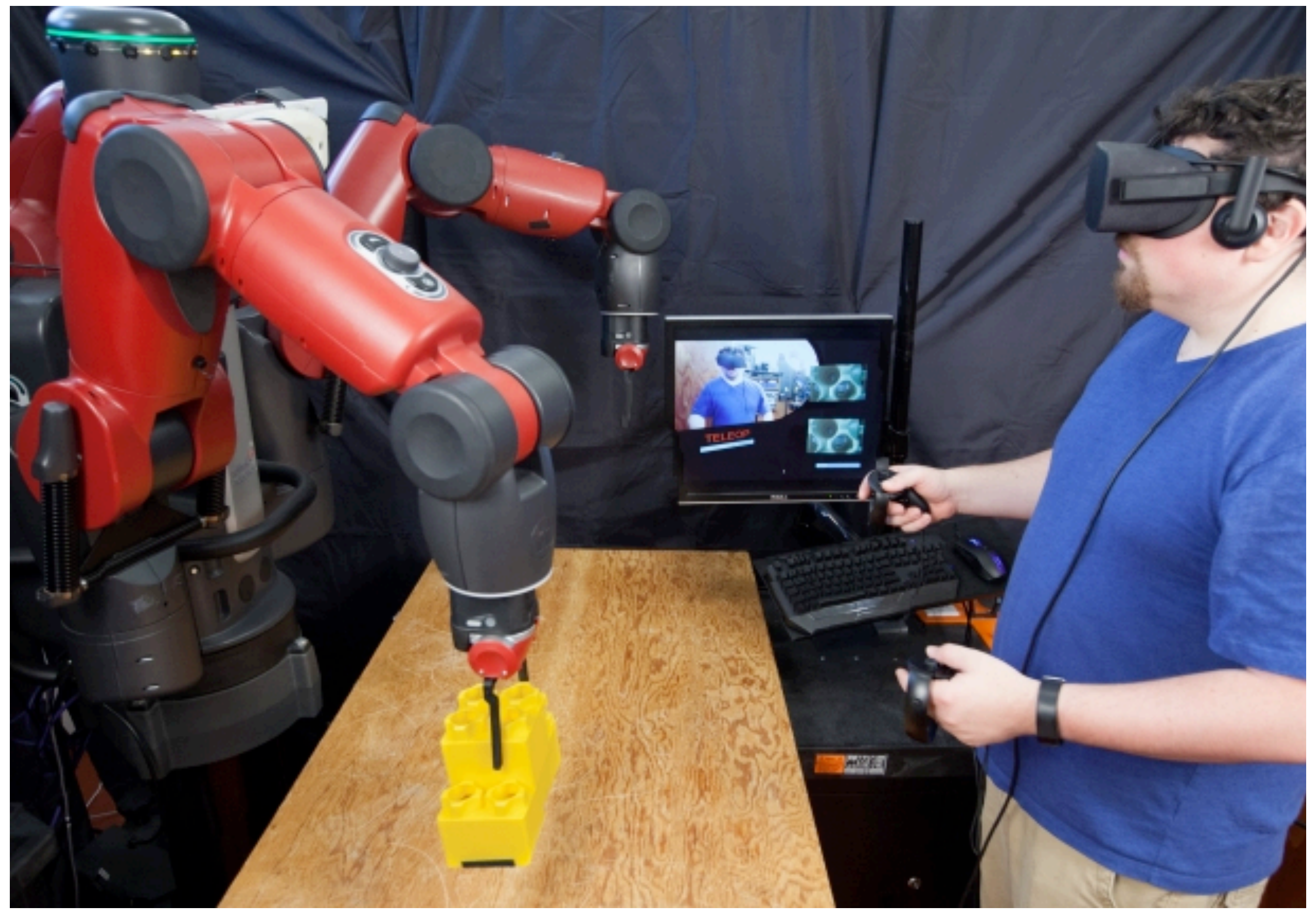
Self-supervised learning



Zeng et al IROS’10

✗ Simple scenarios: low success rate, hard to get initial positive training data.

Learning from demonstration



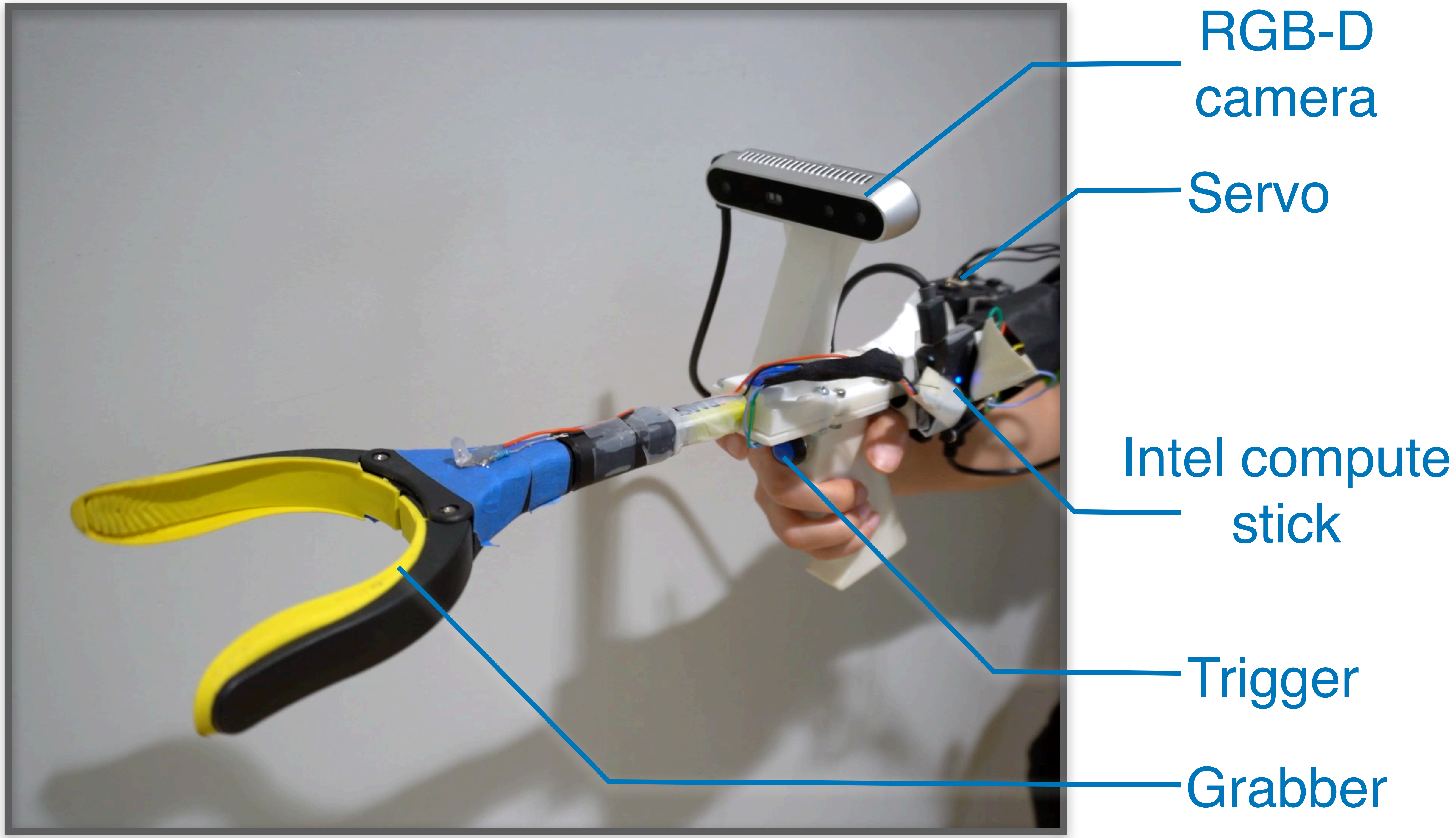
Schneider et al IROS’10

✗ Expensive setup, Limited physical access (robots)

✗ Expert operator

✗ Hard to scale

The Data Problem



Data collection device

The Data Problem

Data collection device

Robot

The Data Problem

Low-friction interface for untrained user:

✓ Collect data everywhere.
(not limited by robot access)

✓ Data for challenging tasks.
(no broken dishes)

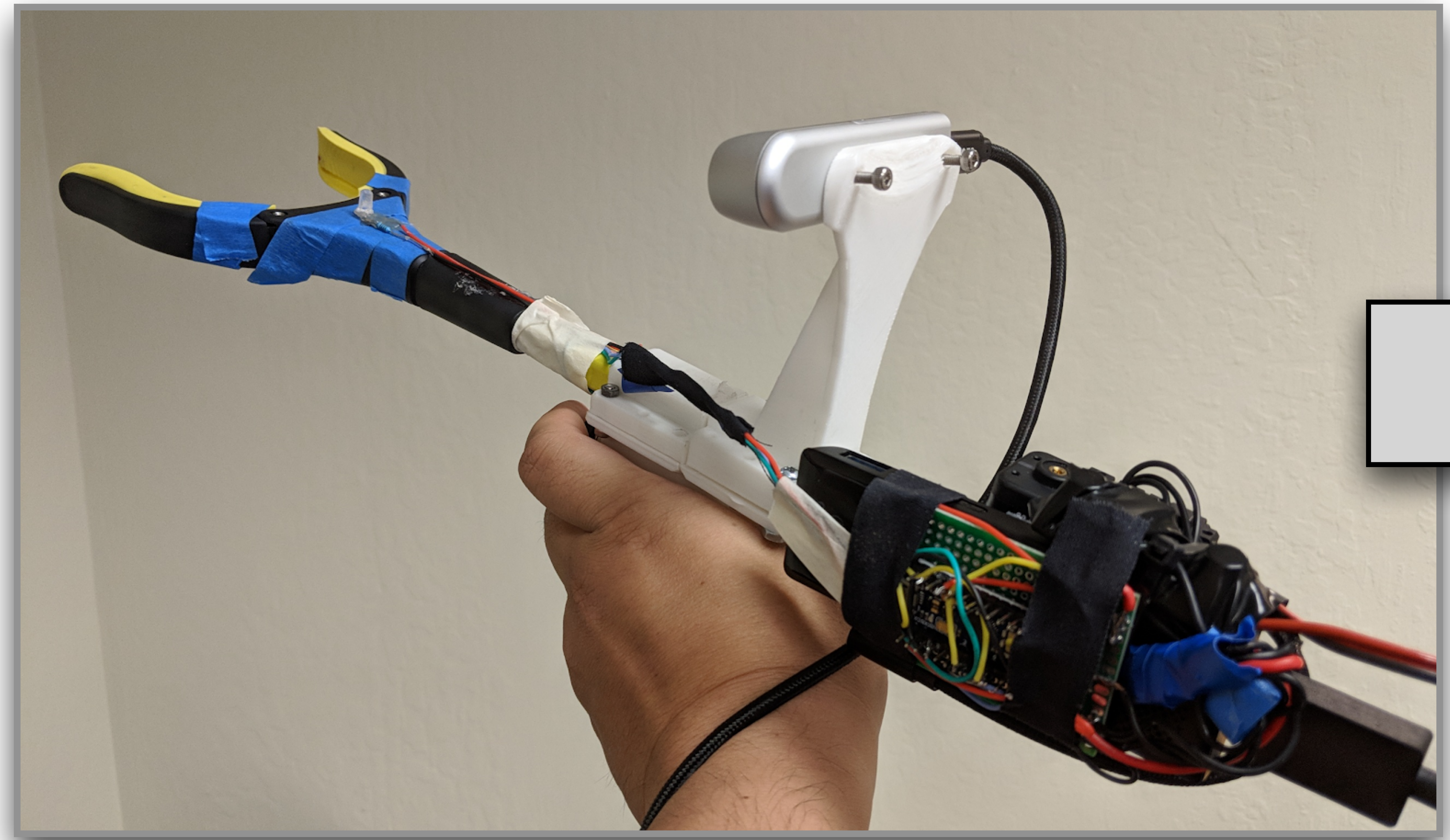
✓ Minimized domain gap.

Data collection device

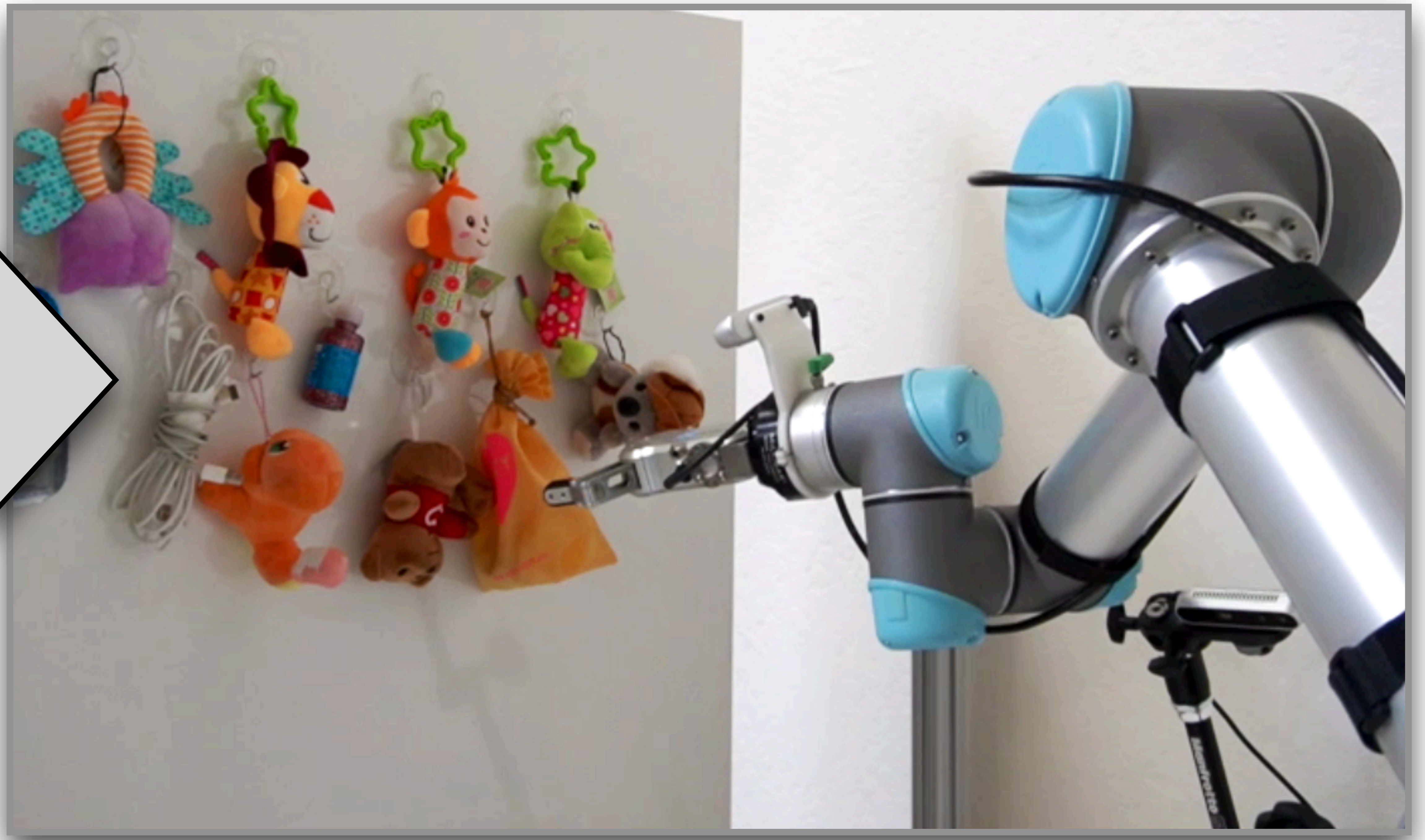
Human demonstrations


Grasping In-the-Wild
Approaching
Step2Grasp: 034
Approaching
Step2Grasp: 010
Approaching
Step2Grasp: 025
Approaching
Step2Grasp: 042
Approaching
Step2Grasp: 044
Approaching
Step2Grasp: 044
Approaching
Step2Grasp: 033
Approaching
Step2Grasp: 044
Approaching
Step2Grasp: 013
Approaching
Step2Grasp: 044
Approaching
Step2Grasp: 044
Approaching
Step2Grasp: 027
Approaching
Step2Grasp: 006
Approaching
Step2Grasp: 041
Approaching
Step2Grasp: 019
Approaching
Step2Grasp: 044
Approaching
Step2Grasp: 017
Approaching
Step2Grasp: 031
Approaching
Step2Grasp: 020
Approaching
Step2Grasp: 014
Approaching
Step2Grasp: 016
Approaching
Step2Grasp: 015
Approaching
Step2Grasp: 026
Approaching
Step2Grasp: 007
Approaching
Step2Grasp: 014

The Learning Problem



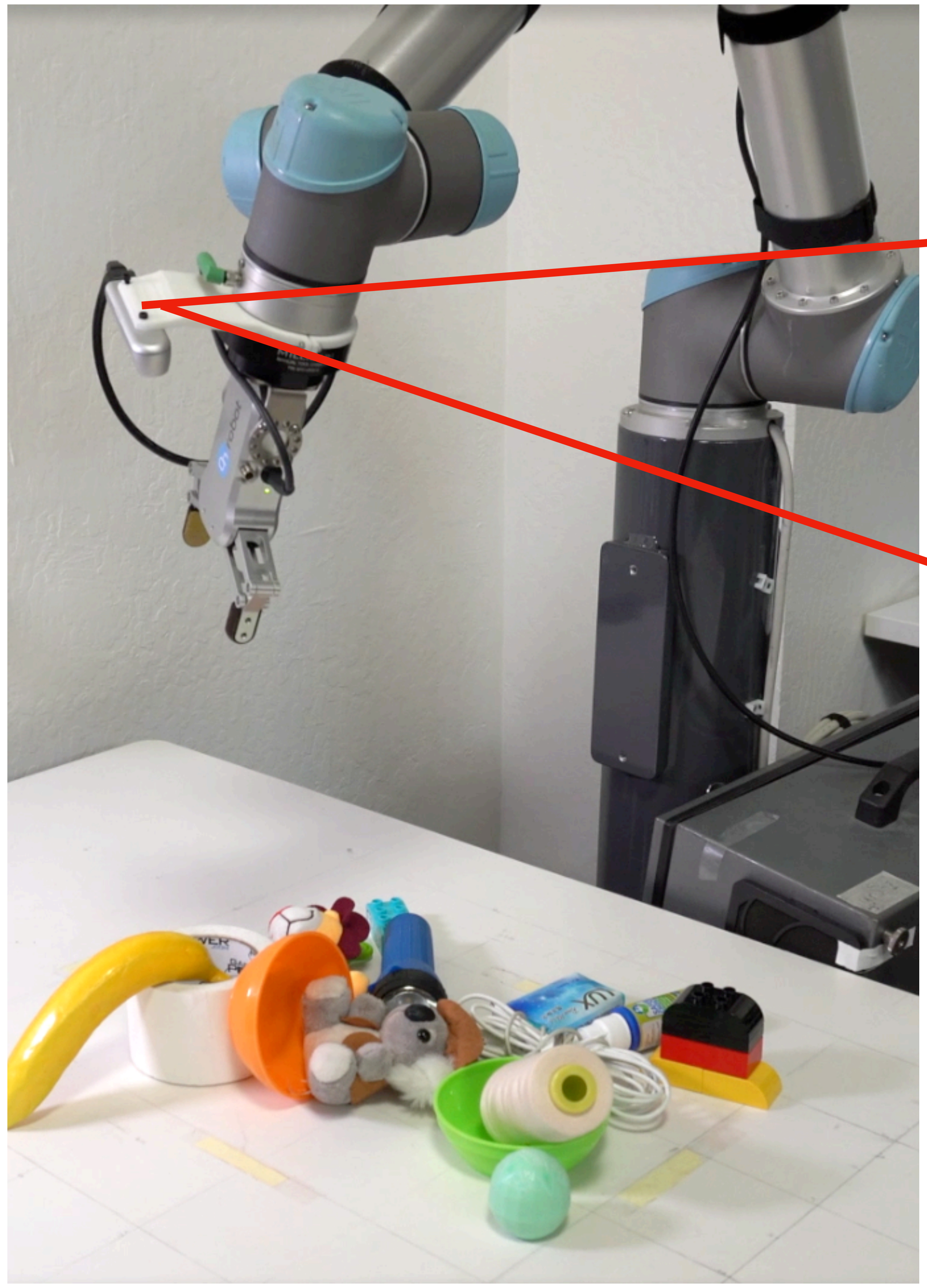
The Data Problem



The Learning Problem

The Learning Problem

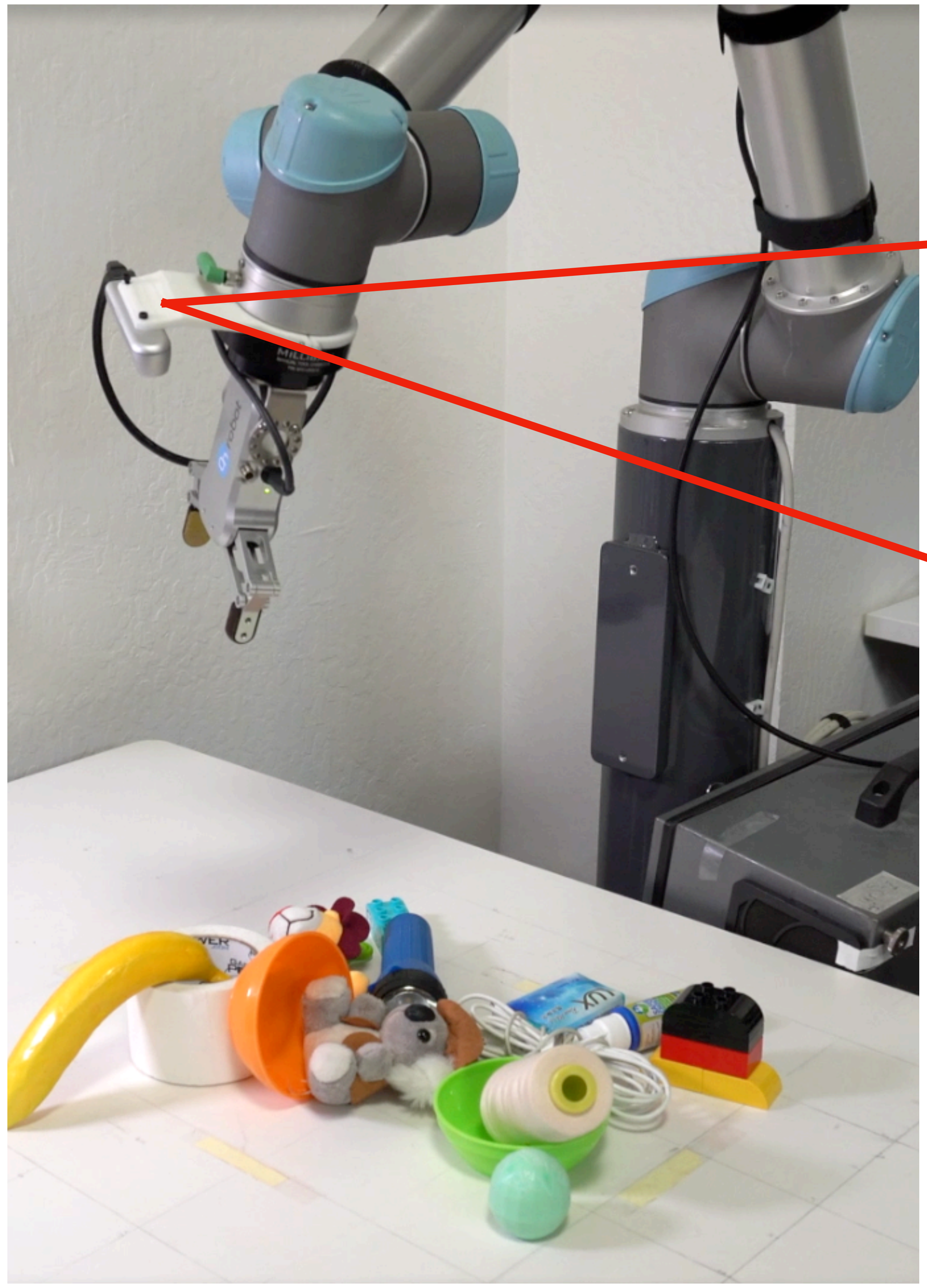
$$f(\text{state}) \rightarrow \text{action}$$



Where to move next?

The Learning Problem

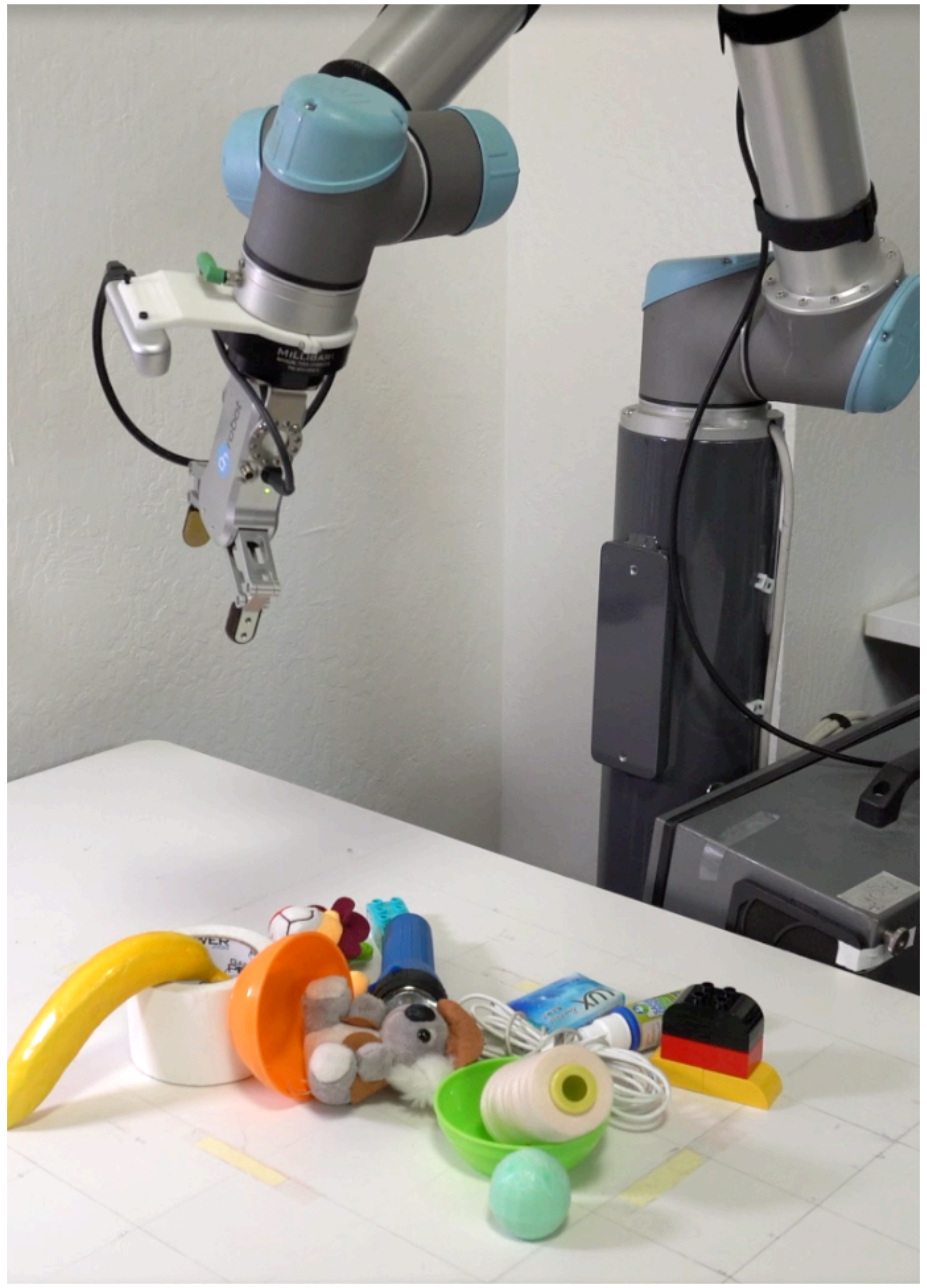
$f(\text{state}) \rightarrow \text{action}$



Where to move next?

How to represent the action?

The Learning Problem



$f(\text{state}) \rightarrow \text{action}$

Where to move next?

How to represent the action?

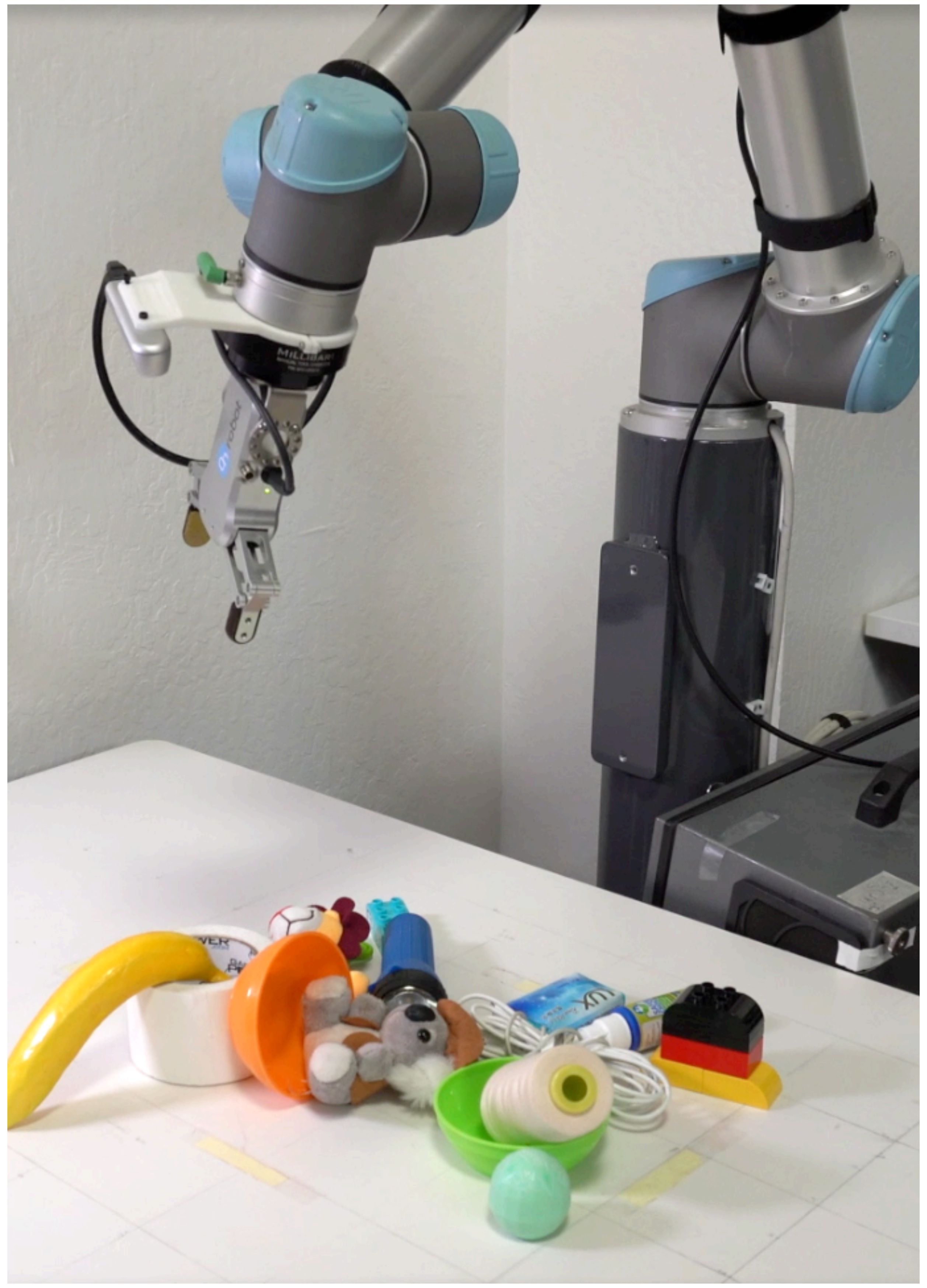
How the action will change the state?

Prior works:

- Joint angles: $[\theta_0\ \theta_1\ \theta_2\ \theta_3\ \theta_4\ \theta_5]$
- Effector offsets: $[d_x\ d_y\ d_z]$
- Motor torques

continuous values that hold abstract meaning

The Learning Problem



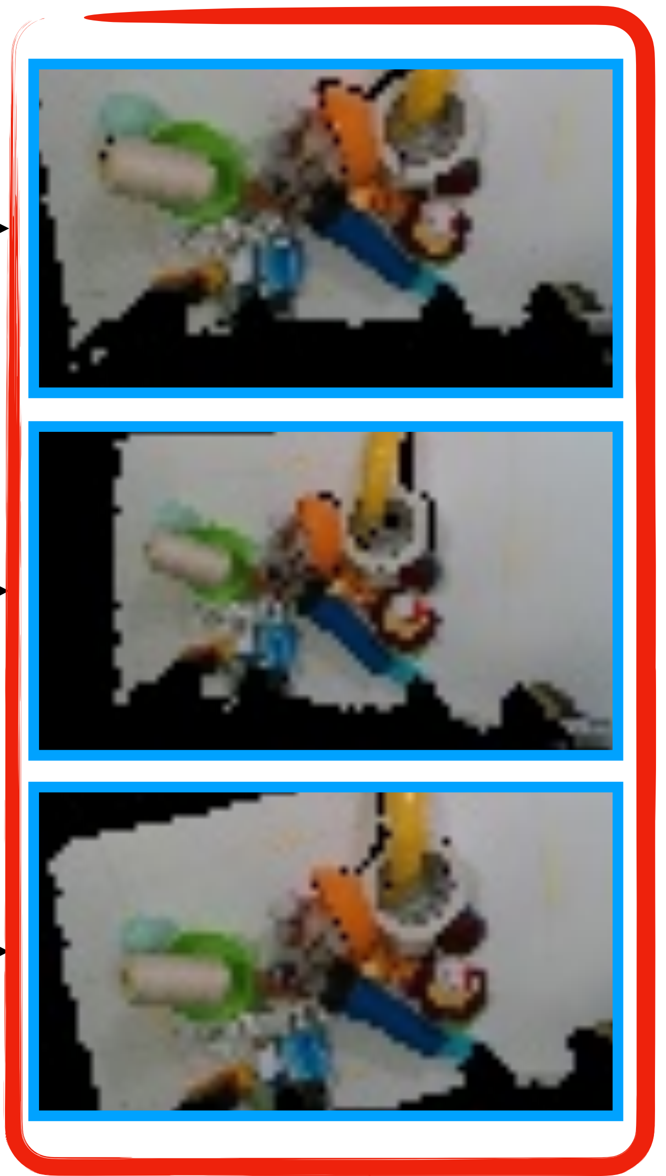
$f(\text{state}) \rightarrow \text{action}$

Prediction
of next state

a_0

a_1

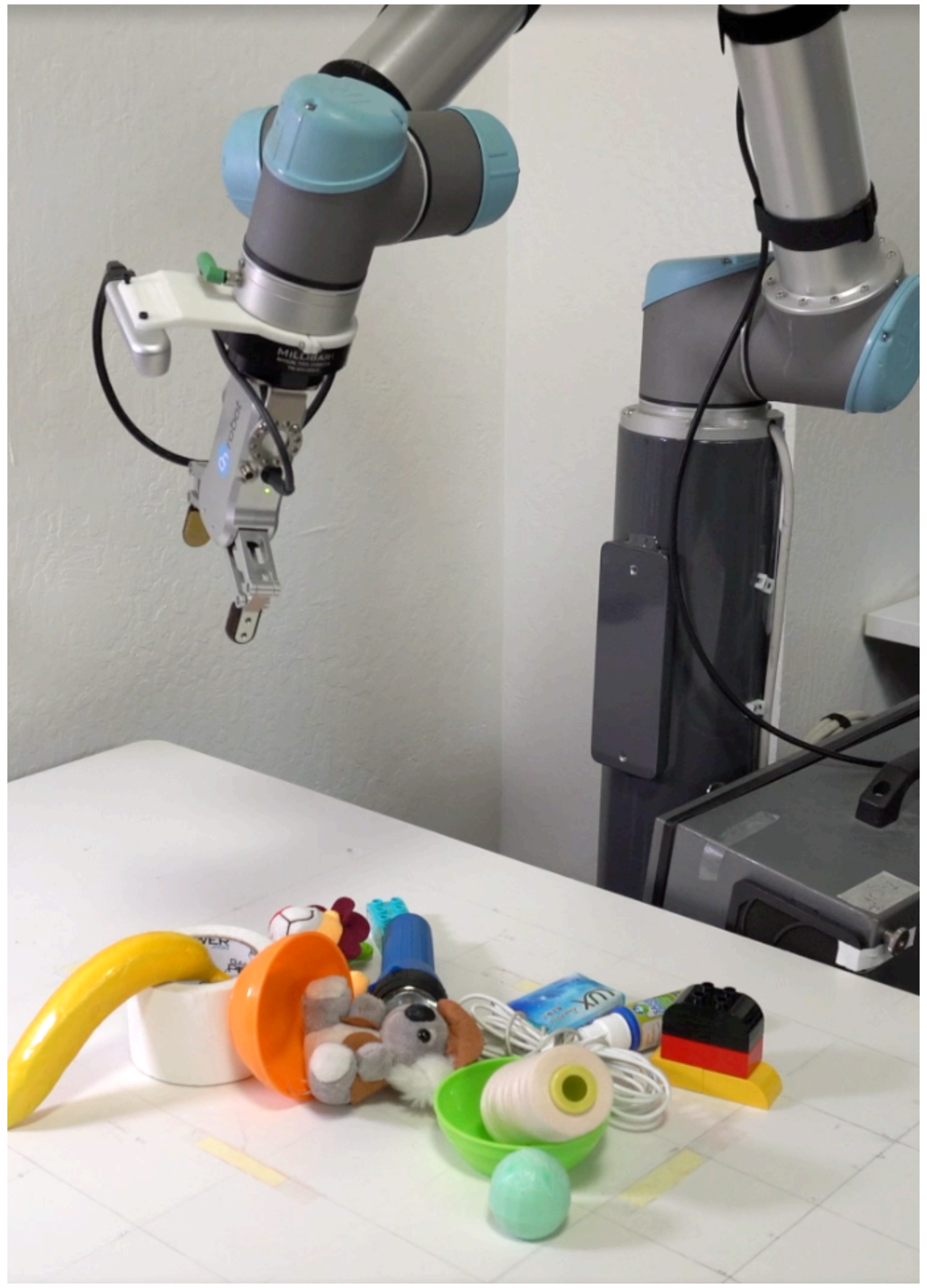
a_2



...

Action-view
representation

The Learning Problem

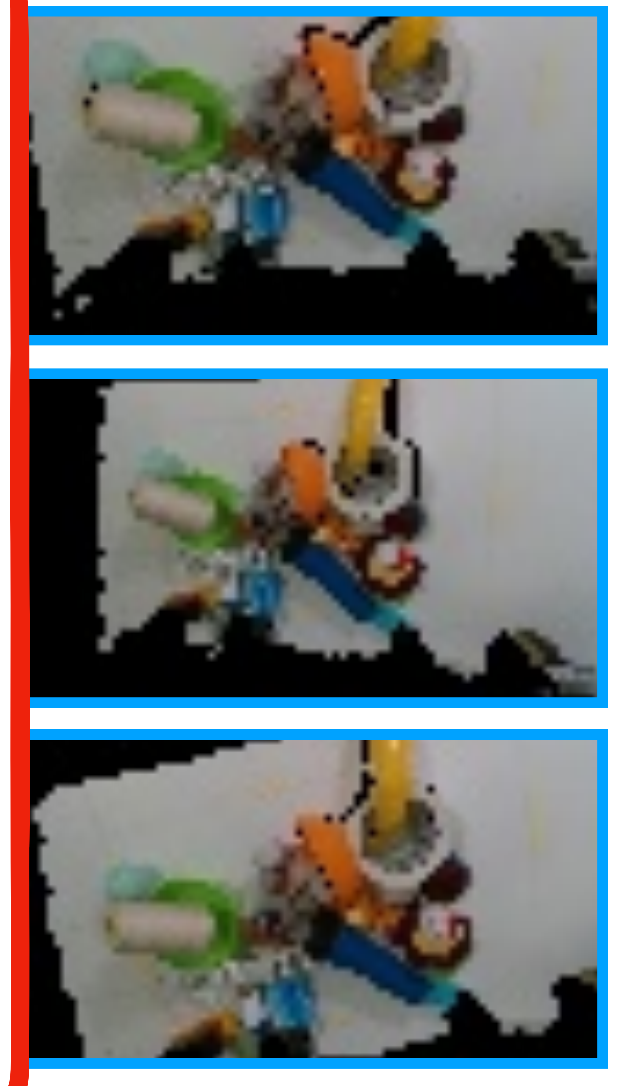


$$f(\text{state}) \rightarrow \text{action}$$

a_0

a_1

a_2

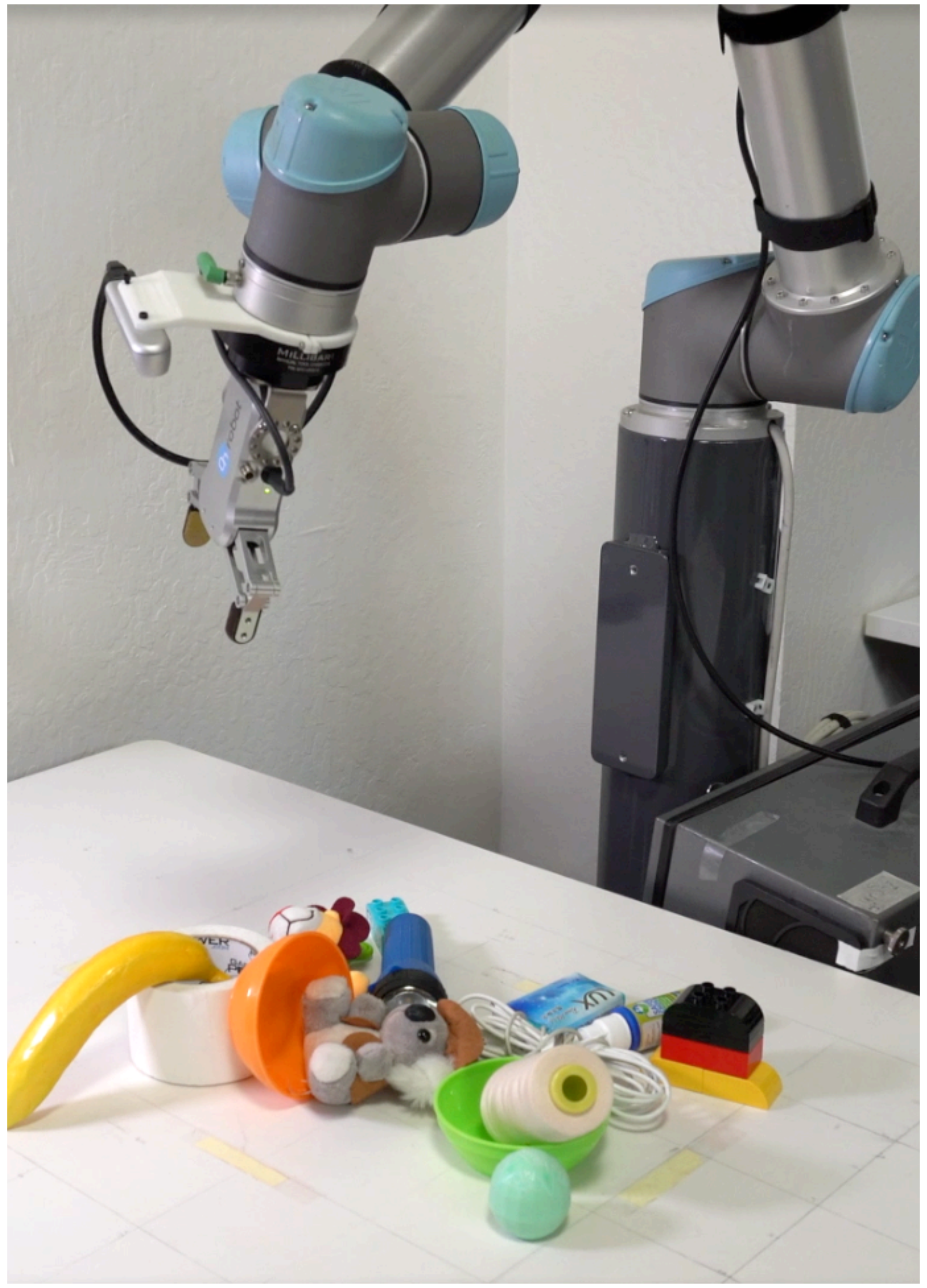


…

6DoF motion
+ camera rigidly mounted

View-based
rendering

Action-view Representation



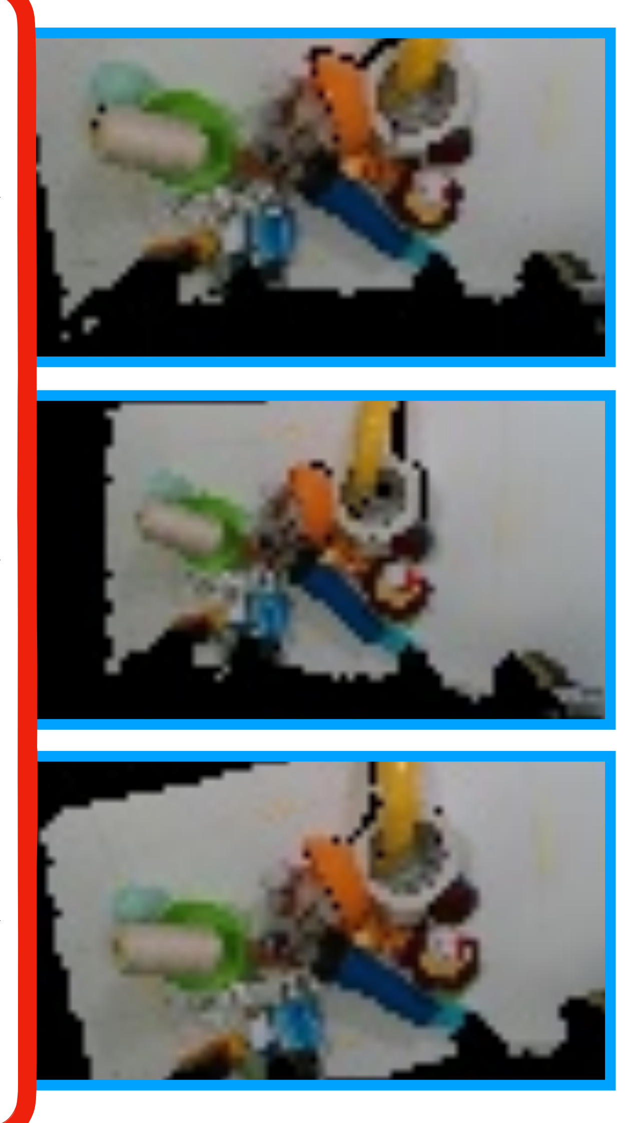
$f(\text{state}) \rightarrow \text{action}$

$a_0 \rightarrow$

$a_1 \rightarrow$

$a_2 \rightarrow$

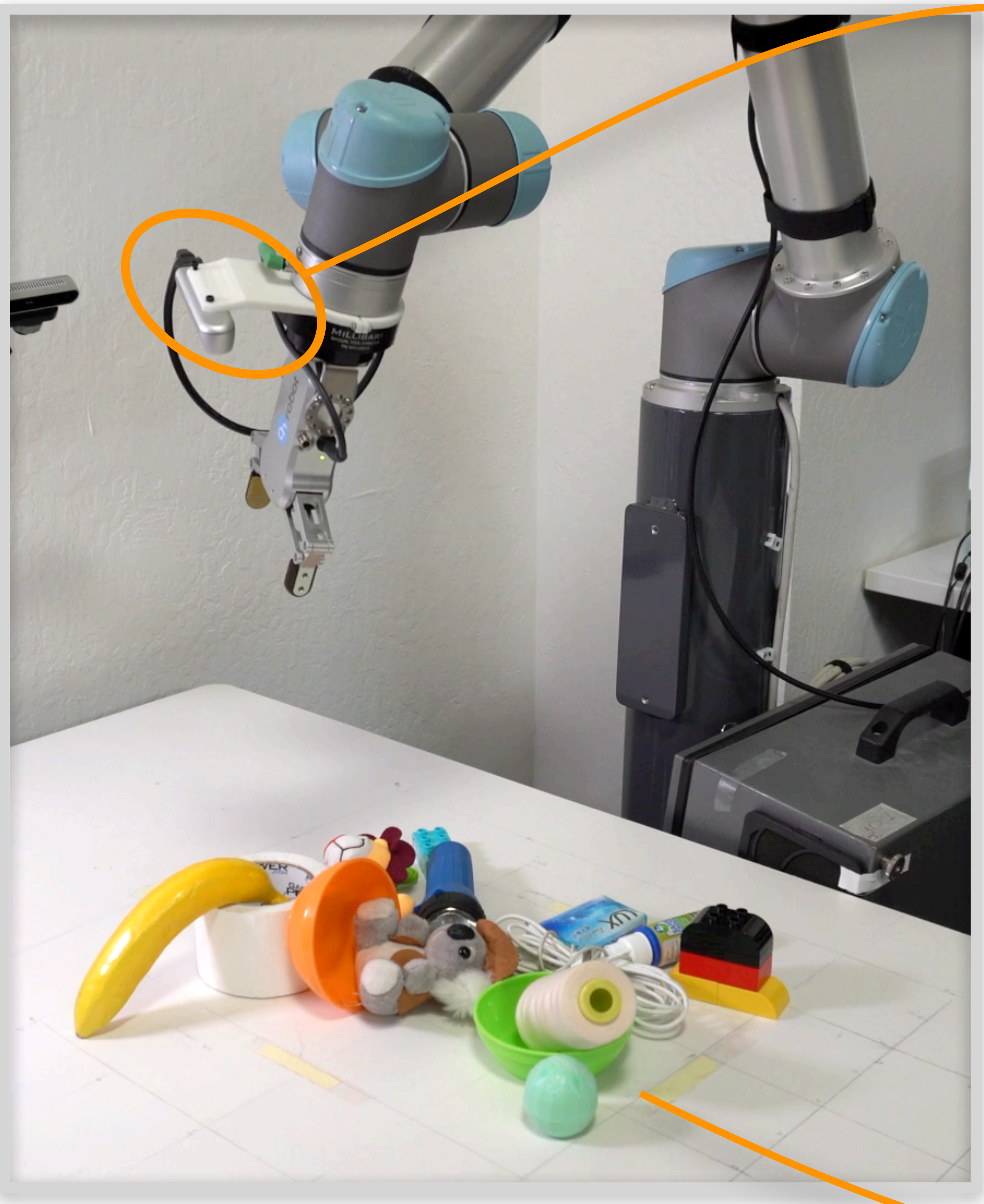
$\cdots$



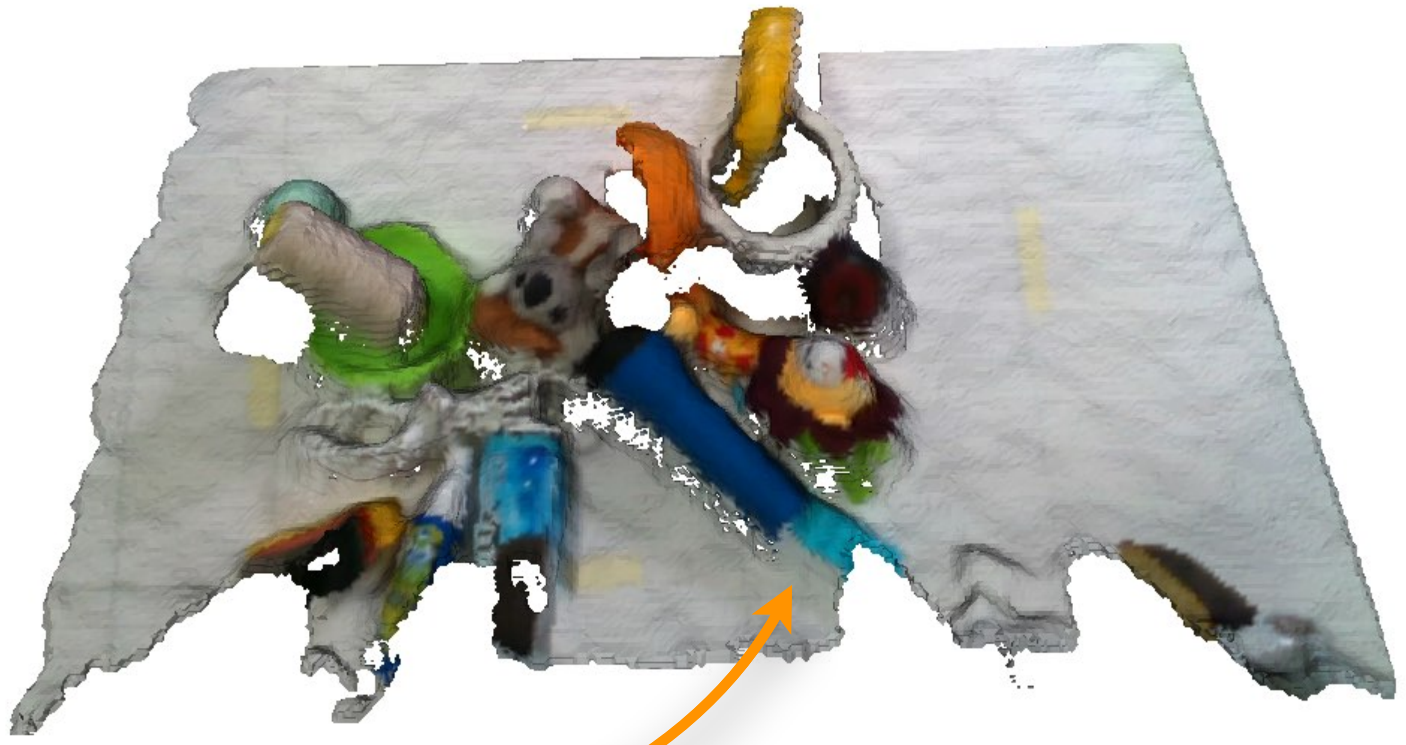
6DoF motion
+ camera rigidly mounted

View-based
rendering

Action-view Representation

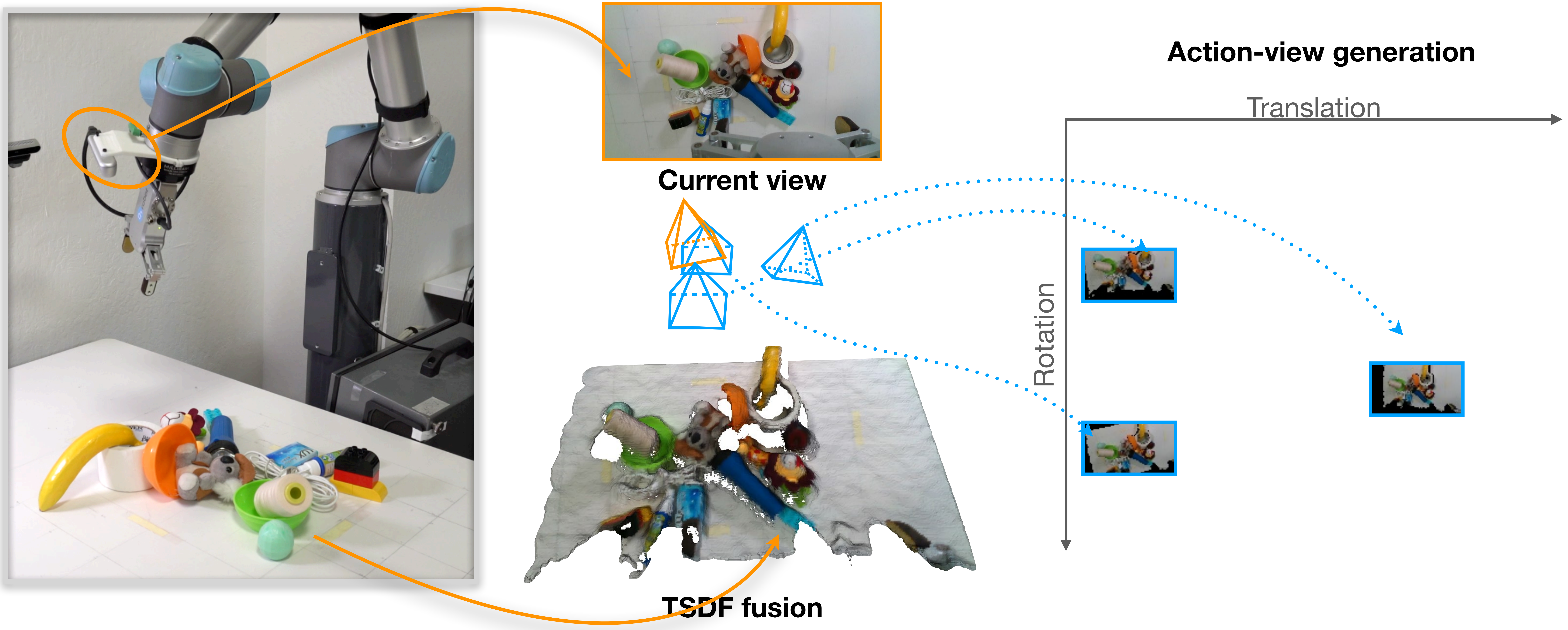


Current view

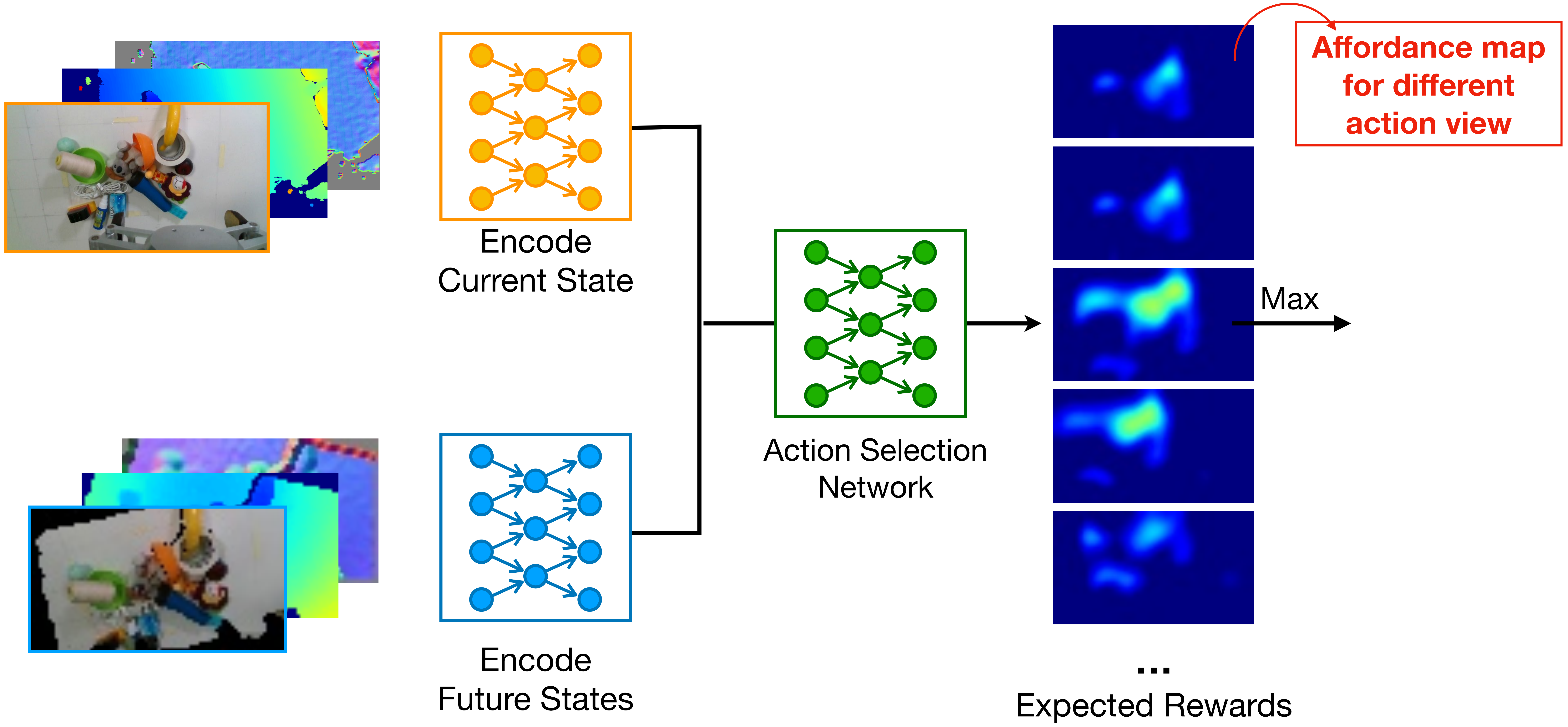


TSDF fusion

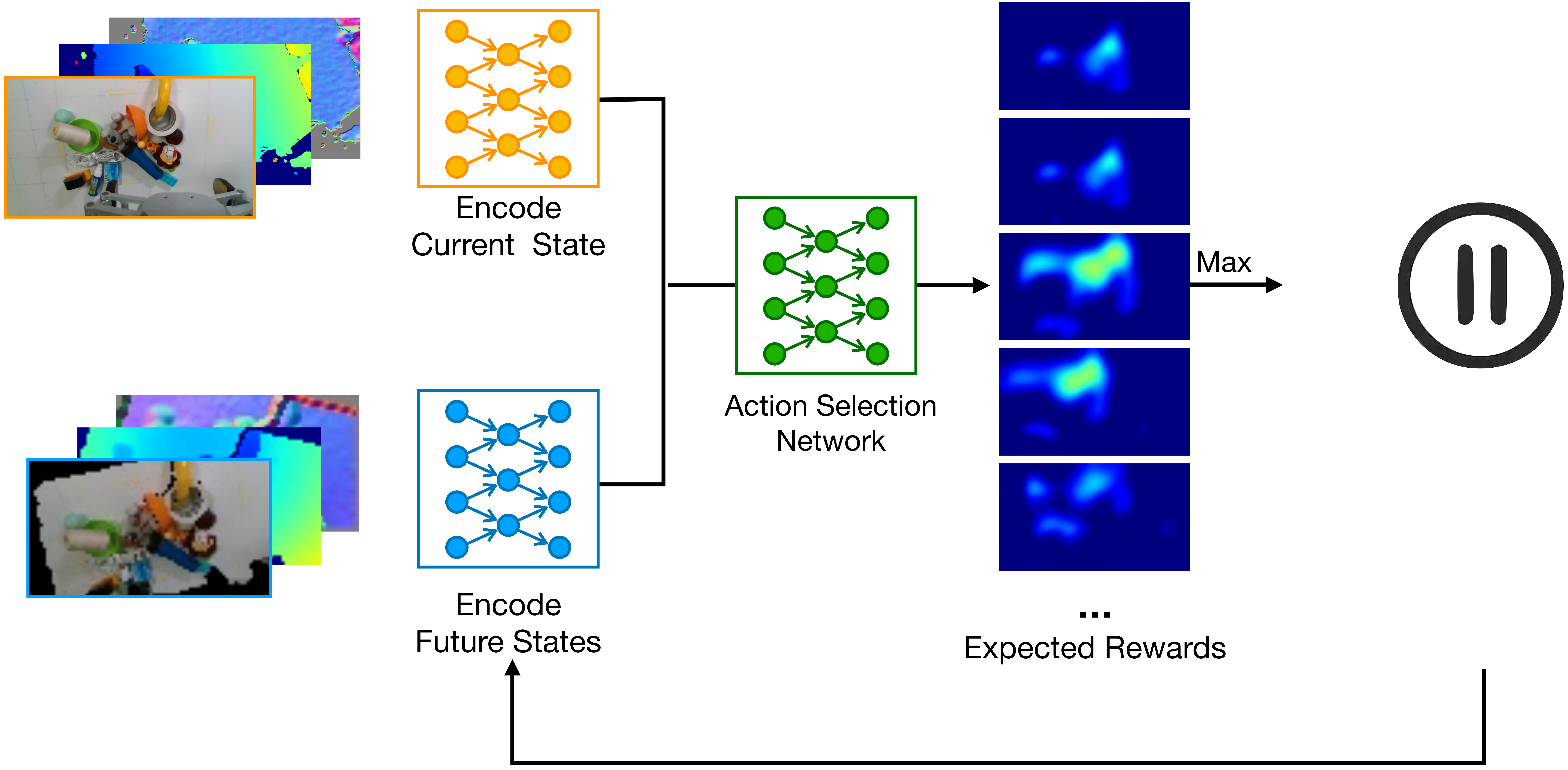
Action-view Grasp Planning



Action-view Grasp Planning



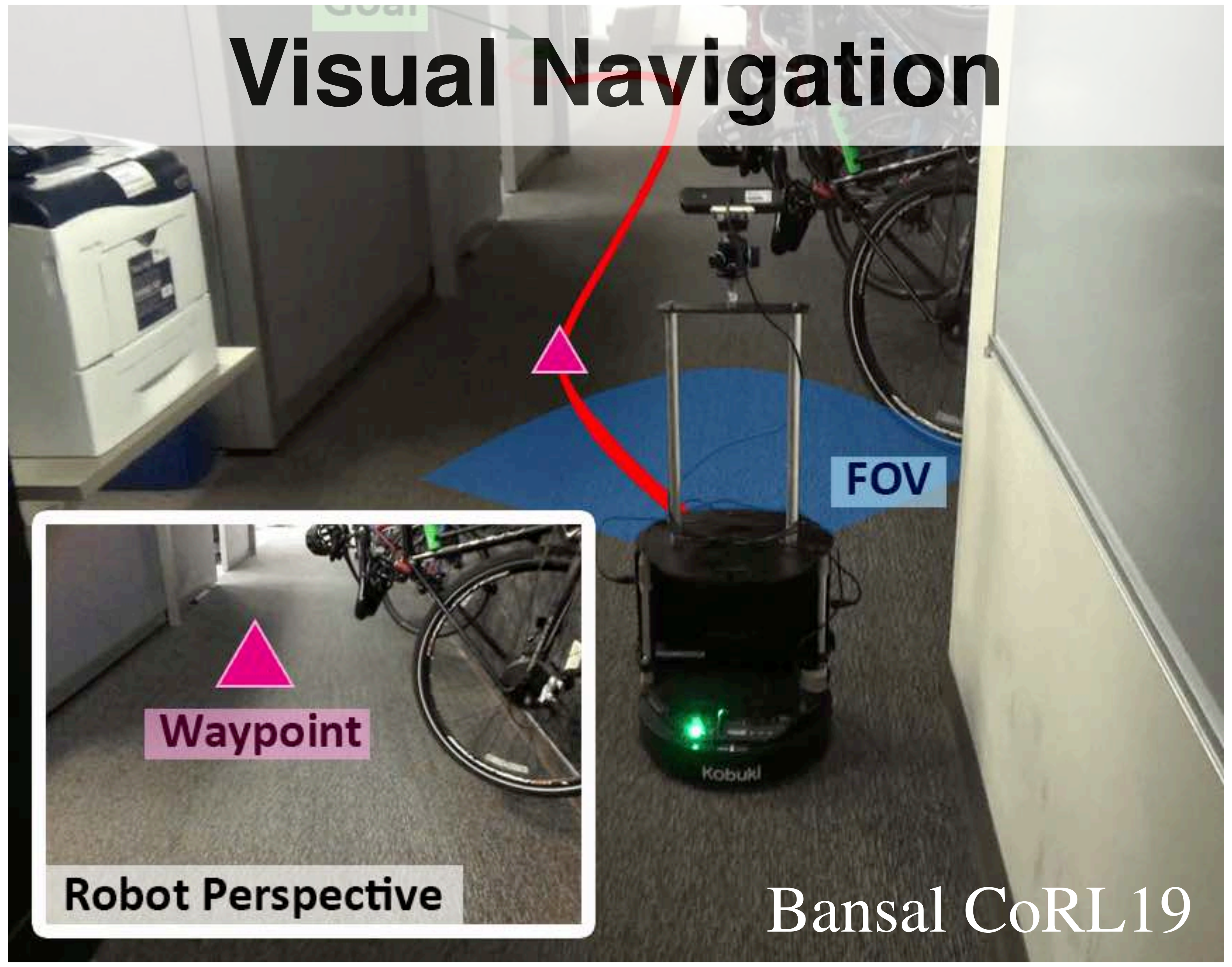
Action-view Grasp Planning

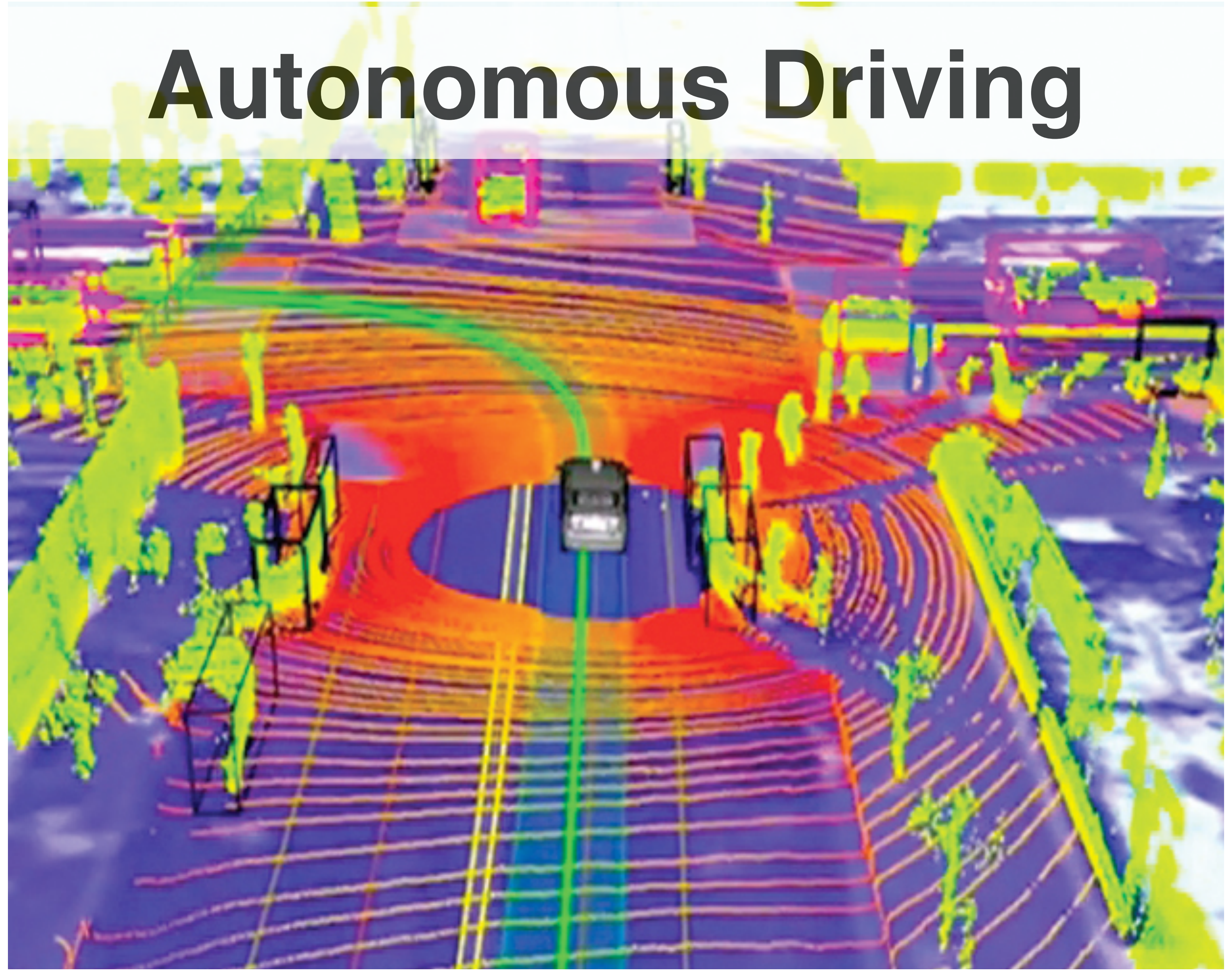


Action-view Grasp Planning

$$f(\text{state}) \rightarrow \text{action}$$

View-based rendering as predictive model



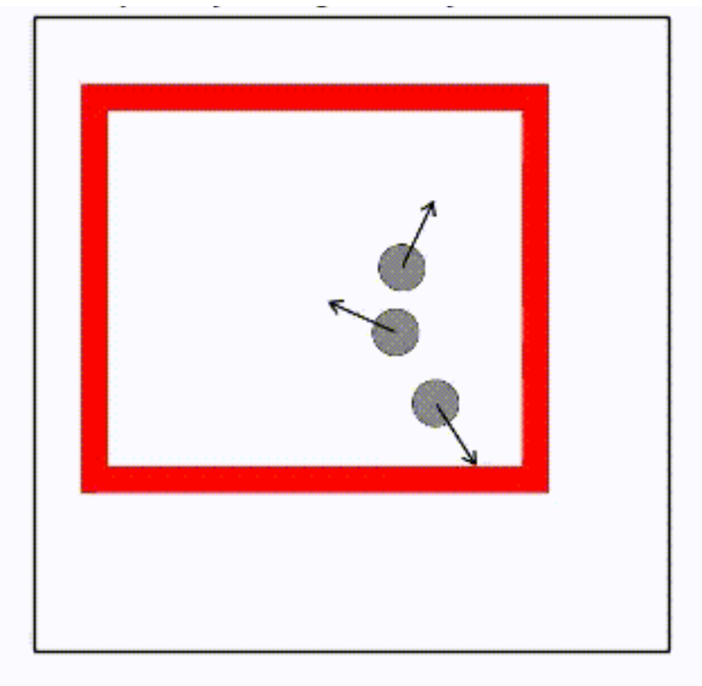


✓ Actions directly lead to ego-centric camera motion

Action-view Grasp Planning

$$f(\text{state}) \rightarrow \text{action}$$

View-based rendering as predictive model



Fragkiadaki et al ICLR16

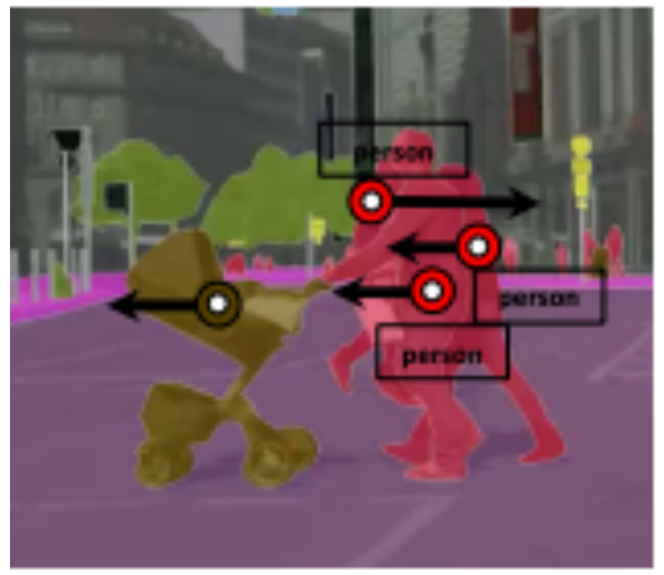
Finn et al ICRA17

Vondrick et al ECCV18

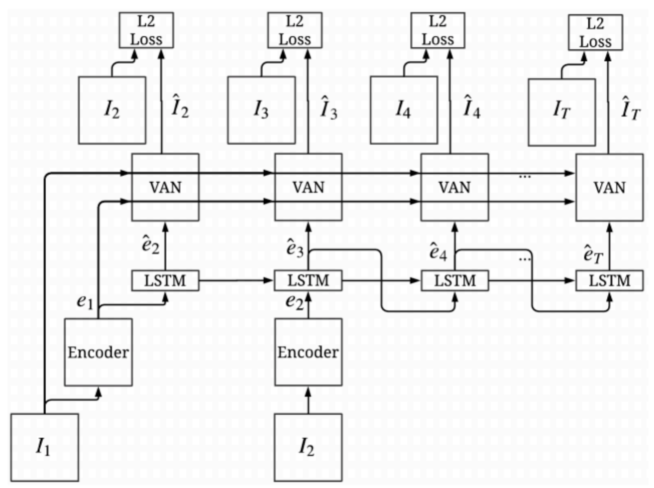
Vondrick et al NIPS18

Tulyakov et al CVPR18

Xue et al NIPS16



Luc et al CVPR17



Wichers et al ICML18

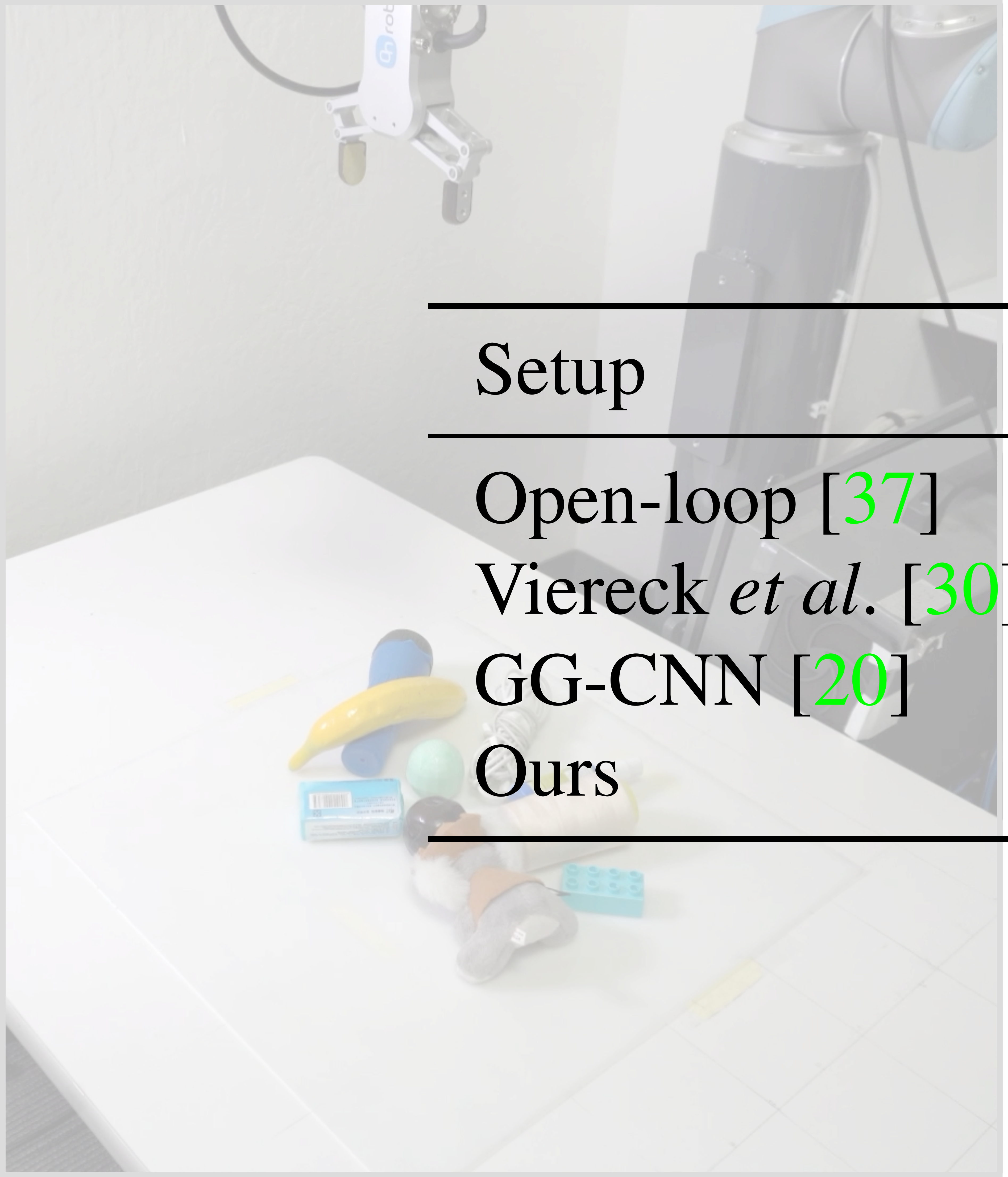
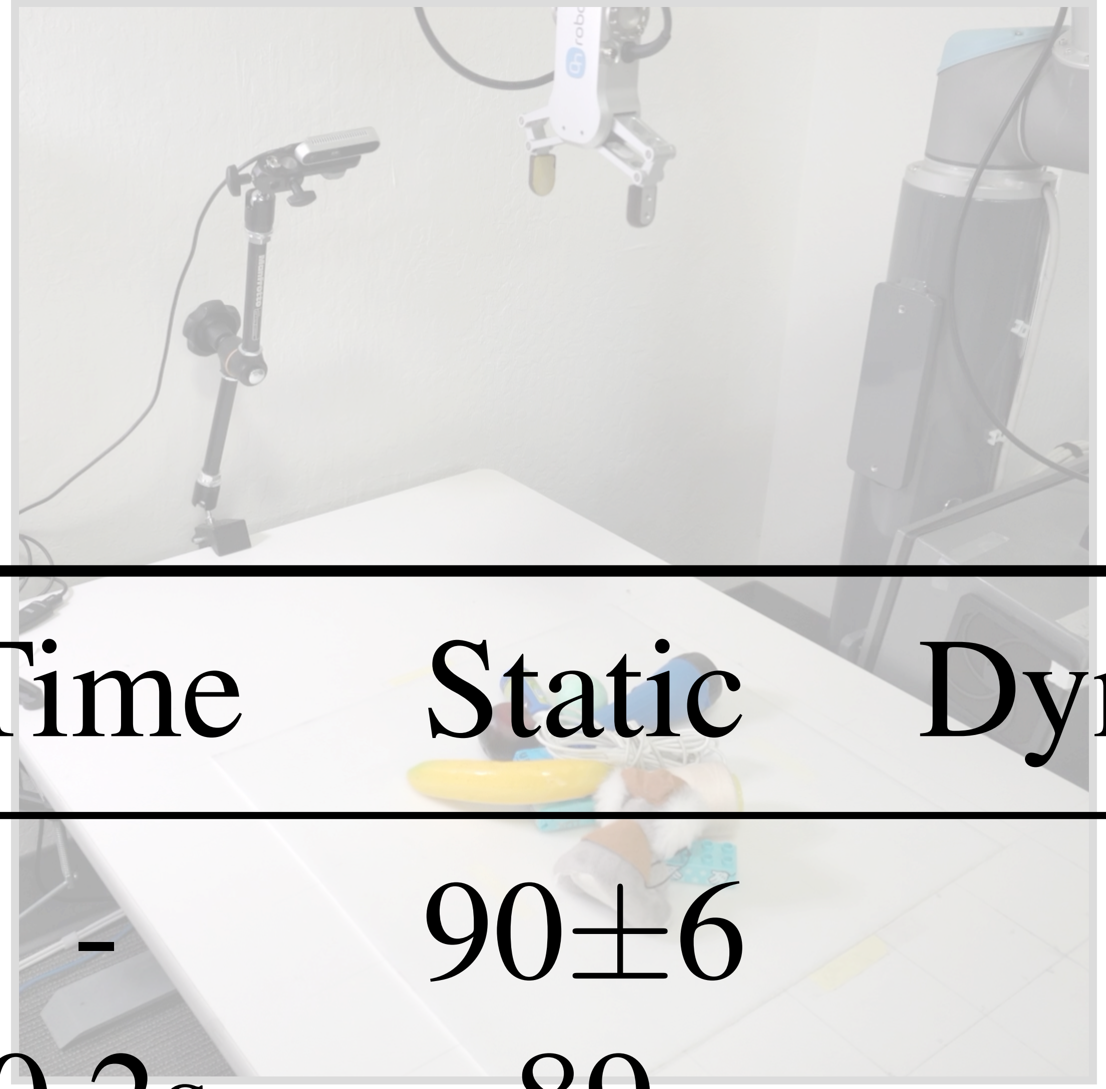
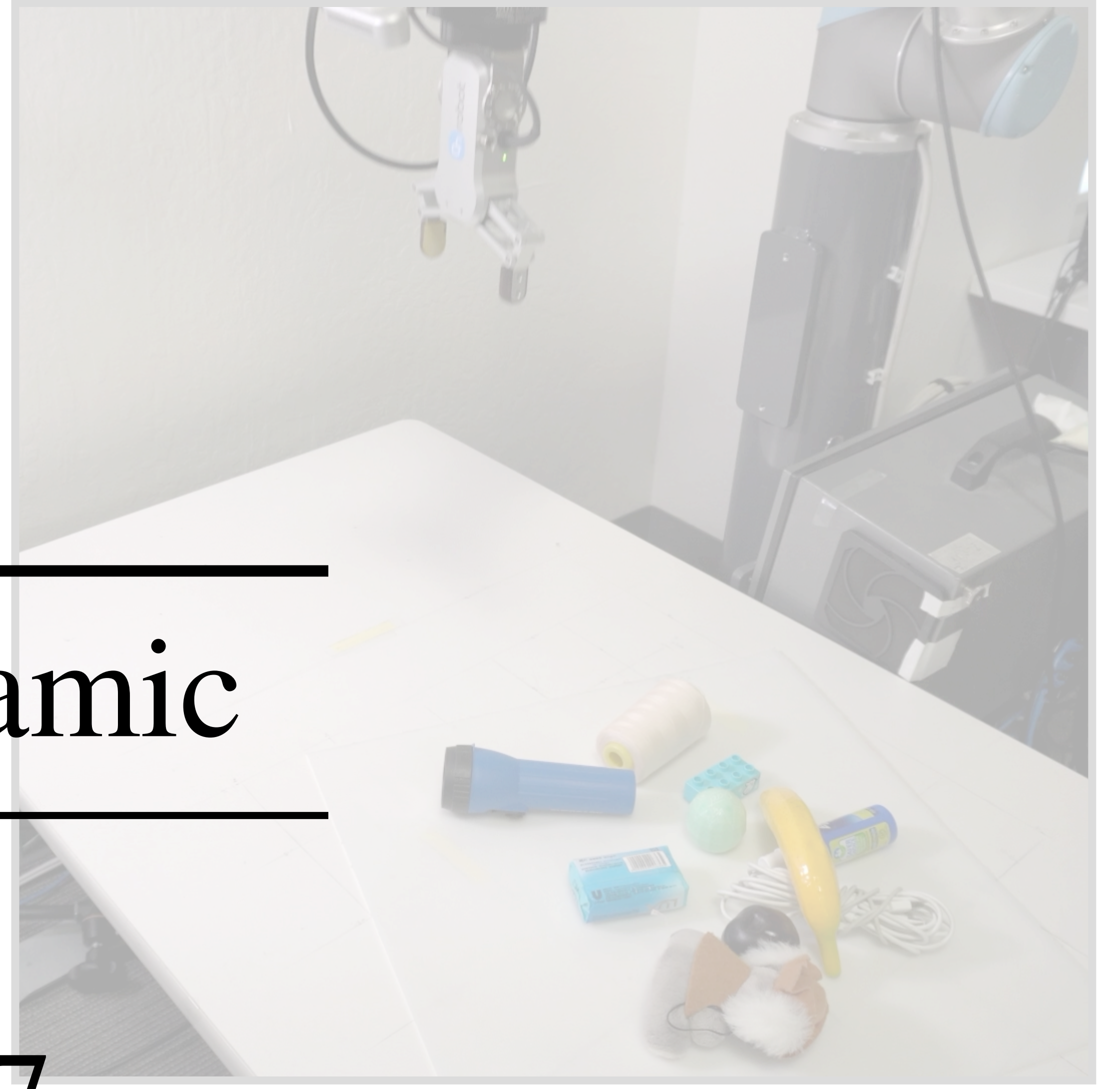
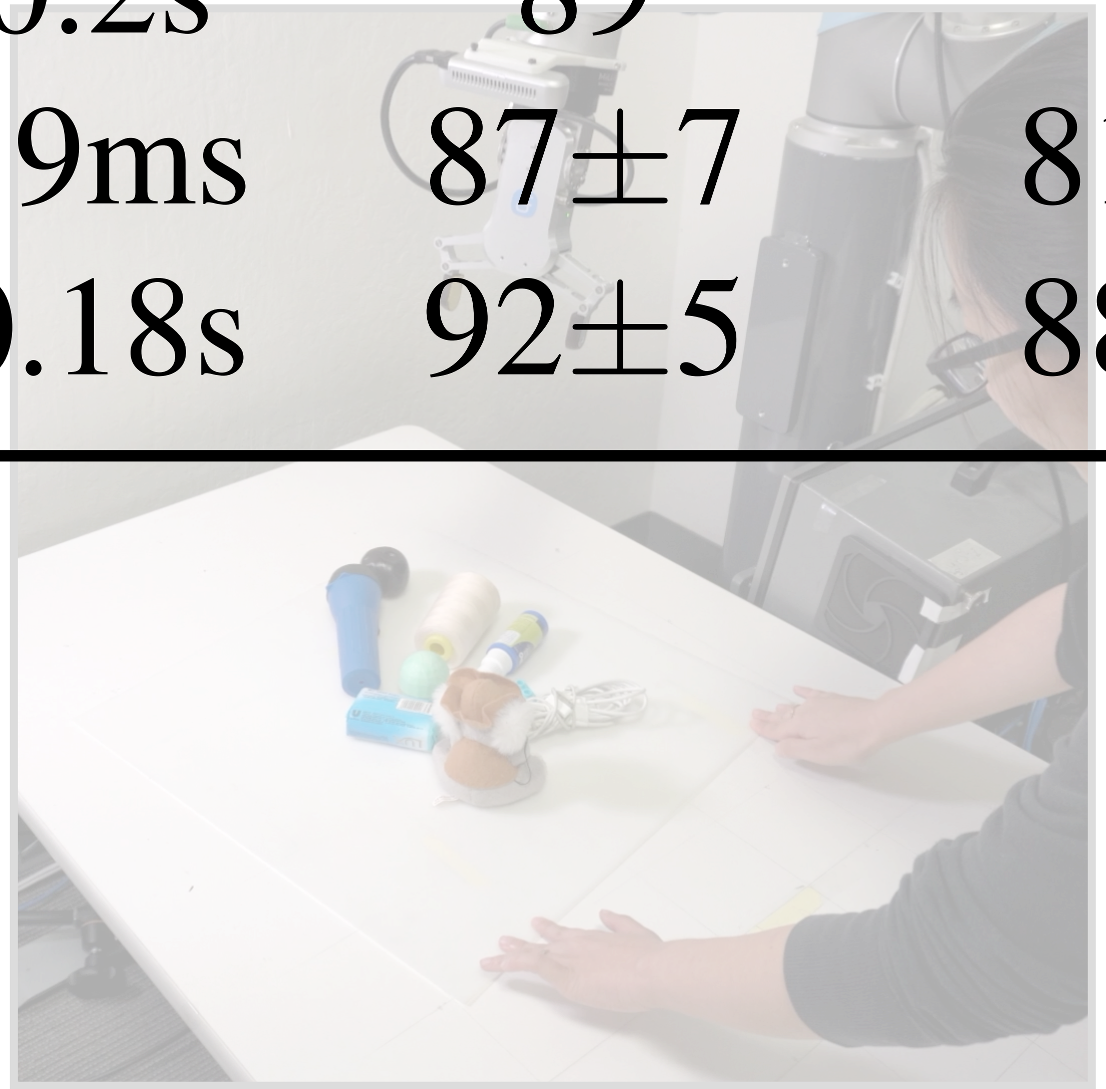
✓ Actions directly lead to ego-centric camera motion

✗ Object and contact physics— learnable predictive model

Experiments

Varying Quasi-static Scenes

Dynamic Scenes


Setup	Time	Static	Dynamic
Open-loop [37]	-	90±6	-
Viereck *et al.* [30]	0.2s	89	77
GG-CNN [20]	19ms	87±7	81±8
Ours	0.18s	92±5	88±8

Summary

✓ Affordance based grasping:

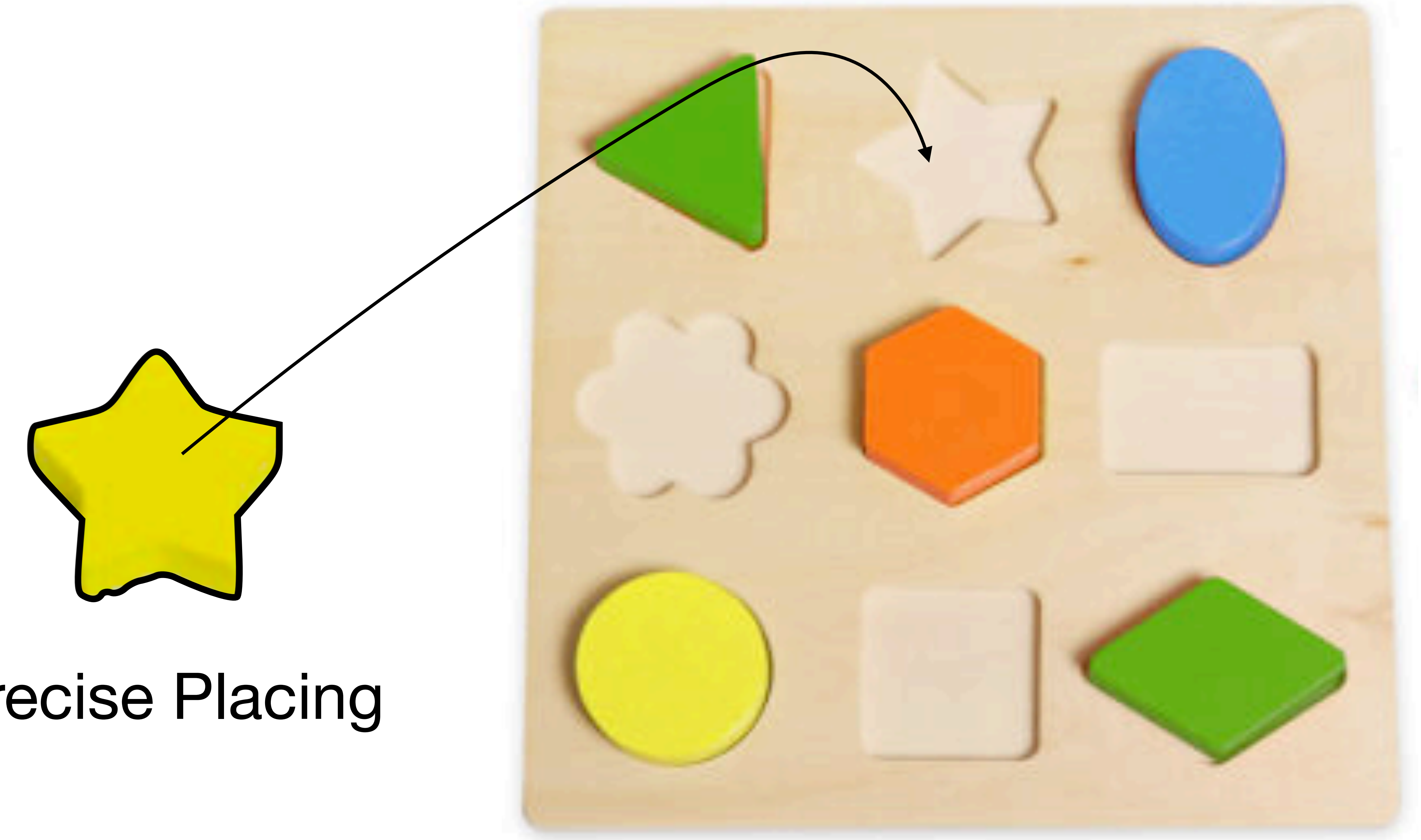
- Good for generalization (No object pose or 3D model needed)

✓ Action-view representation:

- Enables efficient learning of high-degree freedom closed-loop control, by explicitly modeling the action's effect on the state.

Manipulation beyond Grasping

Manipulation tasks beyond grasping: precise placing, assembly …

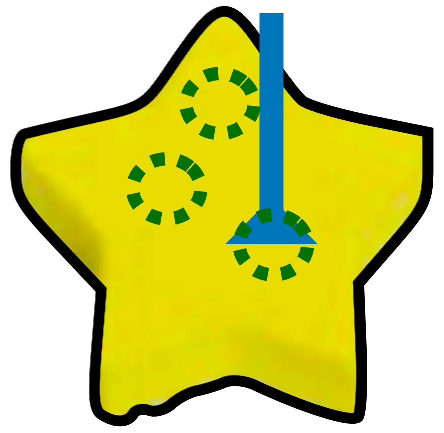


Precise Placing

Manipulation beyond grasping

Manipulation tasks beyond grasping: precise placing, assembly ...

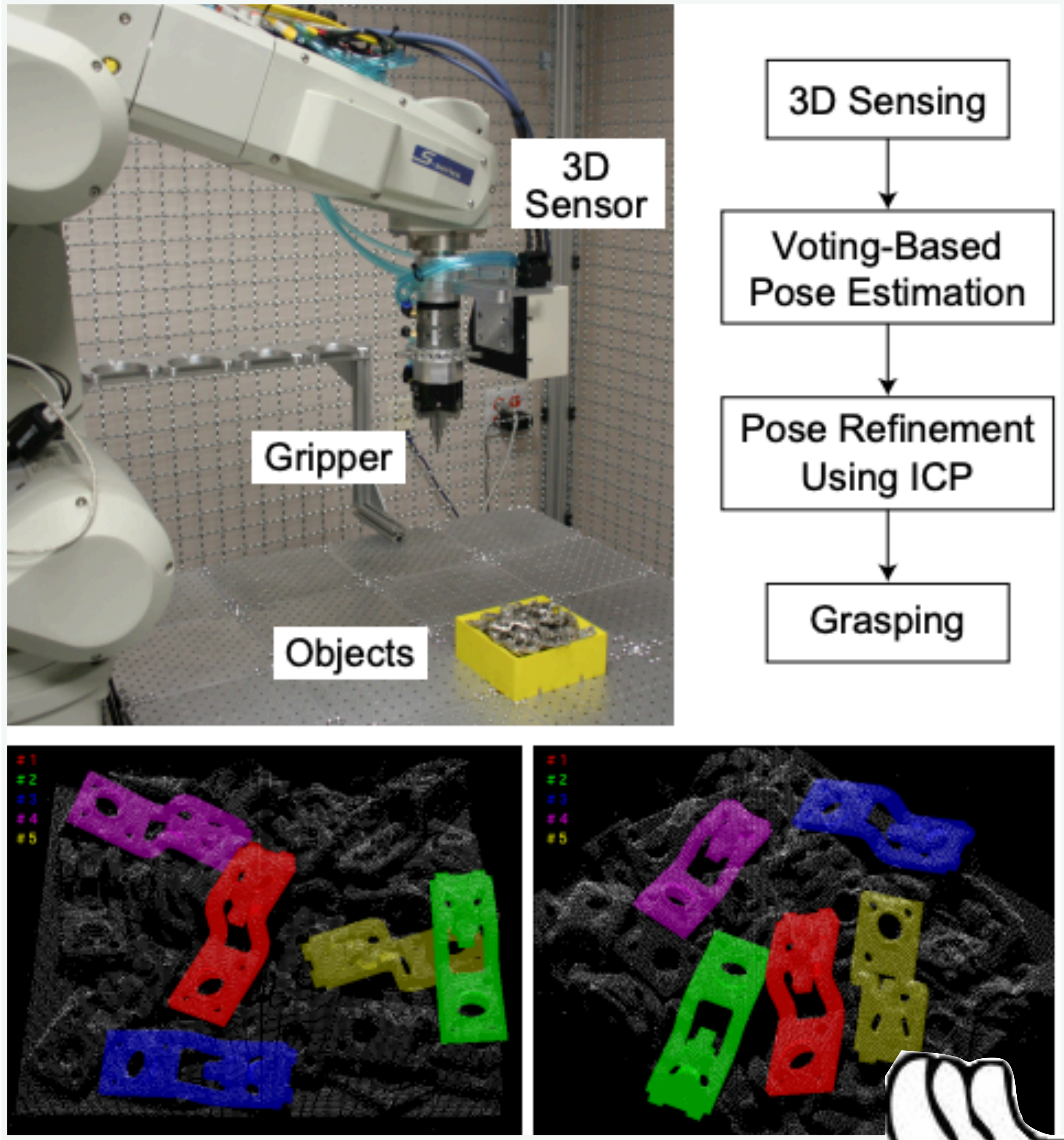
Kit assembly

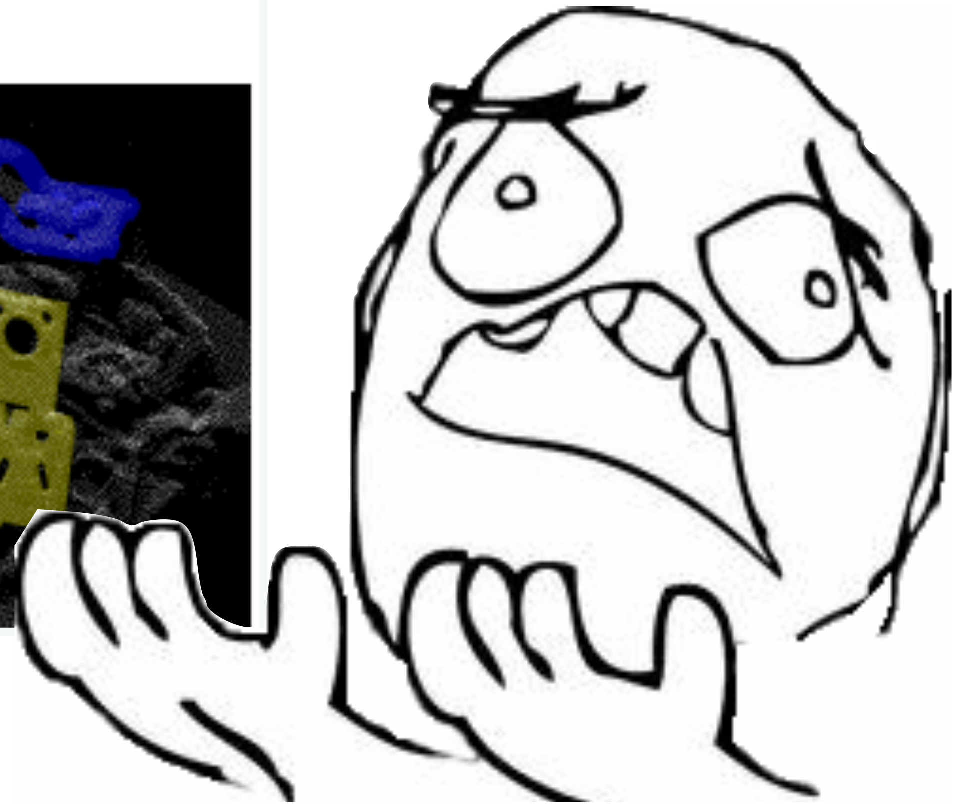


Precise Placing

Kit Assembly

Classic Approach
(Pose estimation)



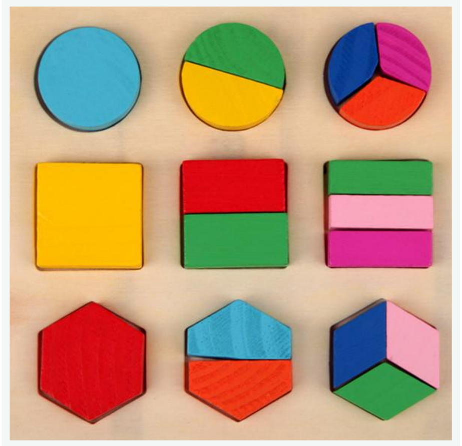


Object Pose Again??

Requires:

- Detailed 3D model
- Extensive Engineering

For every single object

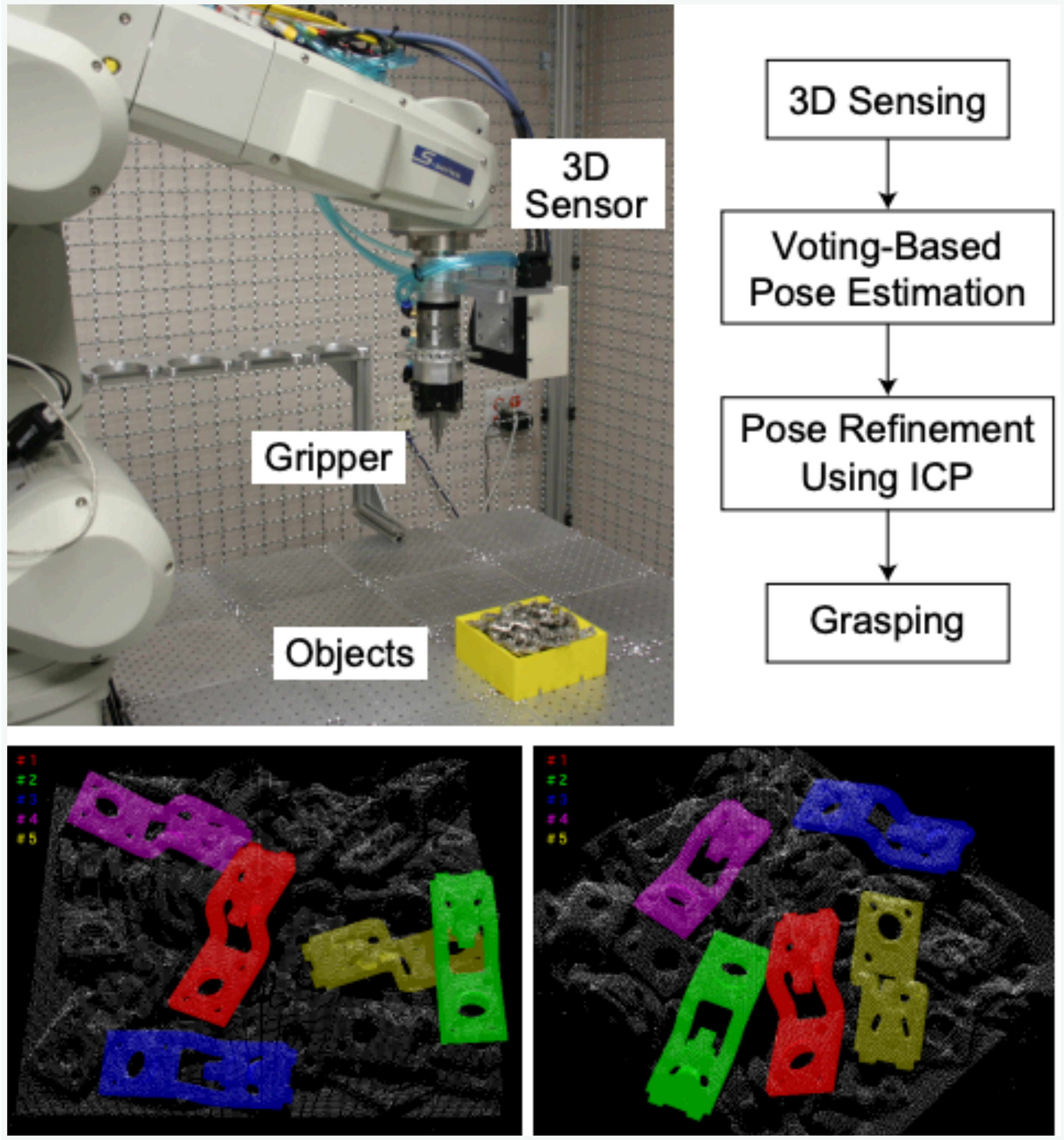


Real-world Applications:

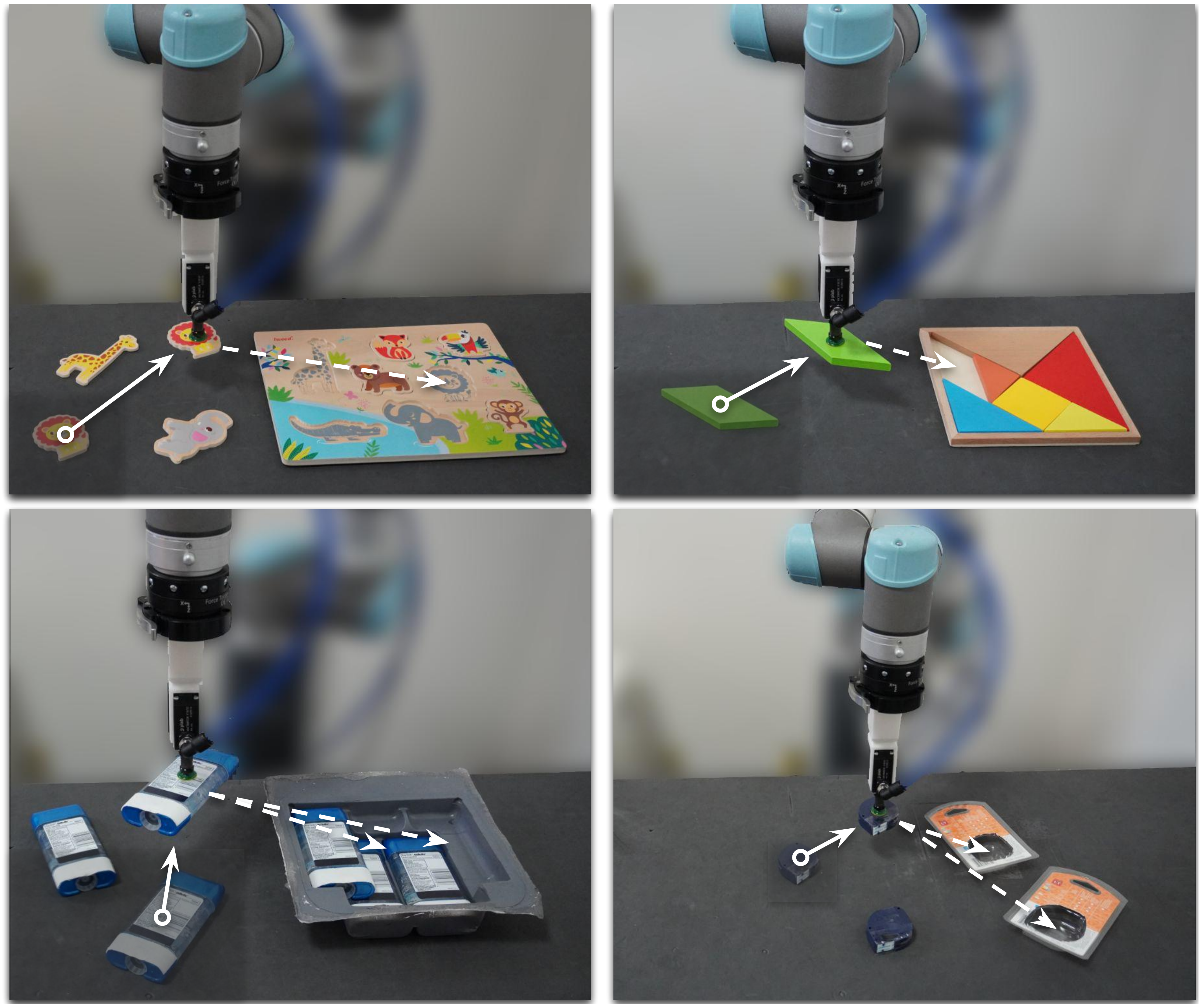
- Fast changing products (promotion/seasonal event)
- Big variation
- Not cost effective

Kit Assembly

Classic Approach
(Pose estimation)



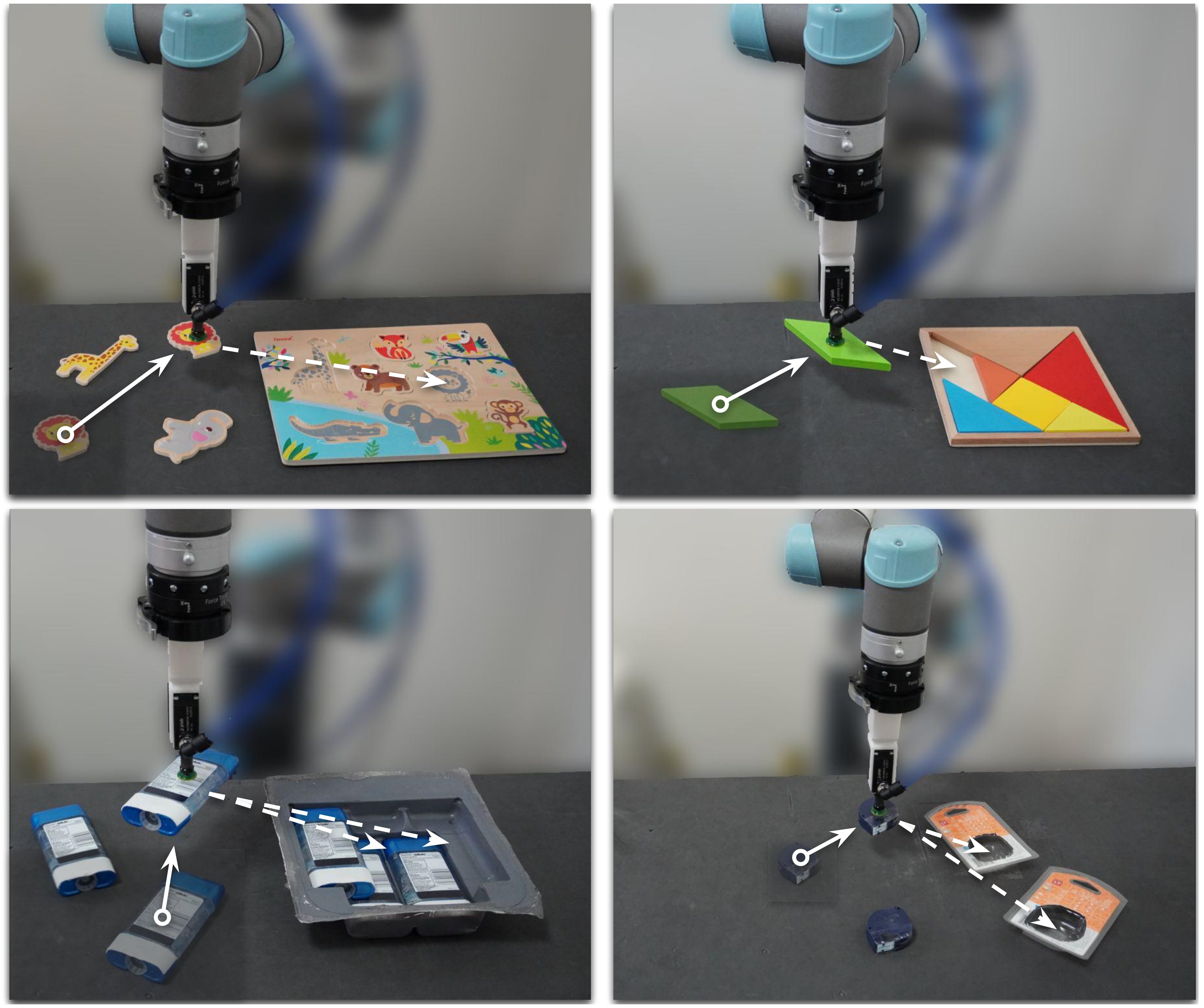
Generalizable Assembly



Goal: develop algorithm that can immediately generalize to new objects

Kit Assembly

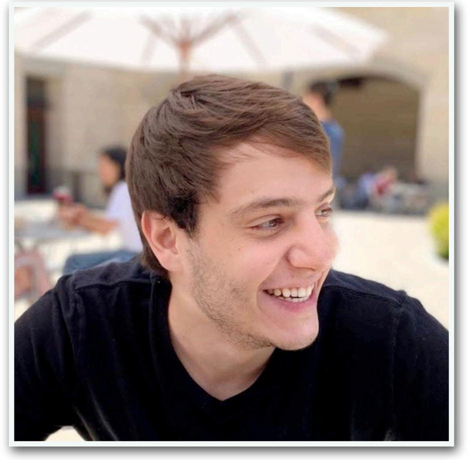
Generalizable Assembly



Form2Fit: Learning Shape Priors for Generalizable Assembly from Disassembly

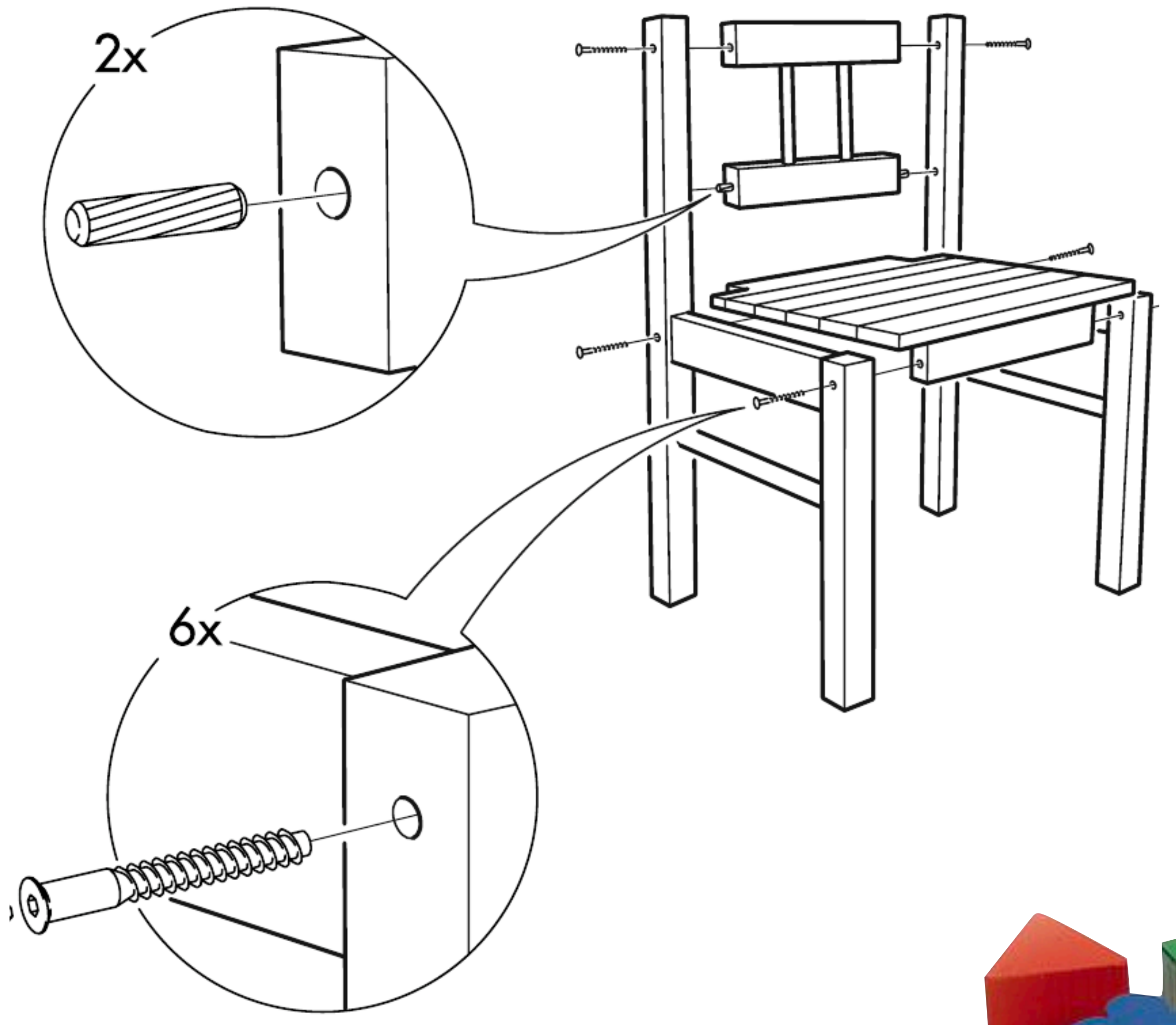
Kevin Zakka, Andy Zeng, Johnny Lee, Shuran Song

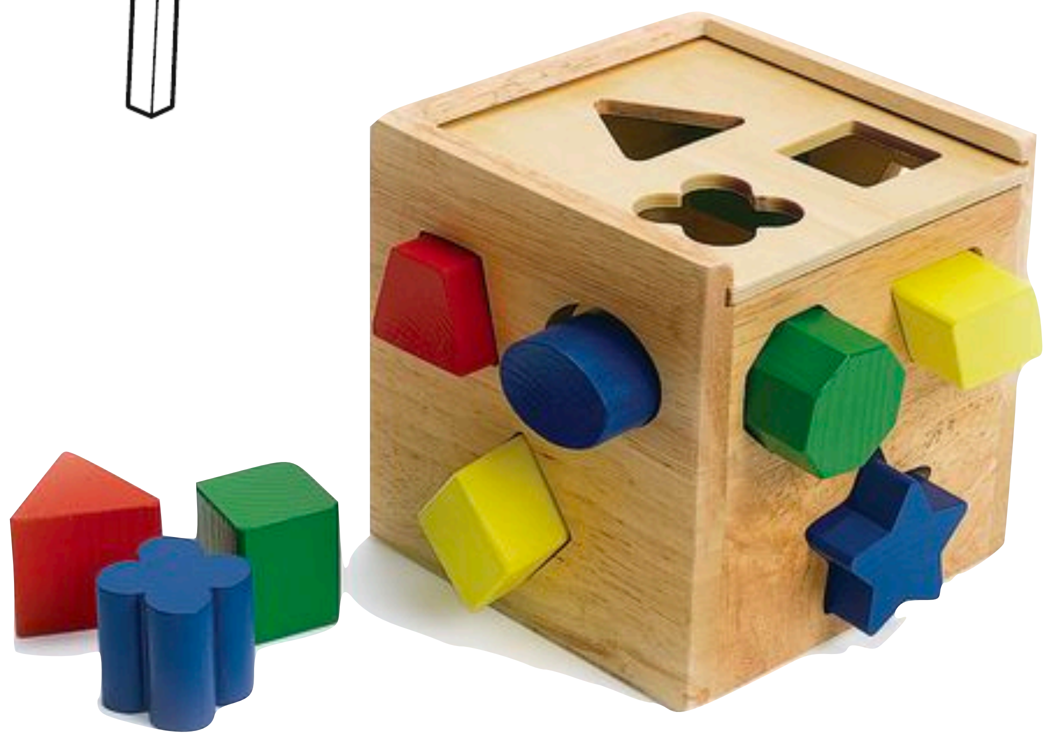
ICRA 2020, Best Paper in Automation Award Finalist

Form2Fit

Learning Shape Prior for Assembly

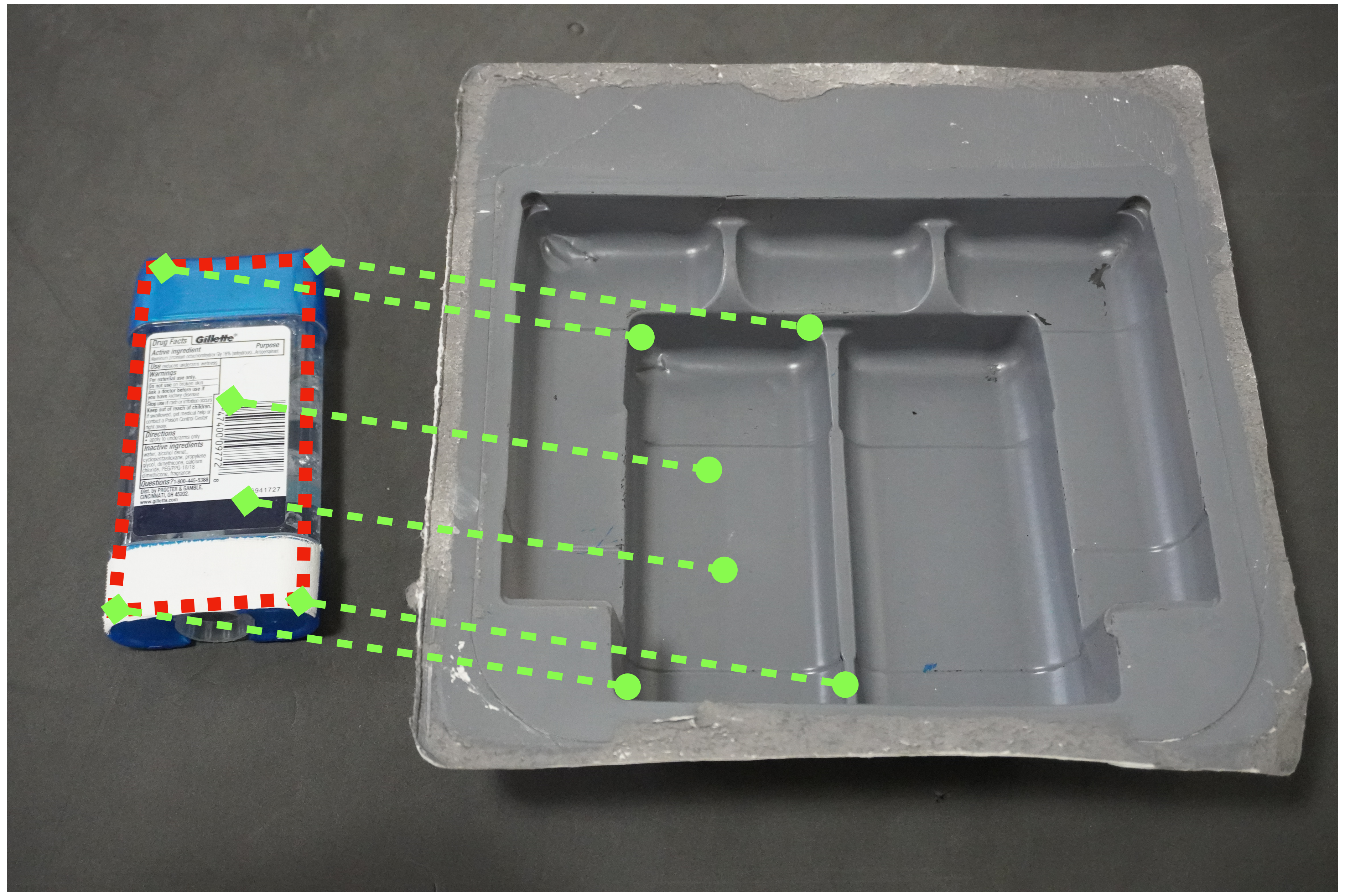




How things fit together?

Form2Fit

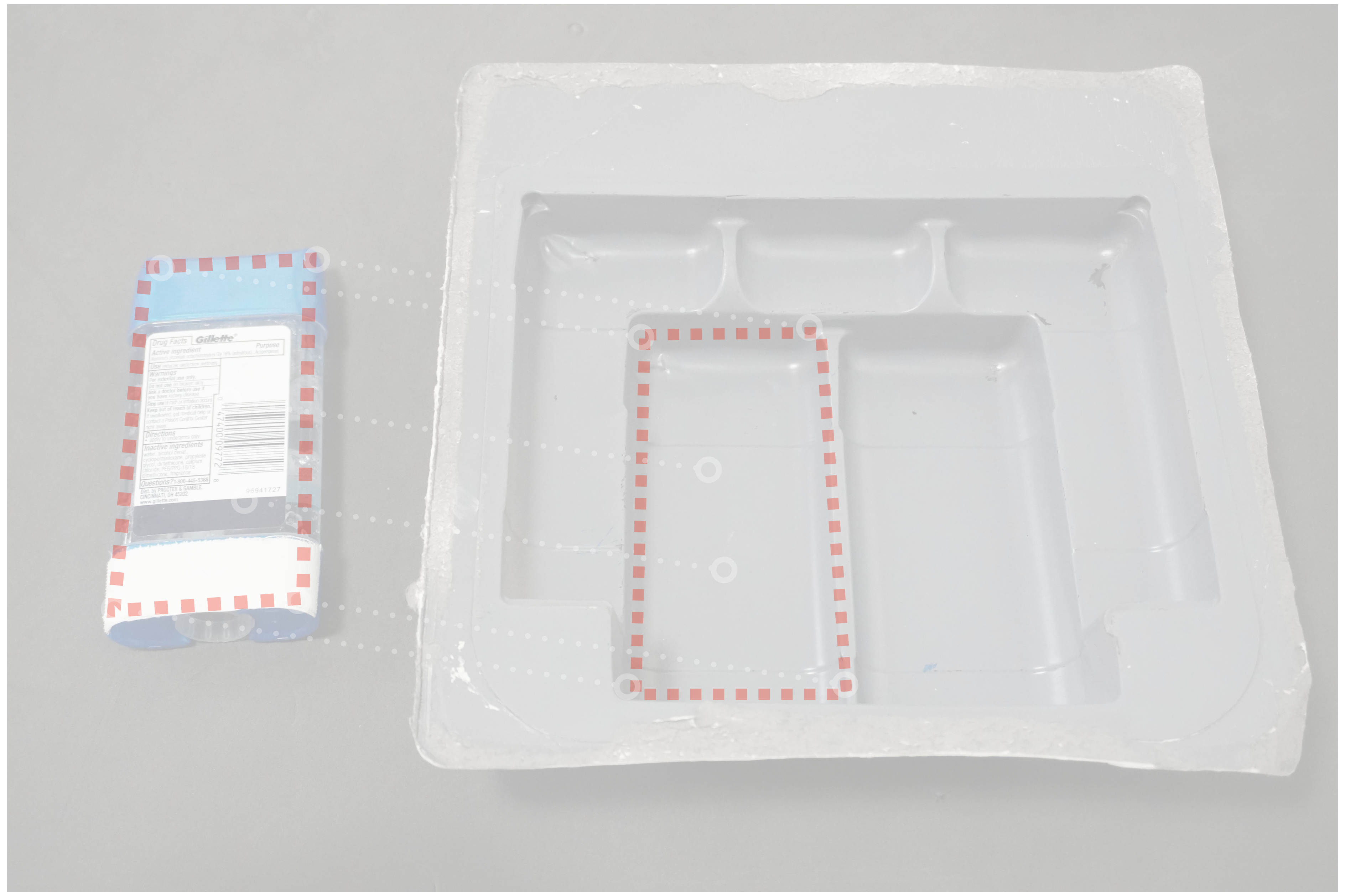
Learning Shape Prior for Assembly



Learns dense shape descriptors to establishes correspondences

Form2Fit

Learning Shape Prior for Assembly



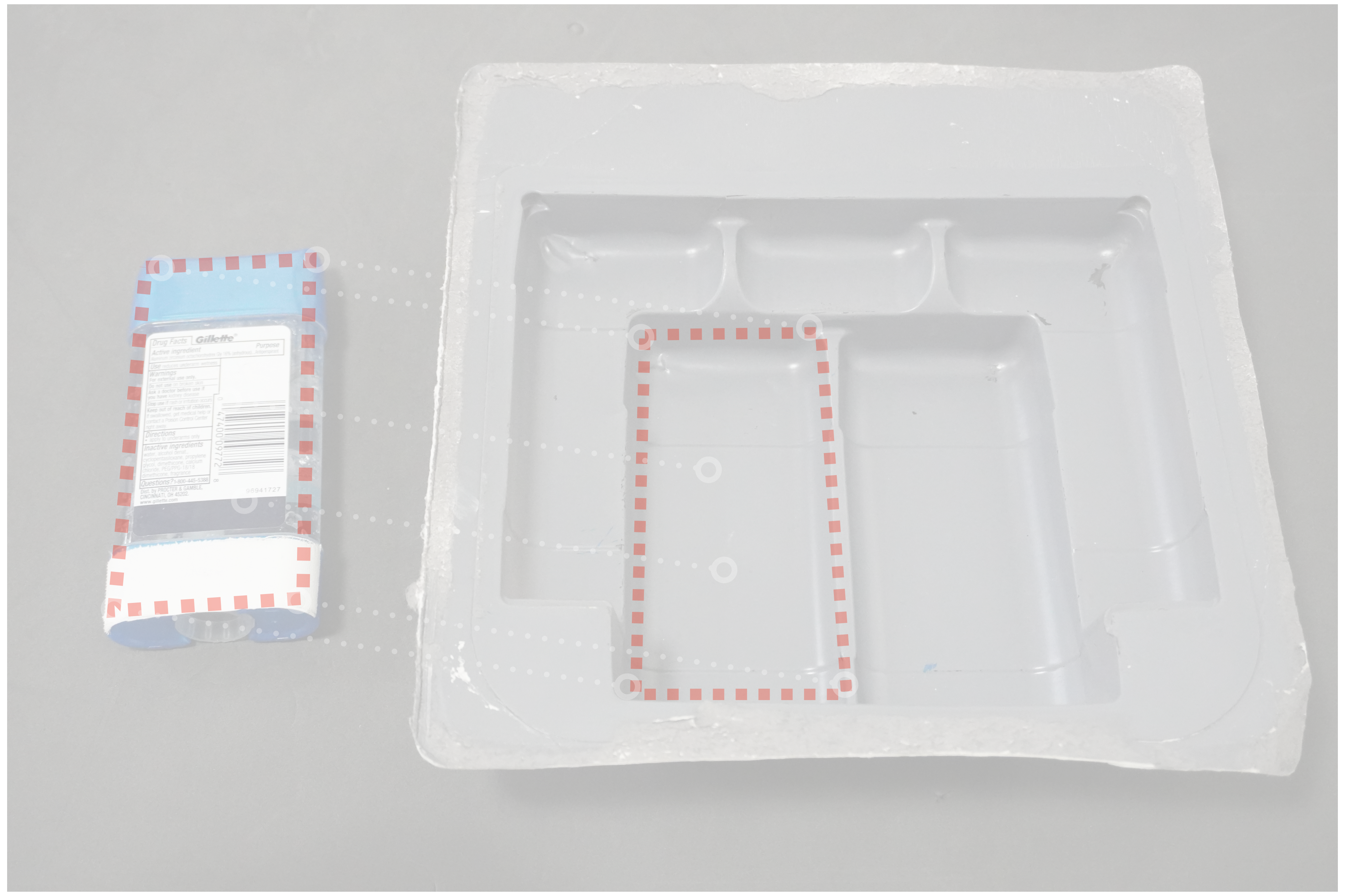
Learns dense shape descriptors to establishes correspondences

Learning Assembly from Disassembly

Training data generation

Form2Fit

Learning Shape Prior for Assembly



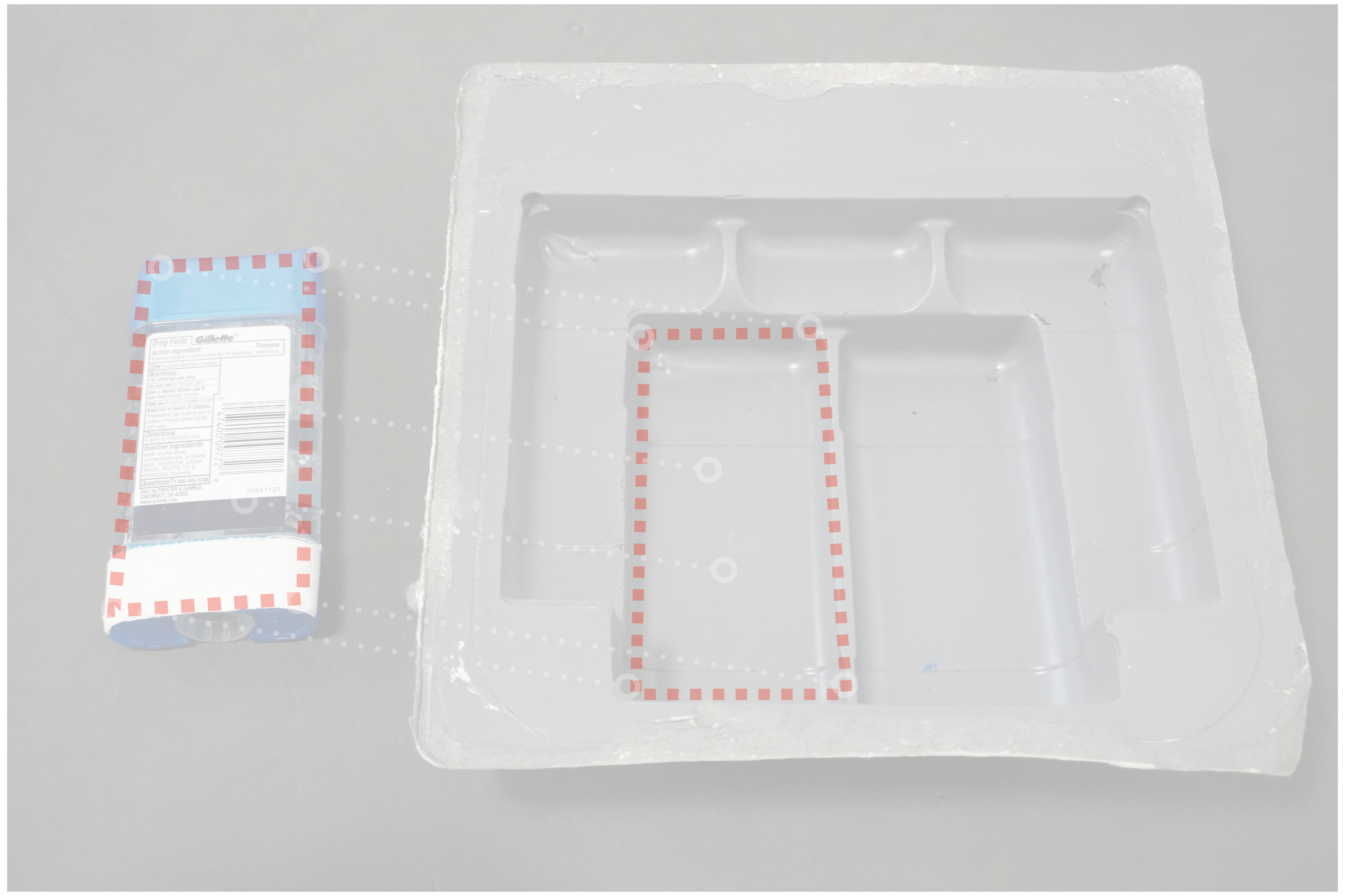
Learns dense shape descriptors to establishes correspondences

Learning Assembly from Disassembly

Disassembly is easier than assembly

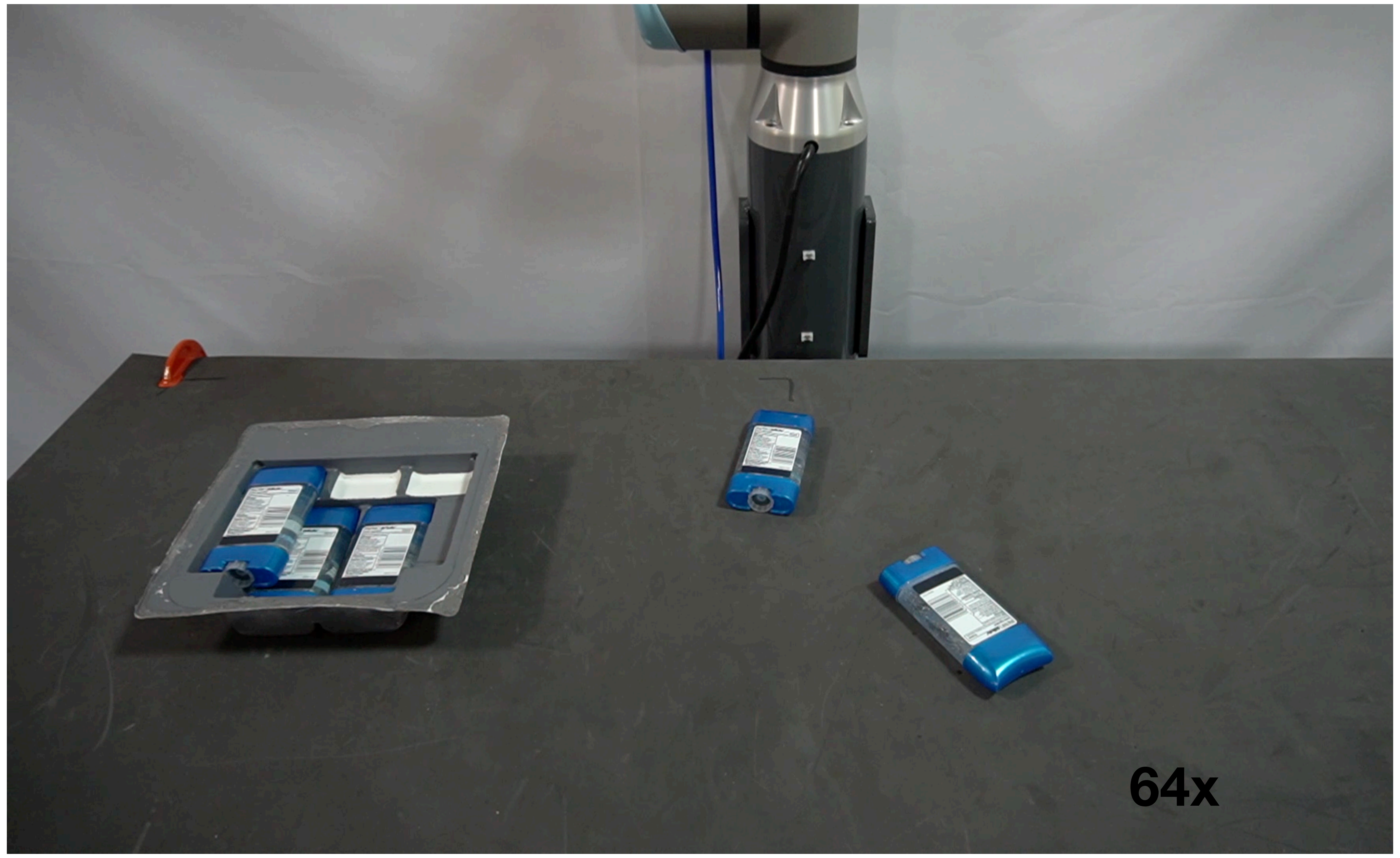
Form2Fit

Learning Shape Prior for Assembly



Learns dense shape descriptors to establishes correspondences

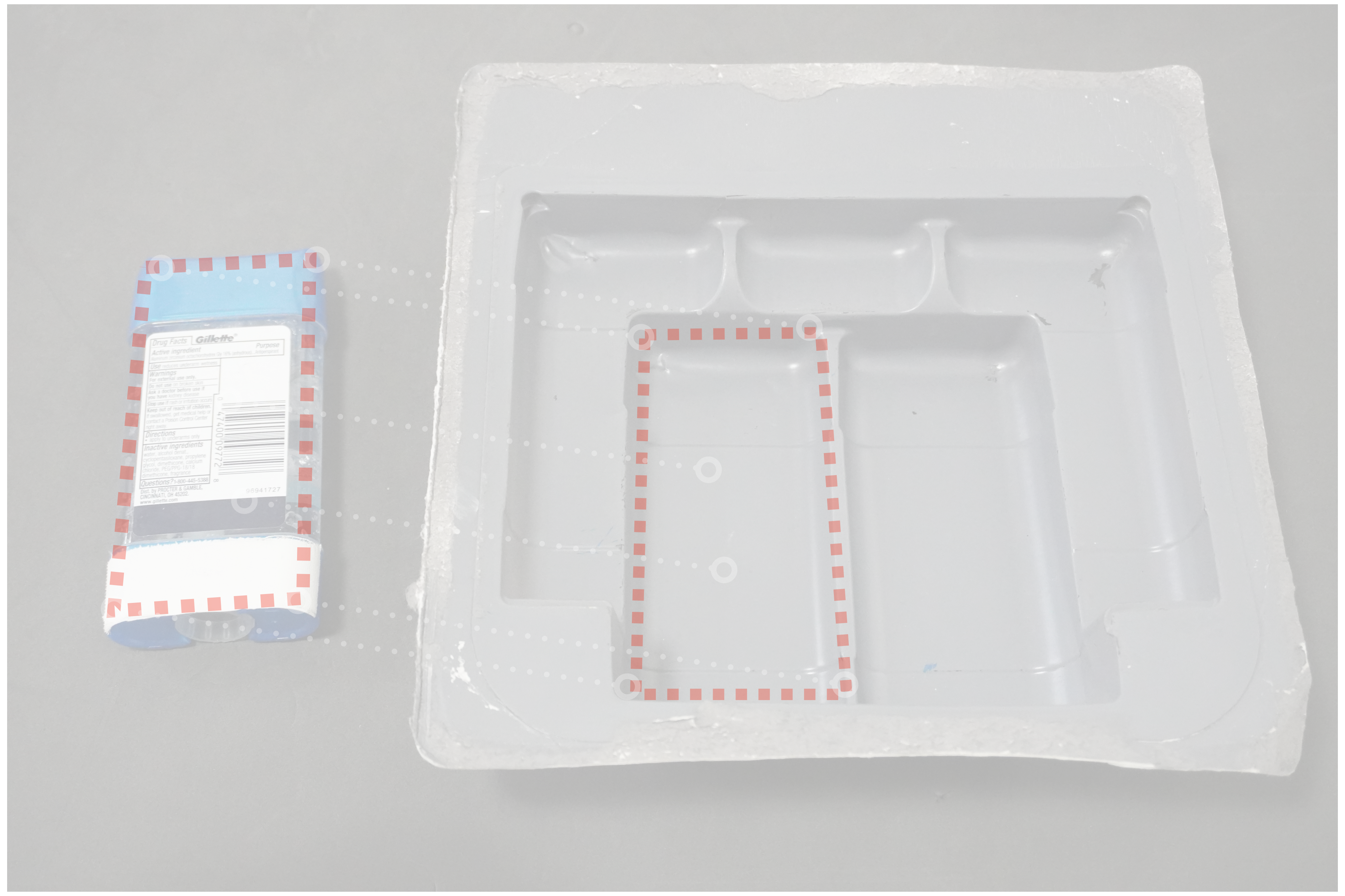
Learning Assembly from Disassembly



Fully self-supervised ground-truth label for shape correspondence

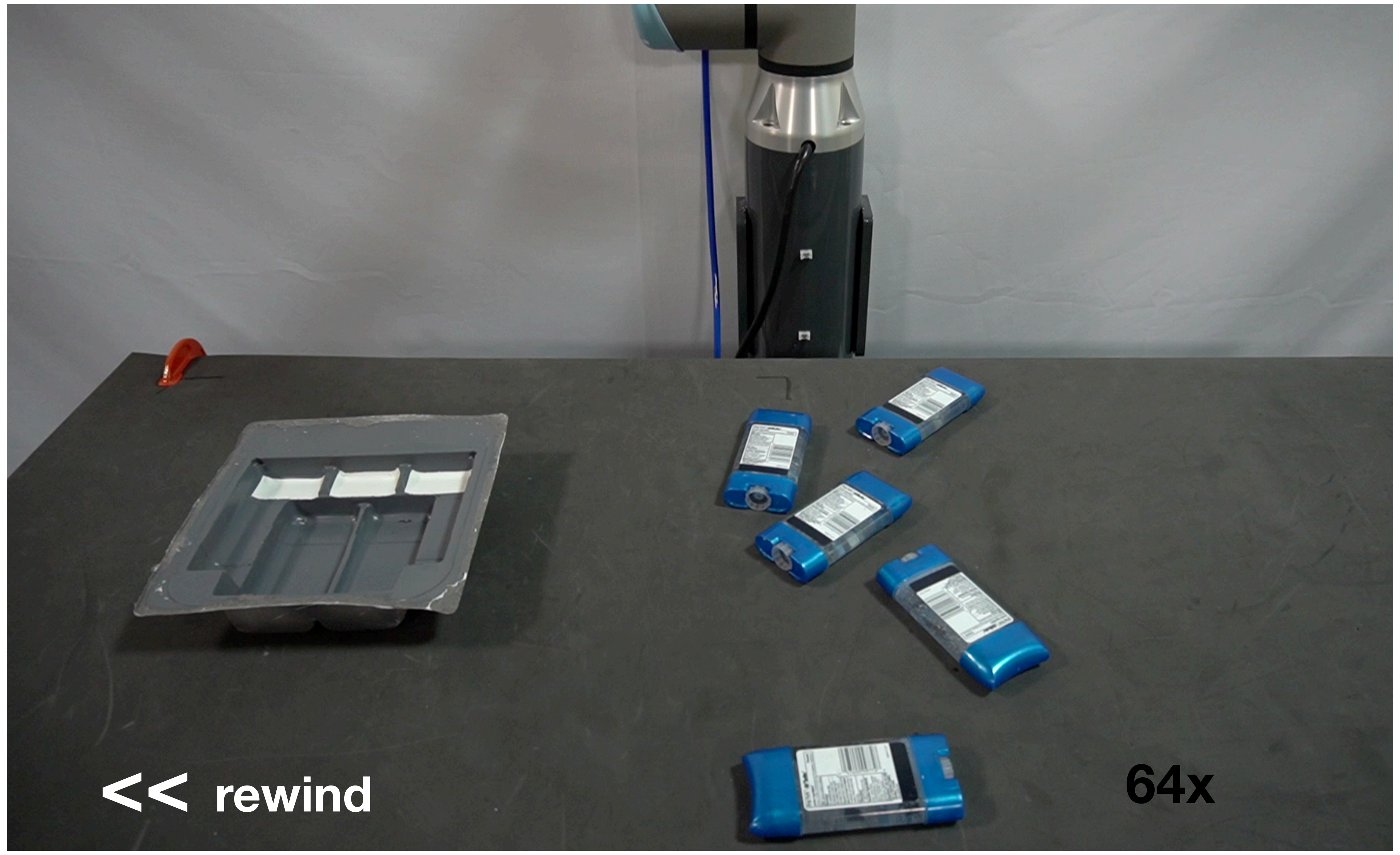
Form2Fit

Learning Shape Prior for Assembly



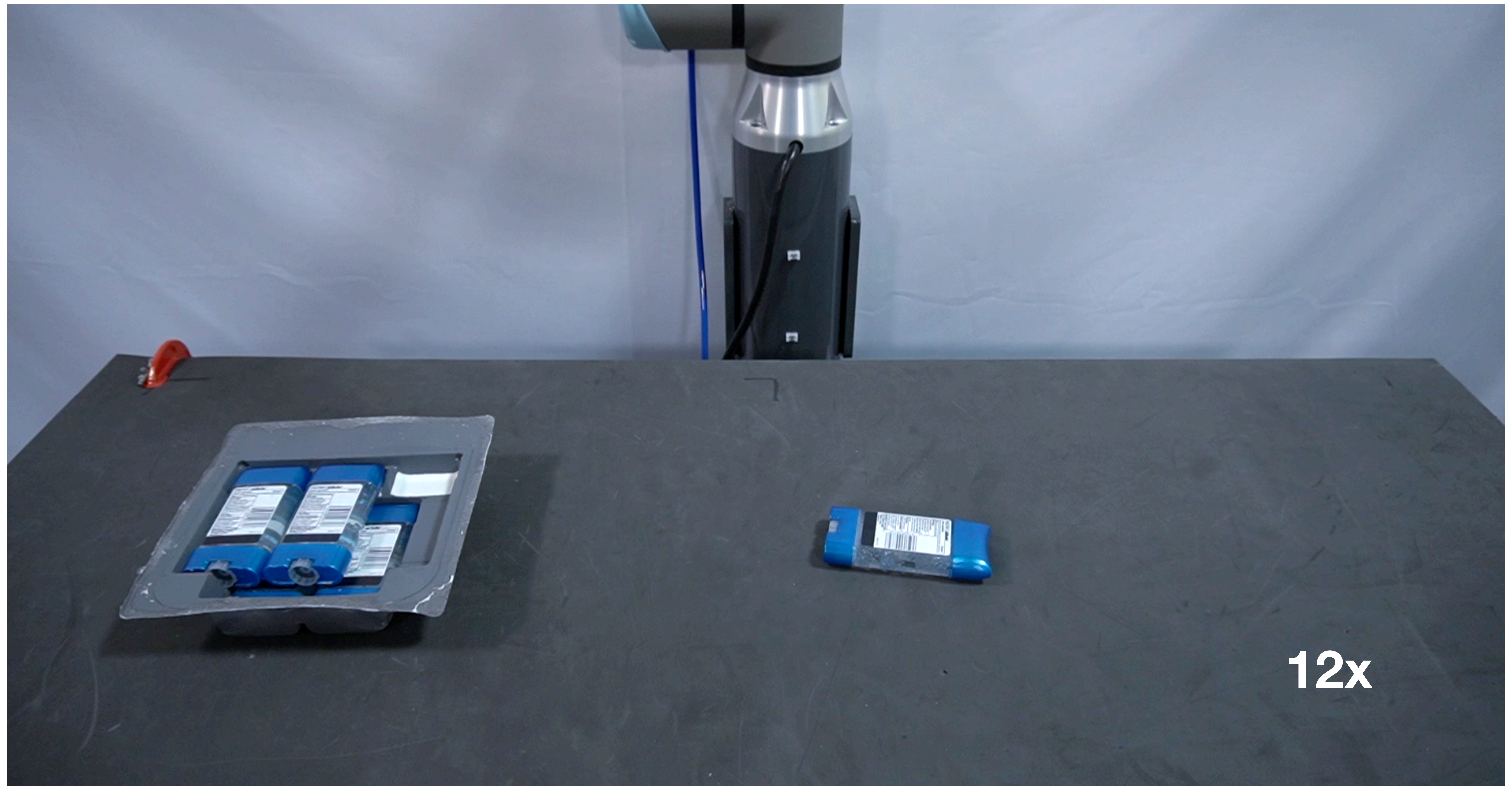
Learns dense shape descriptors to establishes correspondences

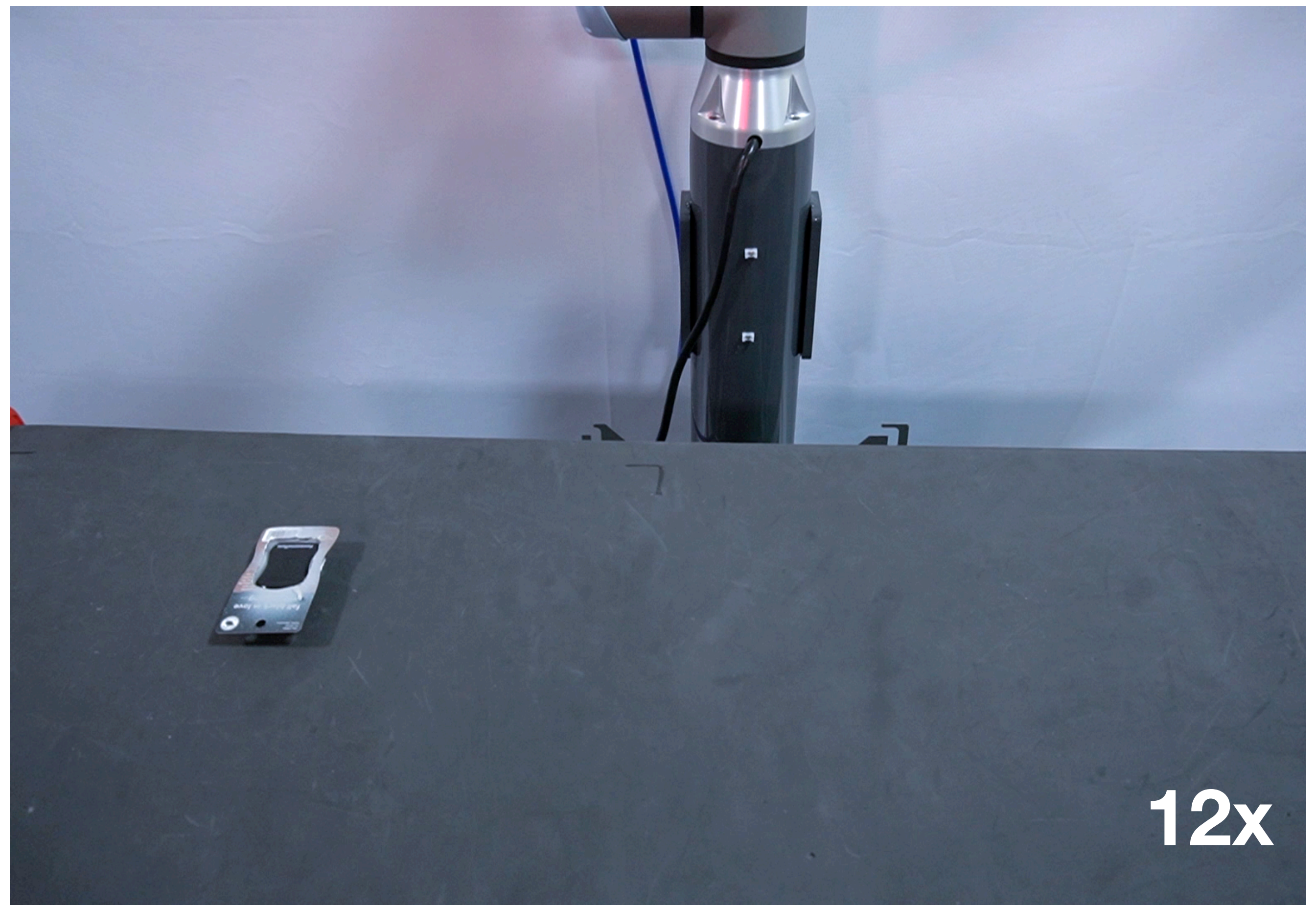
Learning Assembly from Disassembly

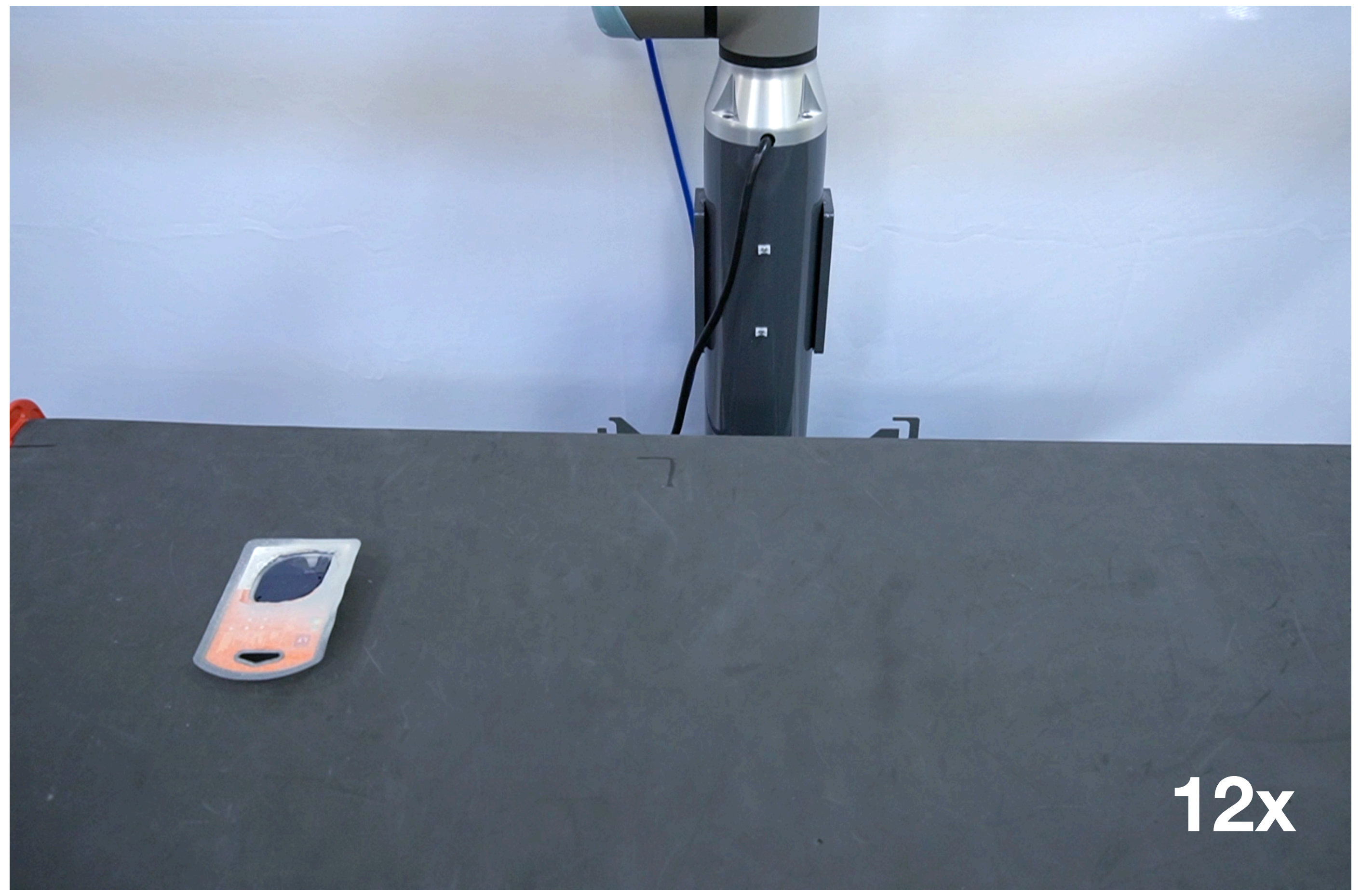


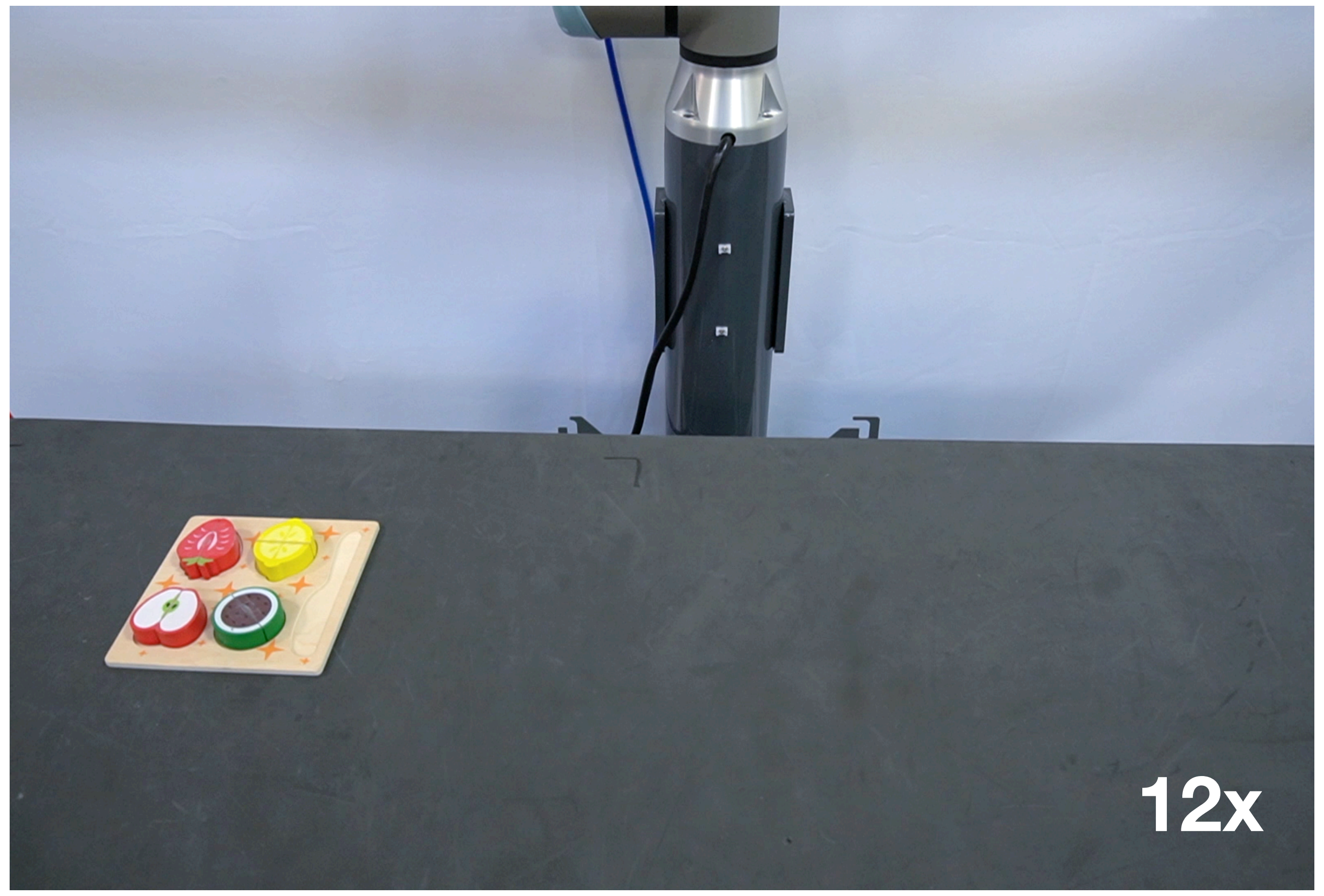
Fully self-supervised ground-truth label for shape correspondence

Data Collection from Disassembly





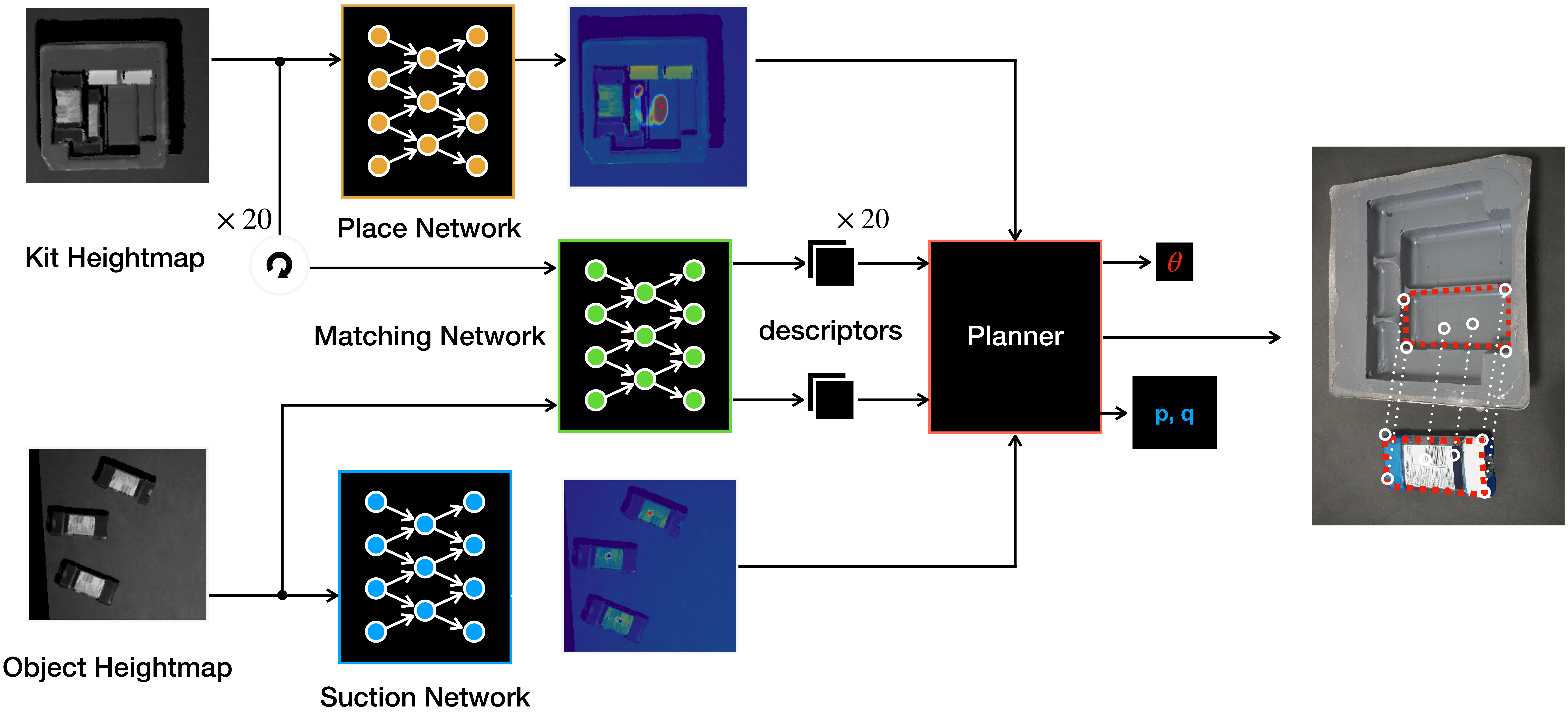




Self-supervised Disassembly

kit is secured to table to prevent accidental displacement from bad suction grasps

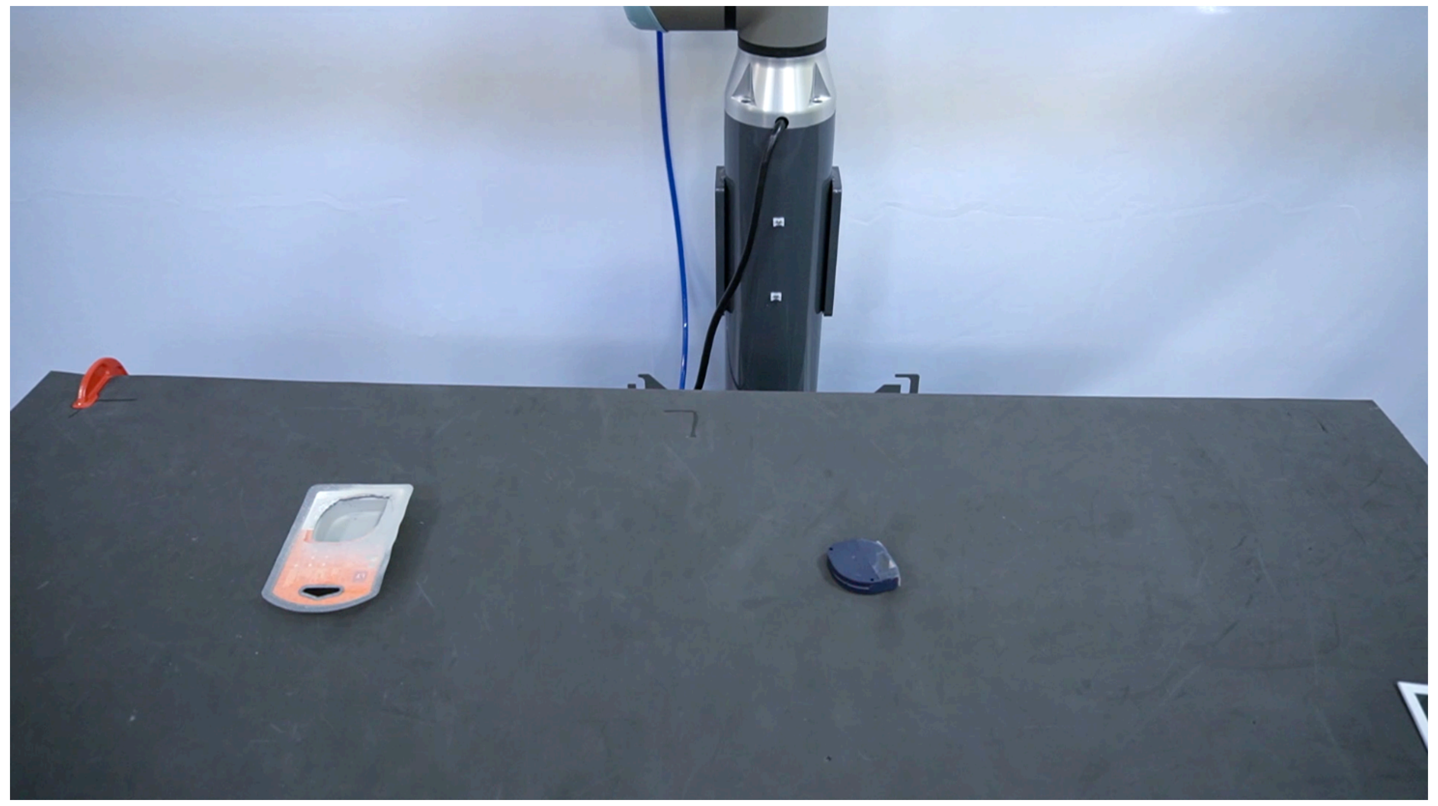
Shape Matching Network



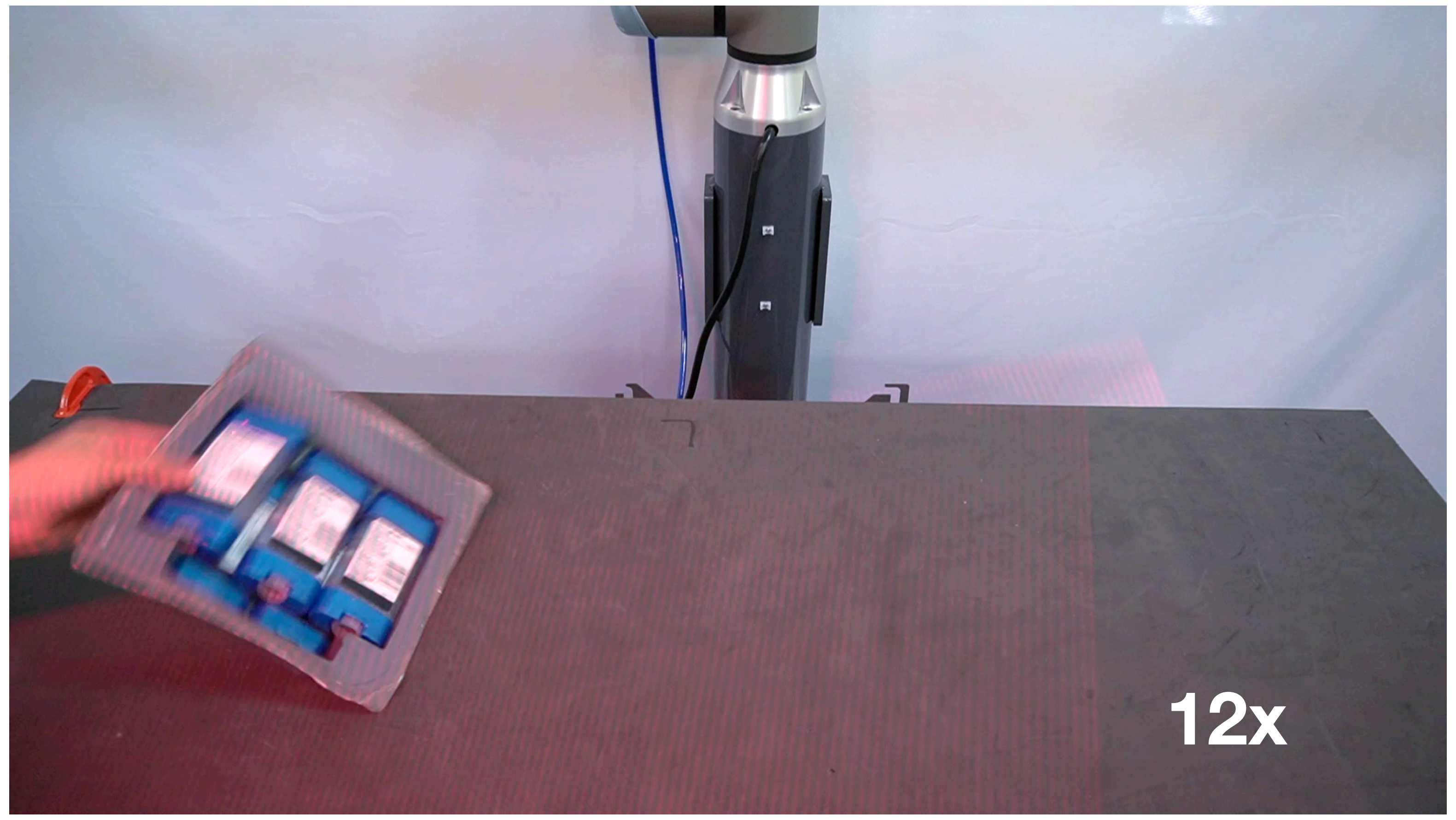
Results

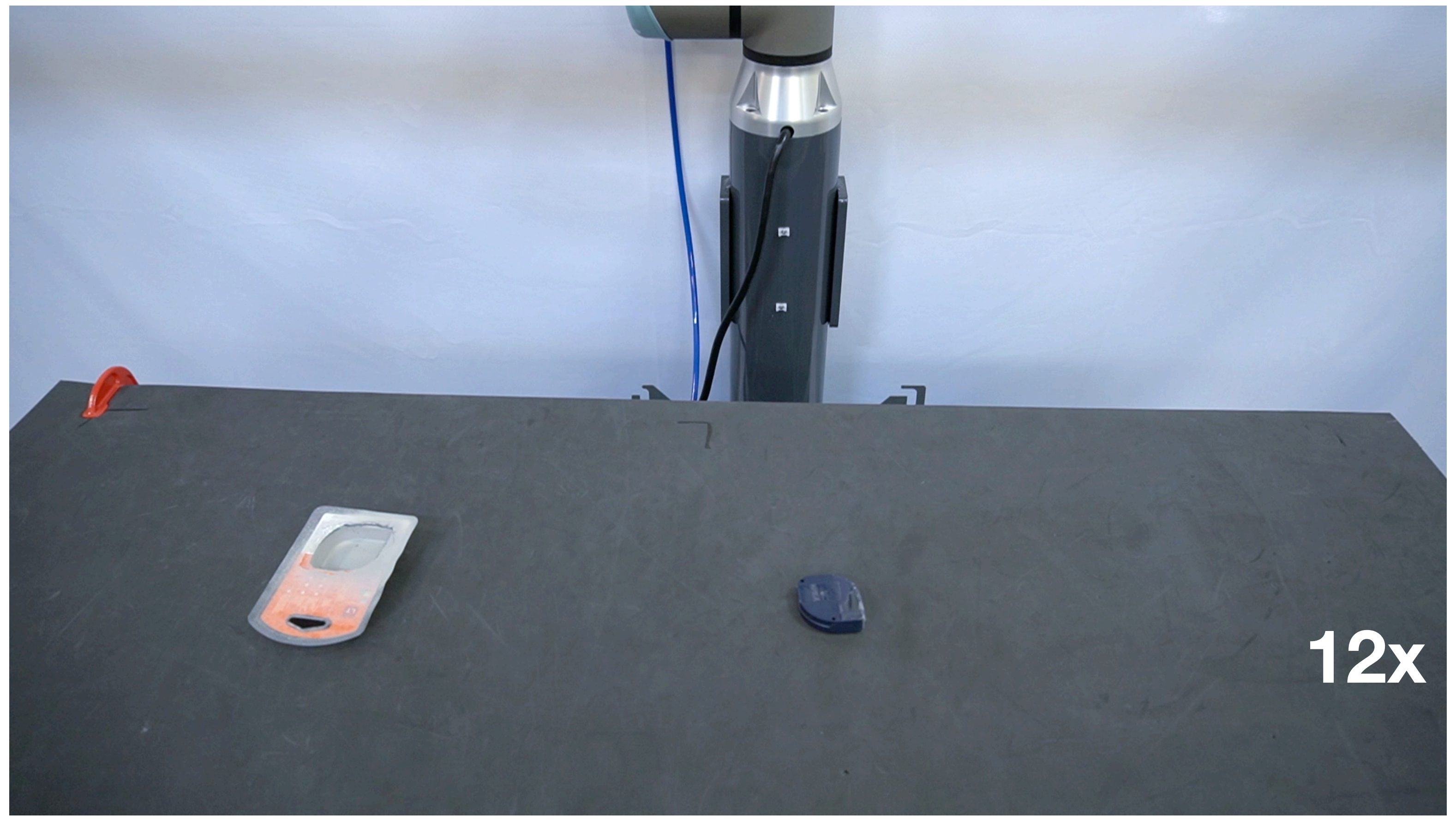
Varying Initial Position - 90%

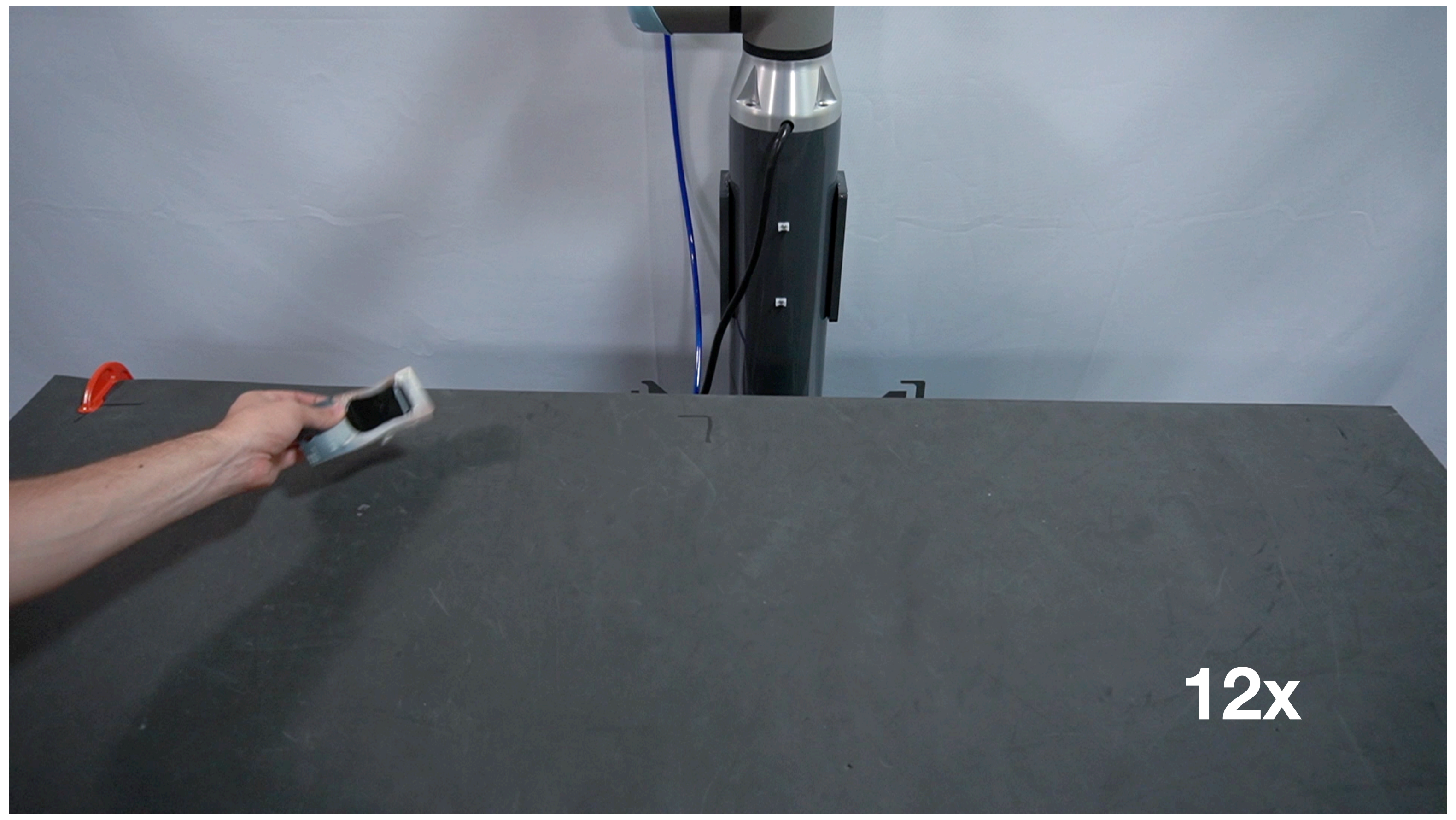
Trained on fixed single kit

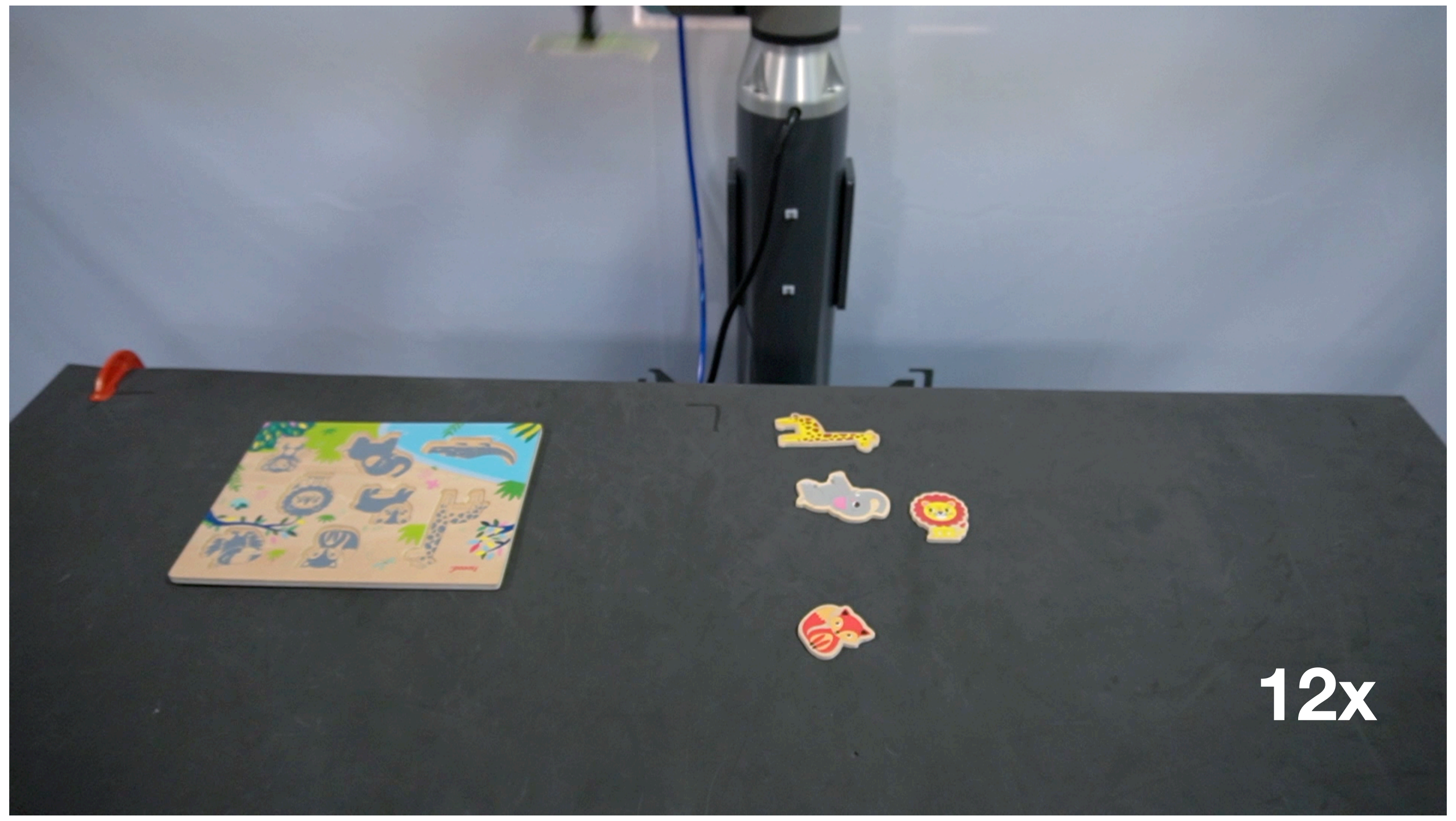


Different kit location and orientation



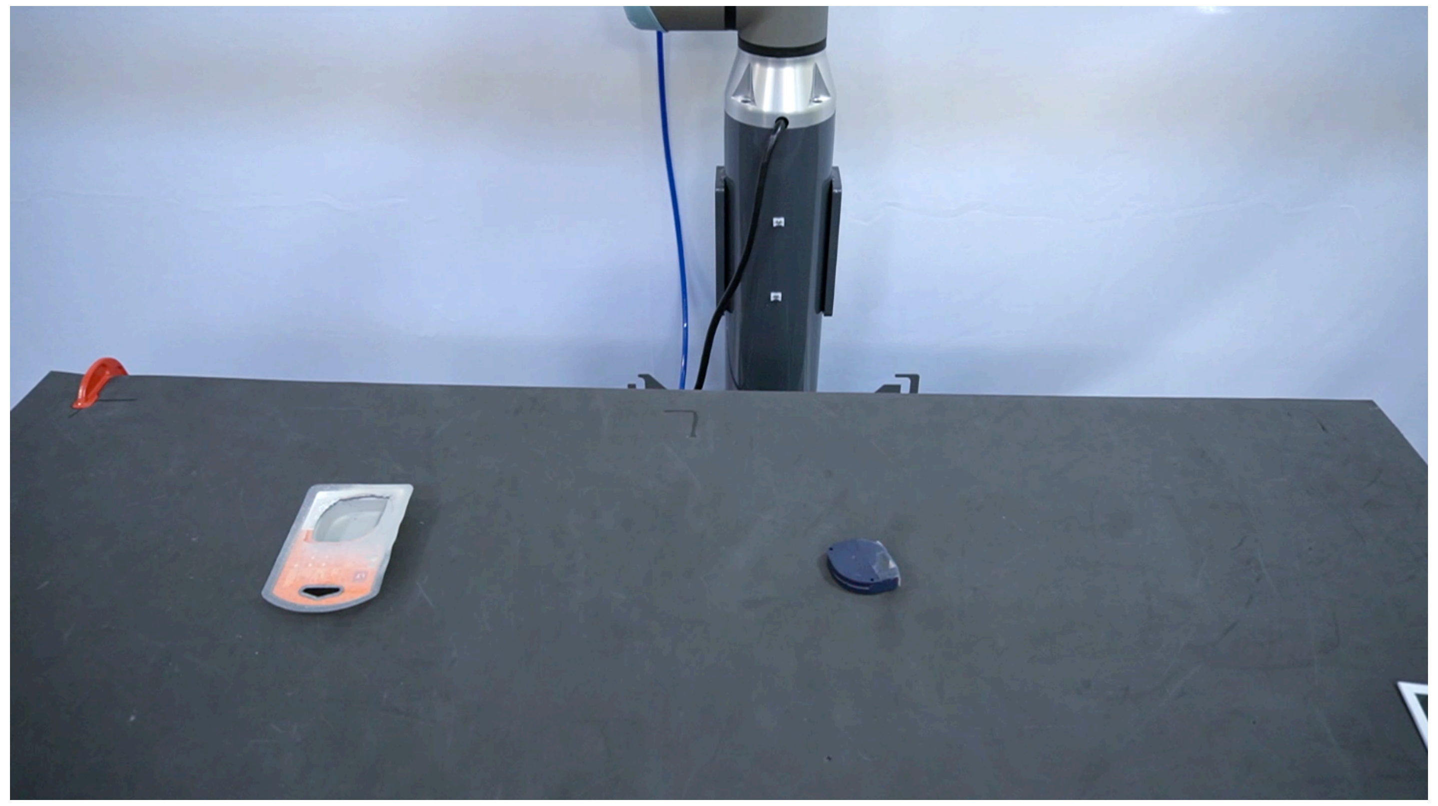




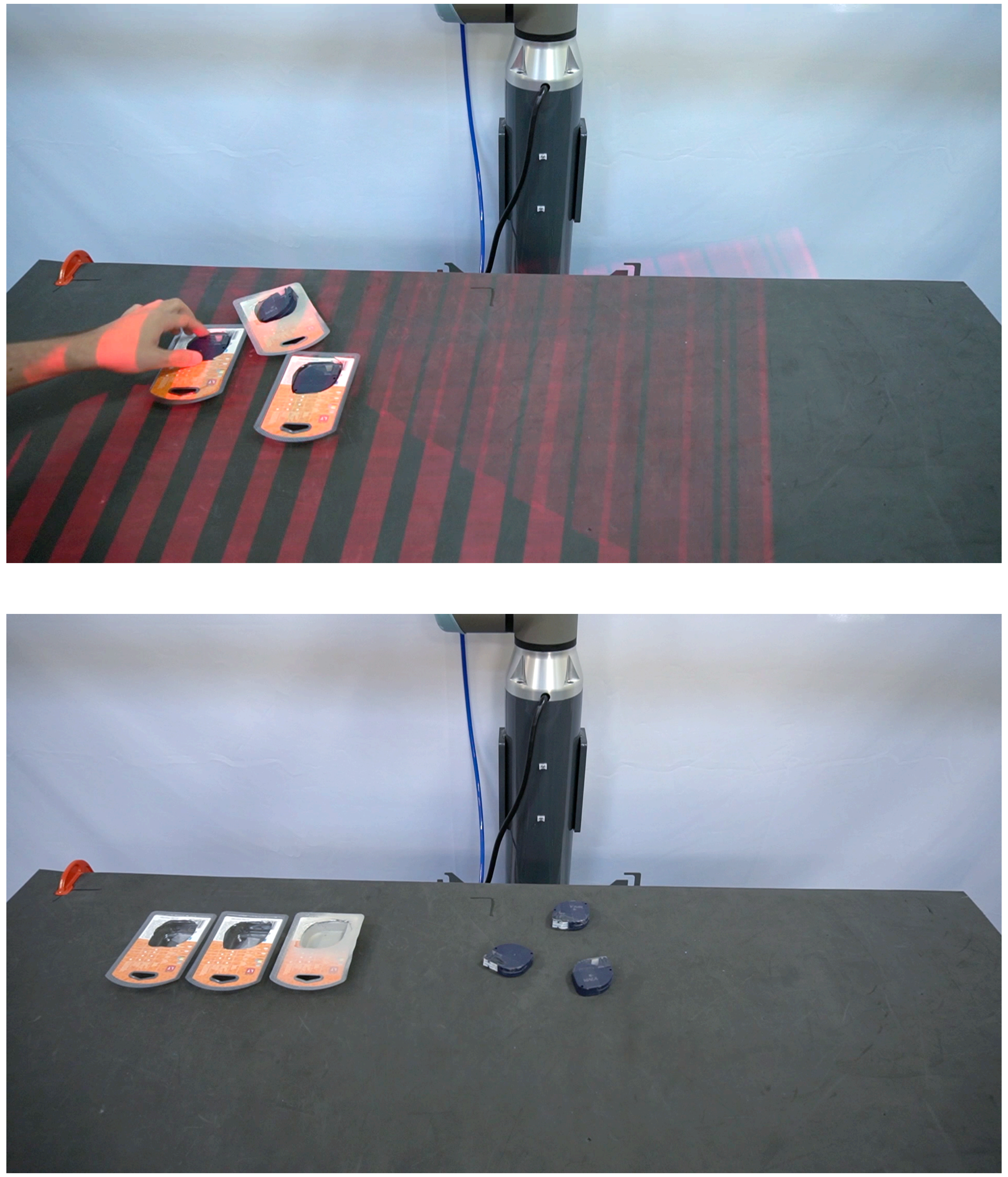


Generalization to Novel Settings - 94%

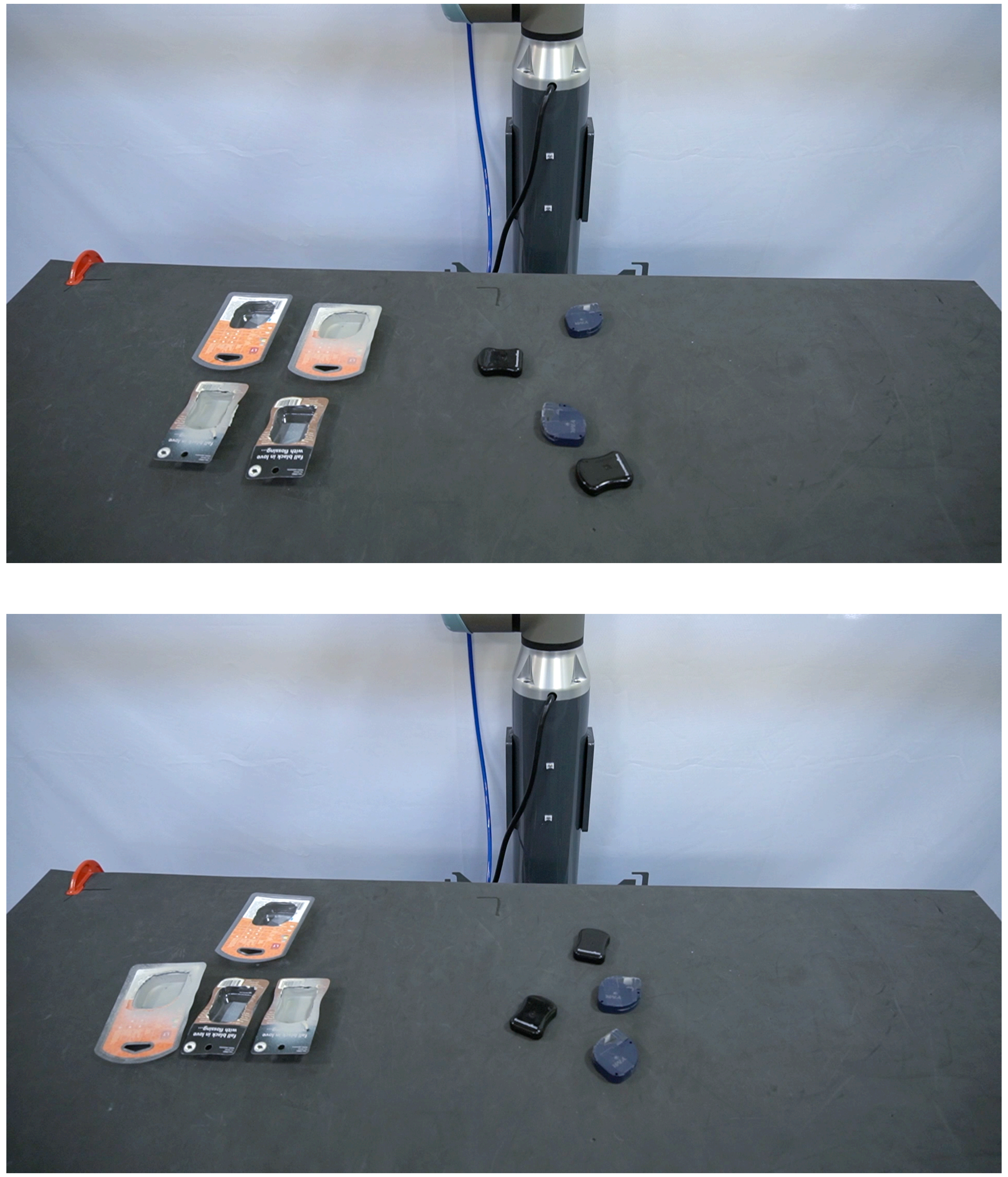
Trained on fixed single kit



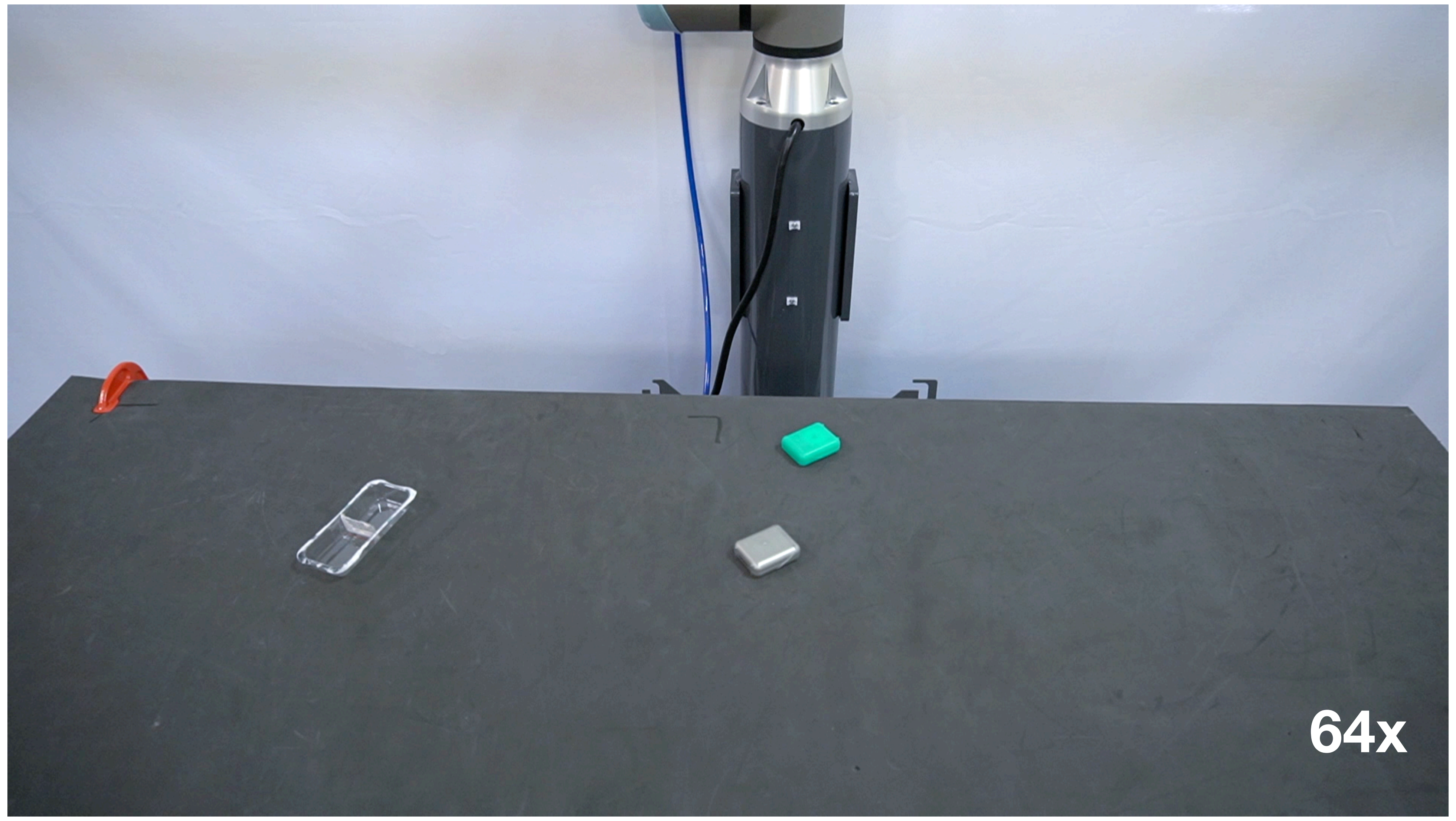
Multiple

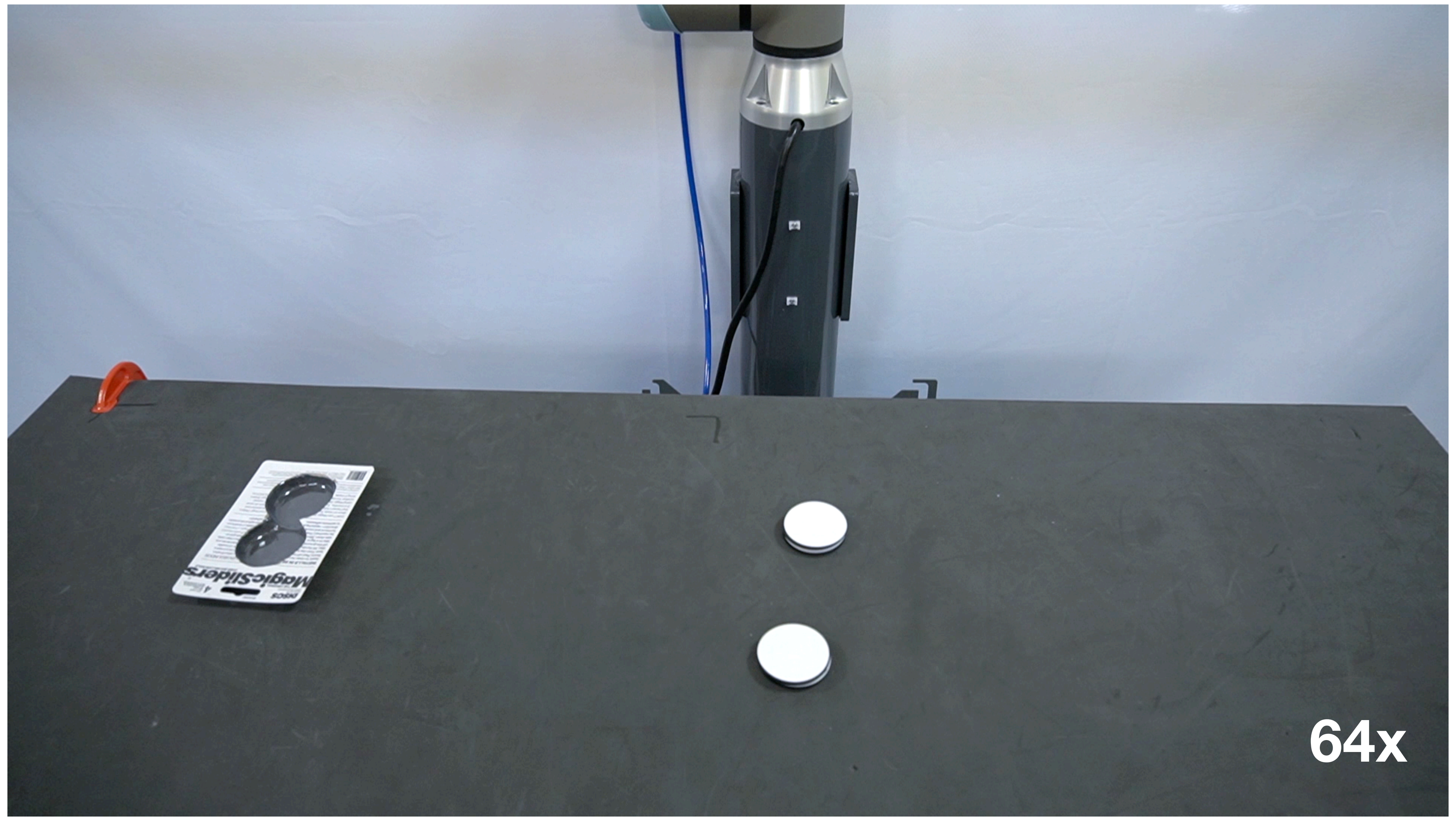


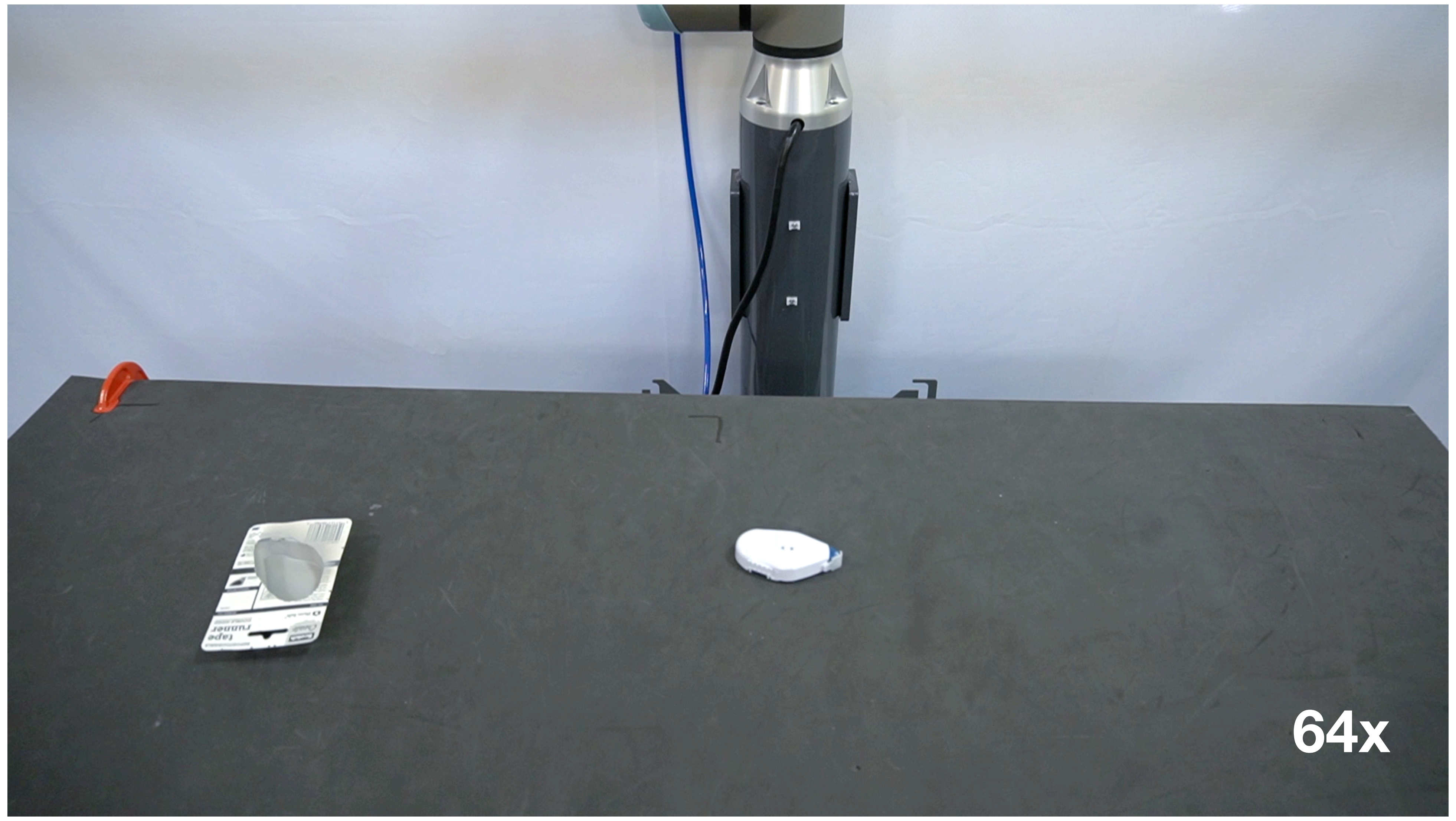
Mixture

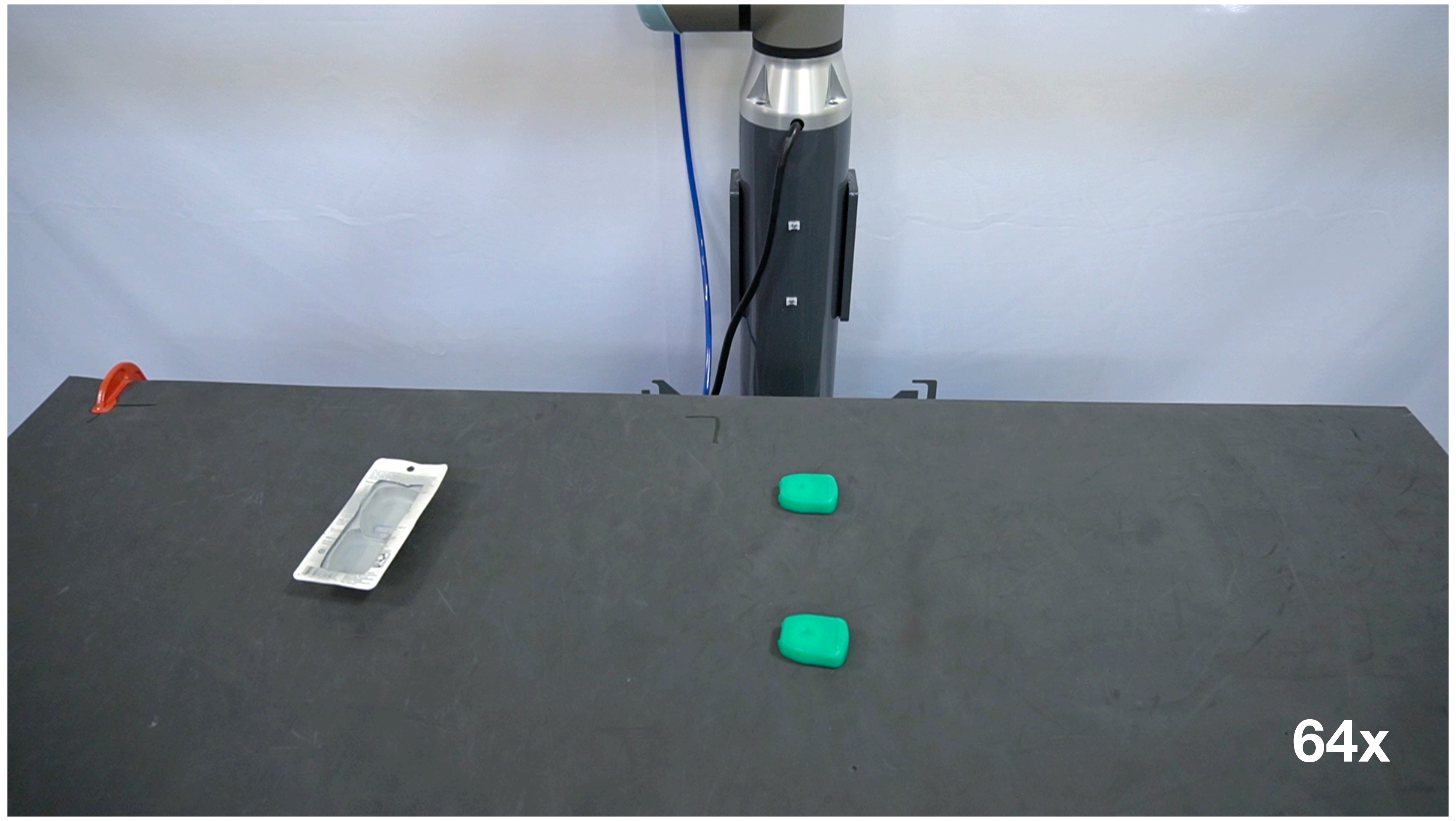


Generalization to Novel Kits - 86%

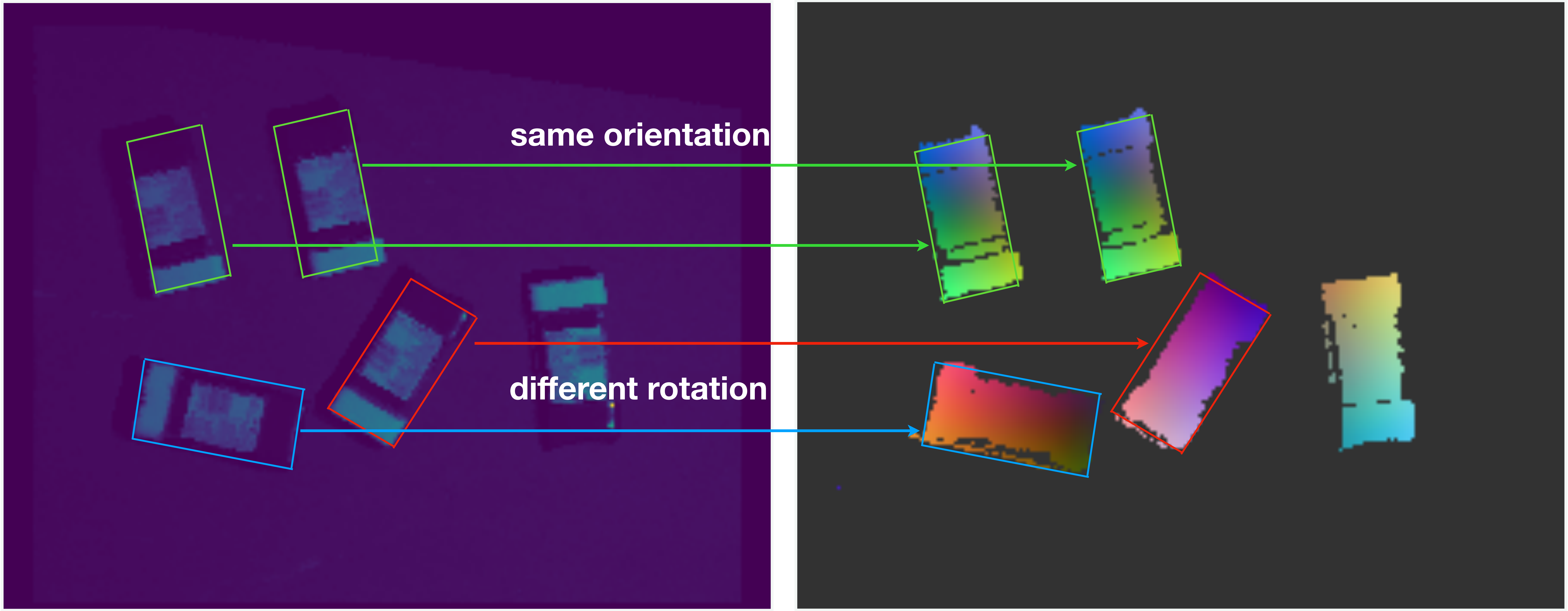






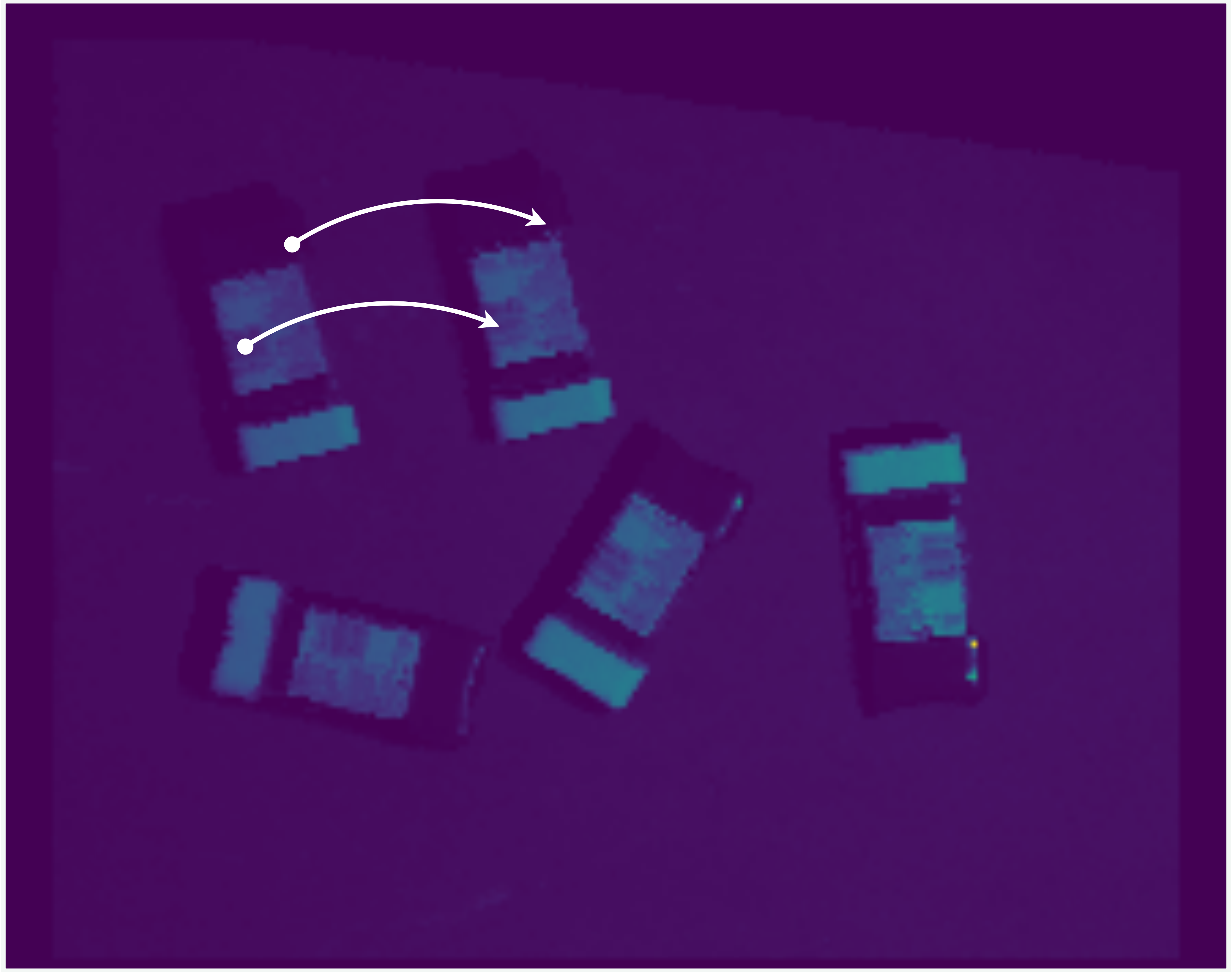


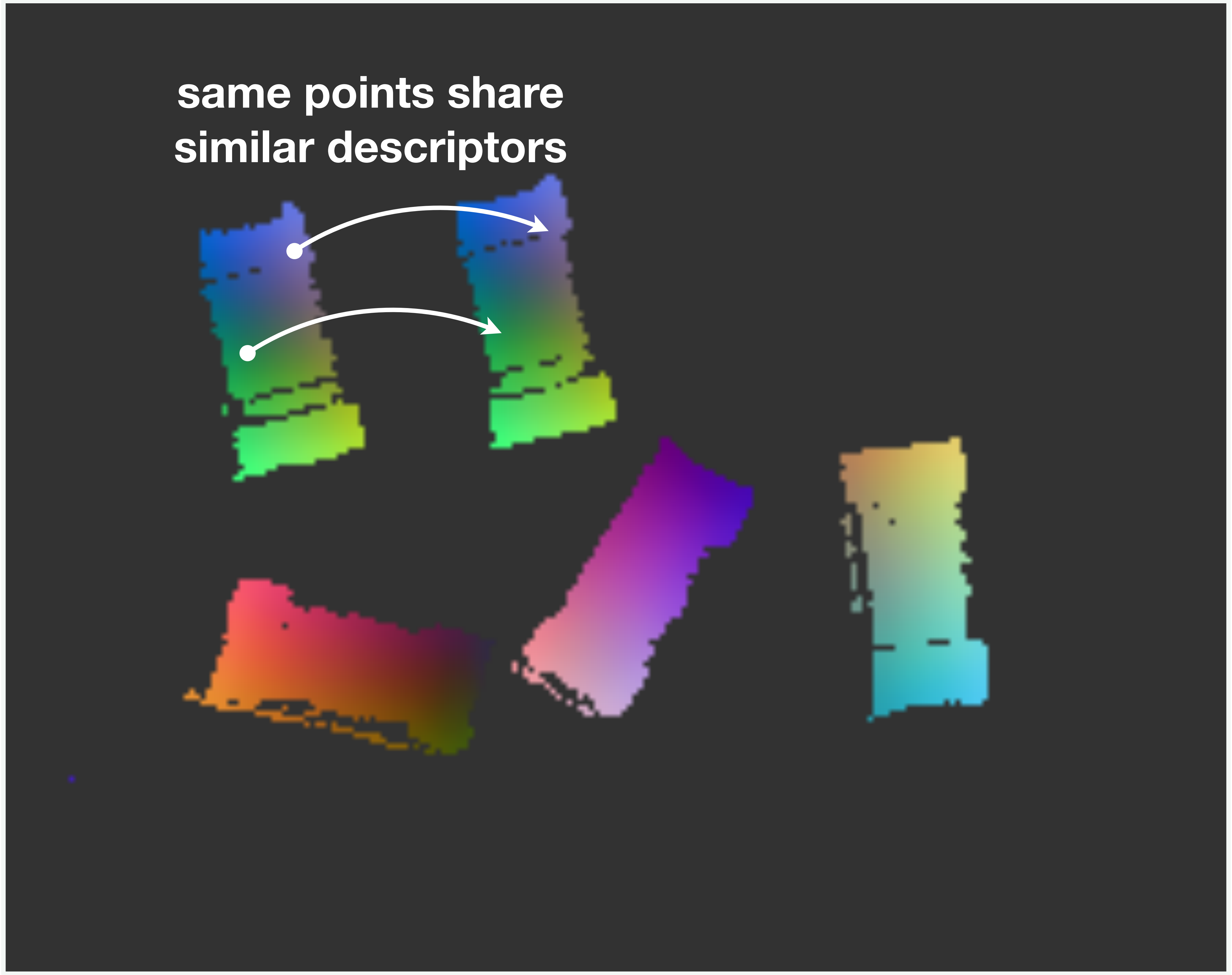
What does Form2Fit Learn?



descriptors encode object orientation

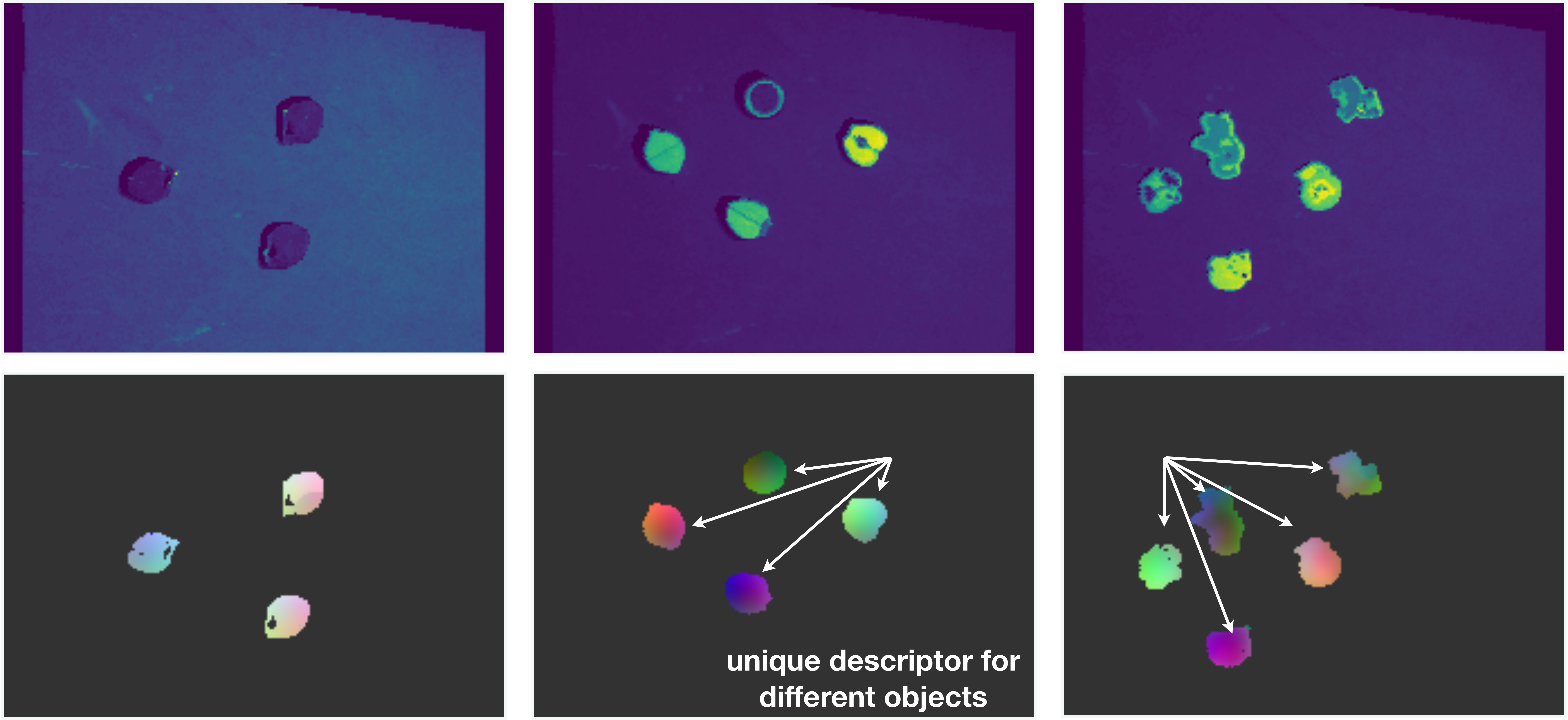
What does Form2Fit Learn?





descriptors encode spatial correspondence

What does Form2Fit Learn?



descriptors encode object identity

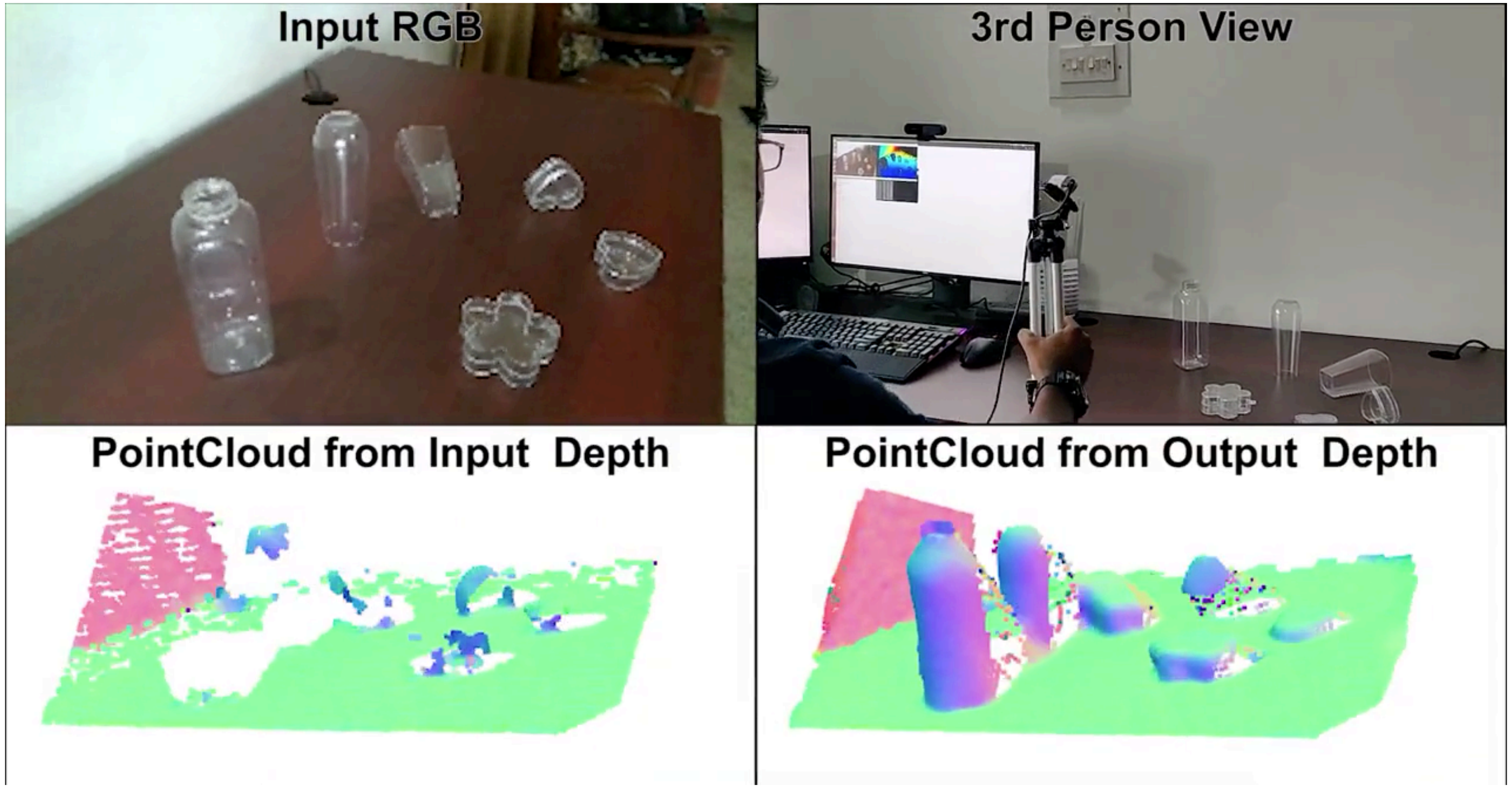
Limitation and Failure Cases

Limitation and Failure Cases

Top-down pick and place with 2D rotation

Transparent packages

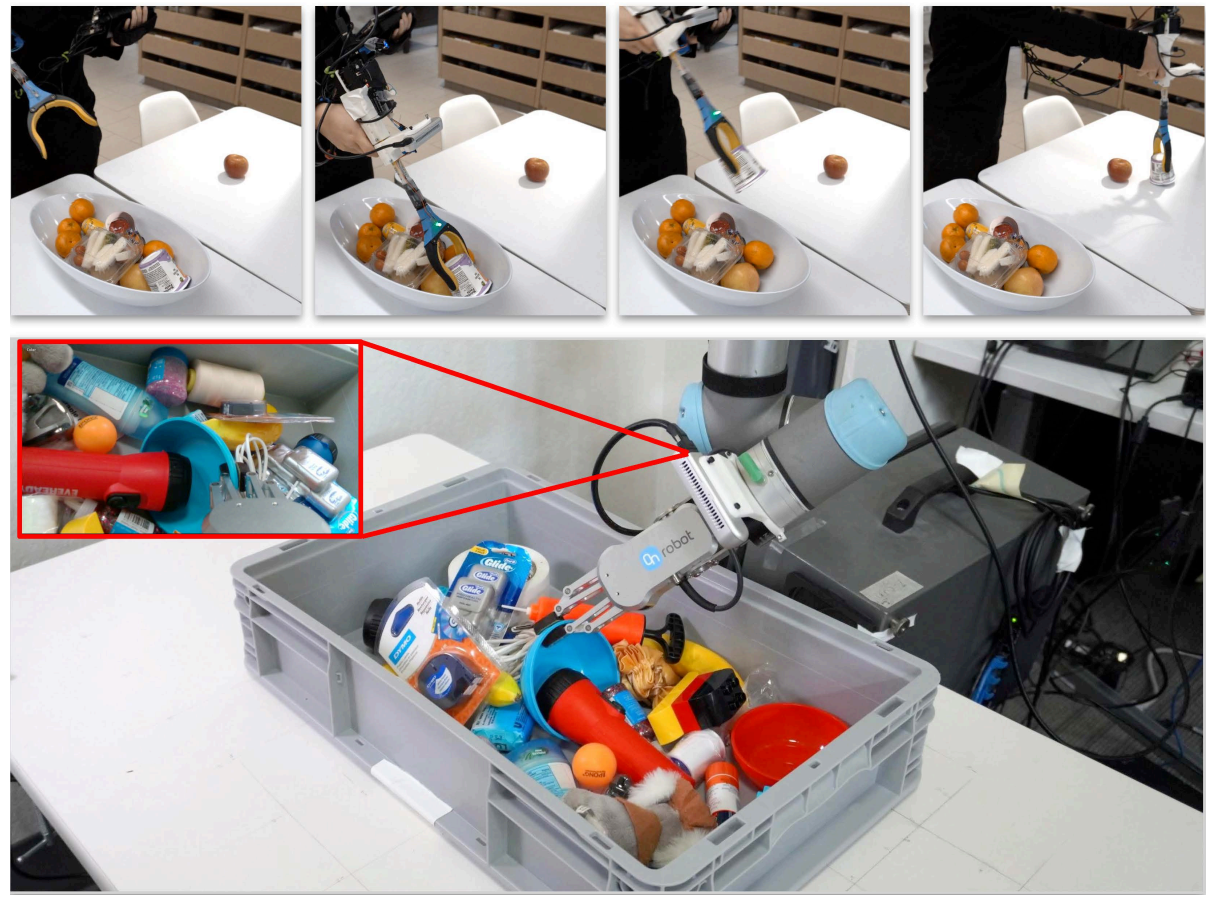
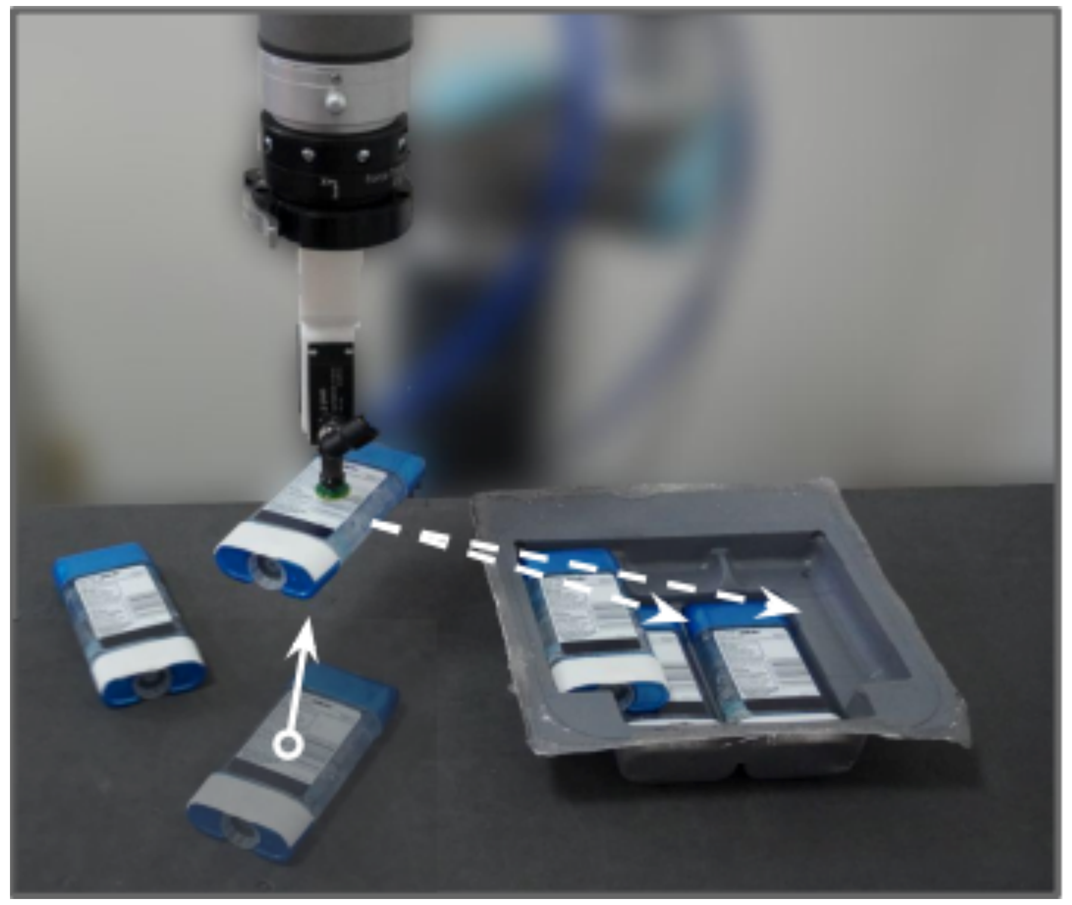
Limitation and Failure Cases



ClearGrasp: 3D Shape Estimation of Transparent Objects for Manipulation
https://sites.google.com/view/cleargrasp, ICRA 2020

Transparent packages

Generalizable Manipulation

	Generalizable Grasping: Grasp In the Wild	**Generalizable Assembly:** Form2Fit
		
Visual representation:	Action affordance Action-view representation	Shape correspondence for object assembly
Obtaining training data:	Low-cost human demonstration	Self-supervised disassembly for assembly

Acknowledgements

Reference:

[1] Grasping in the Wild: Learning 6DoF Closed-Loop Grasping from Low-Cost Demonstrations
Shuran Song, Andy Zeng, Johnny Lee, Thomas Funkhouser

[2] Category-Level Articulated Object Pose Estimation
Xiaolong Li, He Wang, Li Yi, Leonidas Guibas, A. Lynn Abbott, Shuran Song

[3] Form2Fit: Learning Shape Priors for Generalizable Assembly from Disassembly
Kevin Zakka, Andy Zeng, Johnny Lee, Shuran Song (ICRA 2020)

[4] ClearGrasp: 3D Shape Estimation of Transparent Objects for Manipulation
Shreeyak S. Sajjan, Matthew Moore, Mike Pan, Ganesh Nagaraja, Johnny Lee, Andy Zeng, Shuran Song (ICRA 2020)

[5] DensePhysNet: Learning Dense Physical Object Representations via Multi-step Dynamic Interactions
Zhenjia Xu, Jiajun Wu, Andy Zeng, Joshua Tenenbaum, Shuran Song (RSS 2019)

[6] Robotic Pick-and-Place of Novel Objects in Clutter with Multi-Affordance Grasping and Cross-Domain Image Matching
A. Zeng, S. Song, K. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazeli, F. Alet, N. C. Dafle, R. Holladay, I. Morona, P. Q. Nair, D. Green, I. Taylor, W. Liu, T. Funkhouser, A. Rodriguez (ICRA2018)

[7] Multi-view Self-supervised Deep Learning for 6D Pose Estimation in the Amazon Picking Challenge
A. Zeng, K.T. Yu, S. Song, D. Suo, E. Walker Jr., A. Rodriguez, and J. Xiao (ICRA2017)

NSF

Google AI

amazon robotics

COLUMBIA | ENGINEERING
The Fu Foundation School of Engineering and Applied Science

Thank You !