NVIDIA

MODEL-FREE ROBOT OBJECT MANIPULATION

Arsalan Mousavian

NVIDIA Research - Seattle Robotics Lab (SRL)

TODAY’S HOT ROBOTIC APPLICATION AREAS

Moving from A to B without collision

Self-Driving Cars

Warehouse Fulfillment

Robot Delivery

Hotel Hospitality

Healthcare

Inventory Management

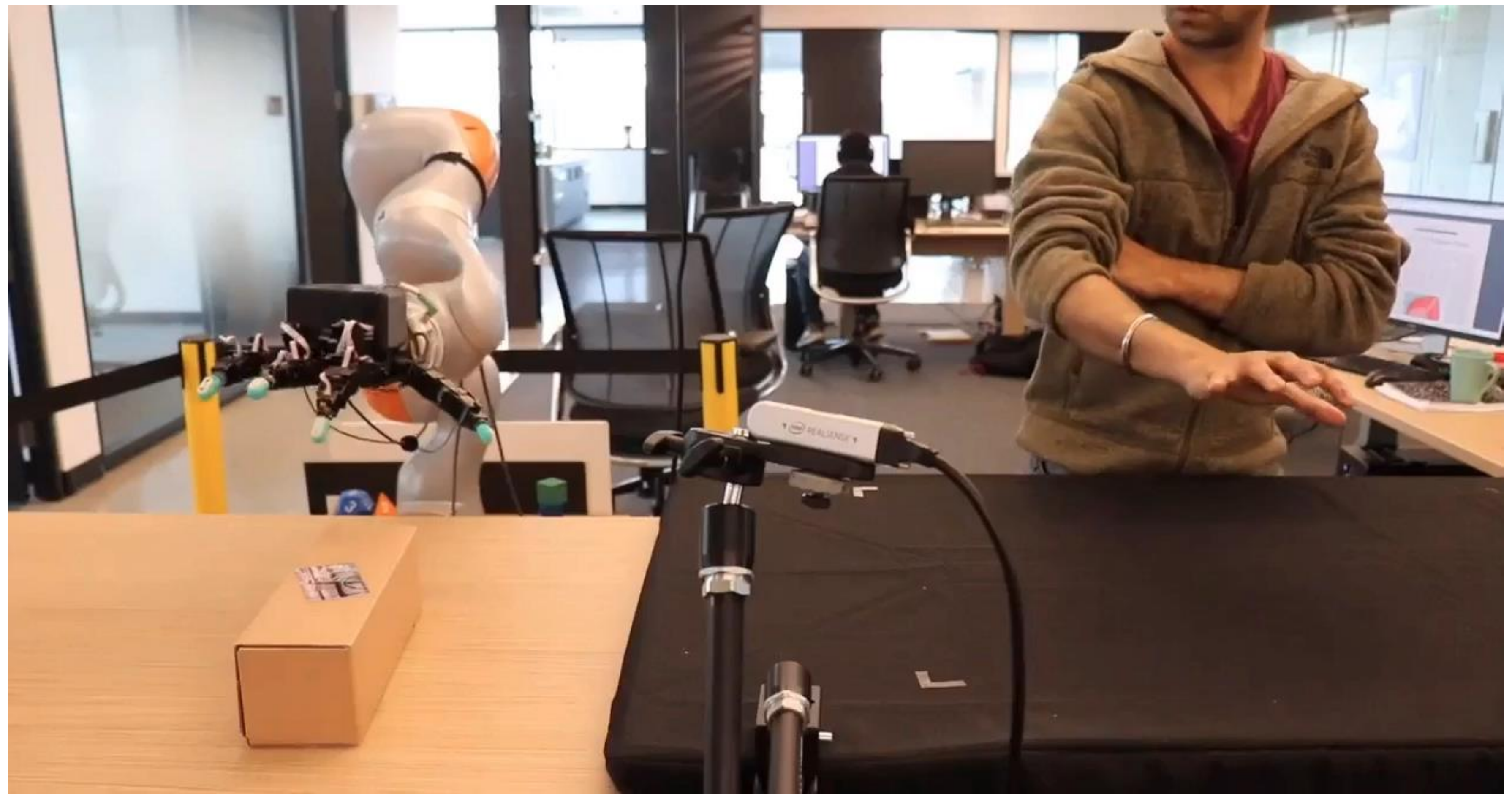
NVIDIA

CAPABLE ROBOT HARDWARE

Robot hardware is quite capable.

Boston Dynamics

Shadow Hand



DexPilot, NVIDIA

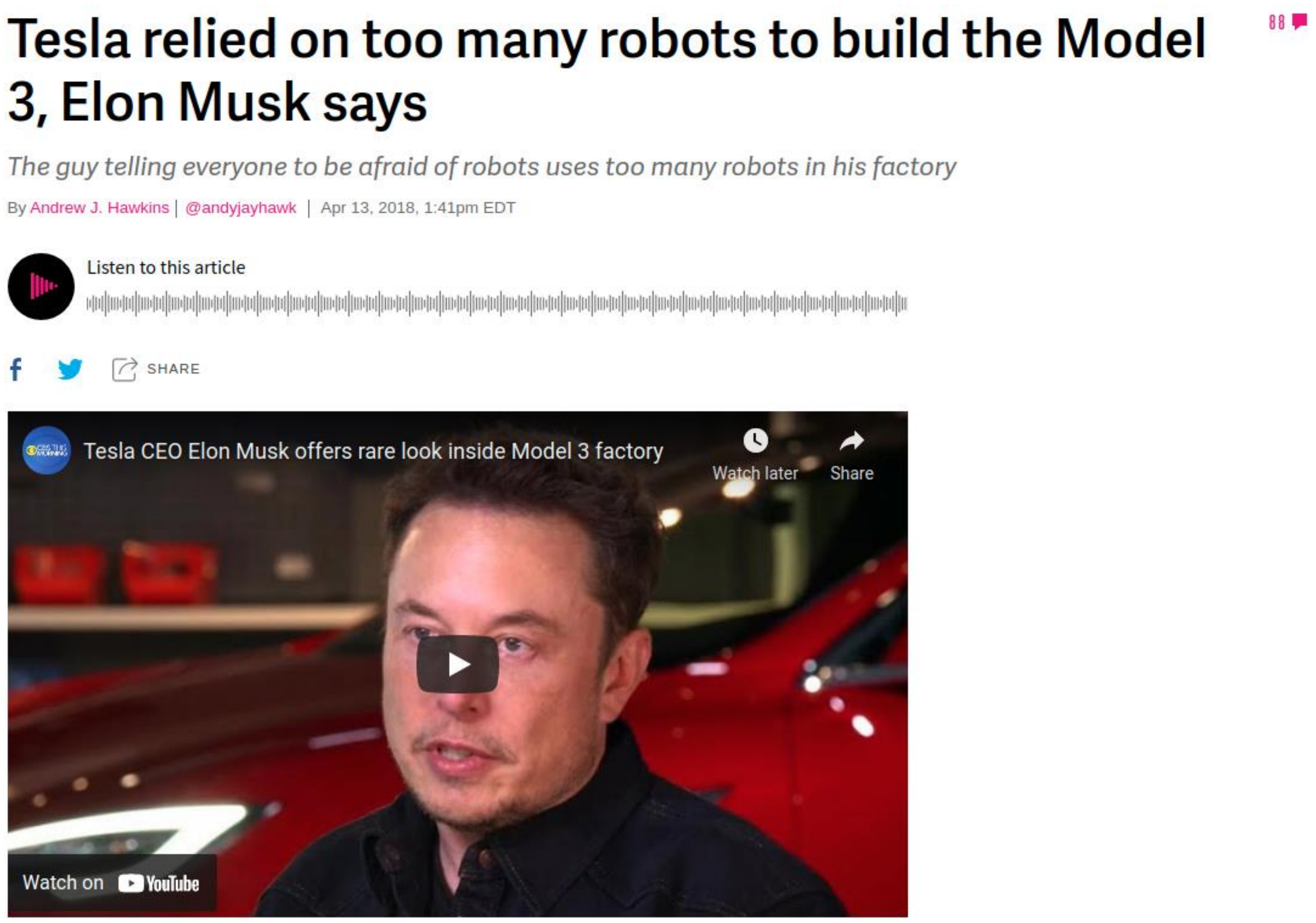
Handy Robot, Samsung

INDUSTRIAL ROBOT MANIPULATION

Robots move blindly with high accuracy in controlled environments at Factories.

STATE OF ROBOT MANIPULATION

Reality check: Gap between robot hardware capability and manipulation capability


Tesla relied on too many robots to build the Model 3, Elon Musk says

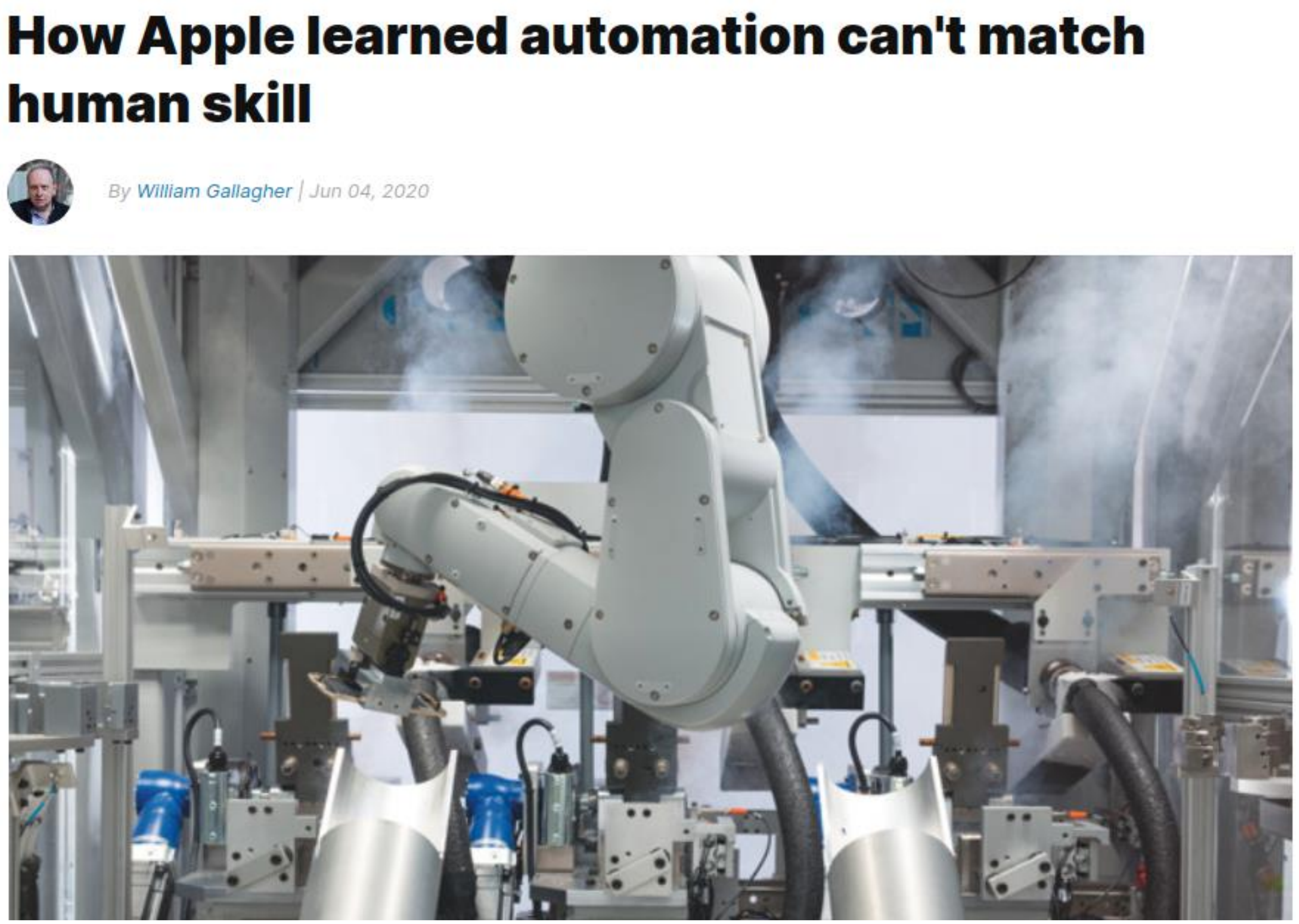
The guy telling everyone to be afraid of robots uses too many robots in his factory

By Andrew J. Hawkins | @andyjayhawk | Apr 13, 2018, 1:41pm EDT

Listen to this article

SHARE

The Verge, April 2020


How Apple learned automation can't match human skill

By William Gallagher | Jun 04, 2020

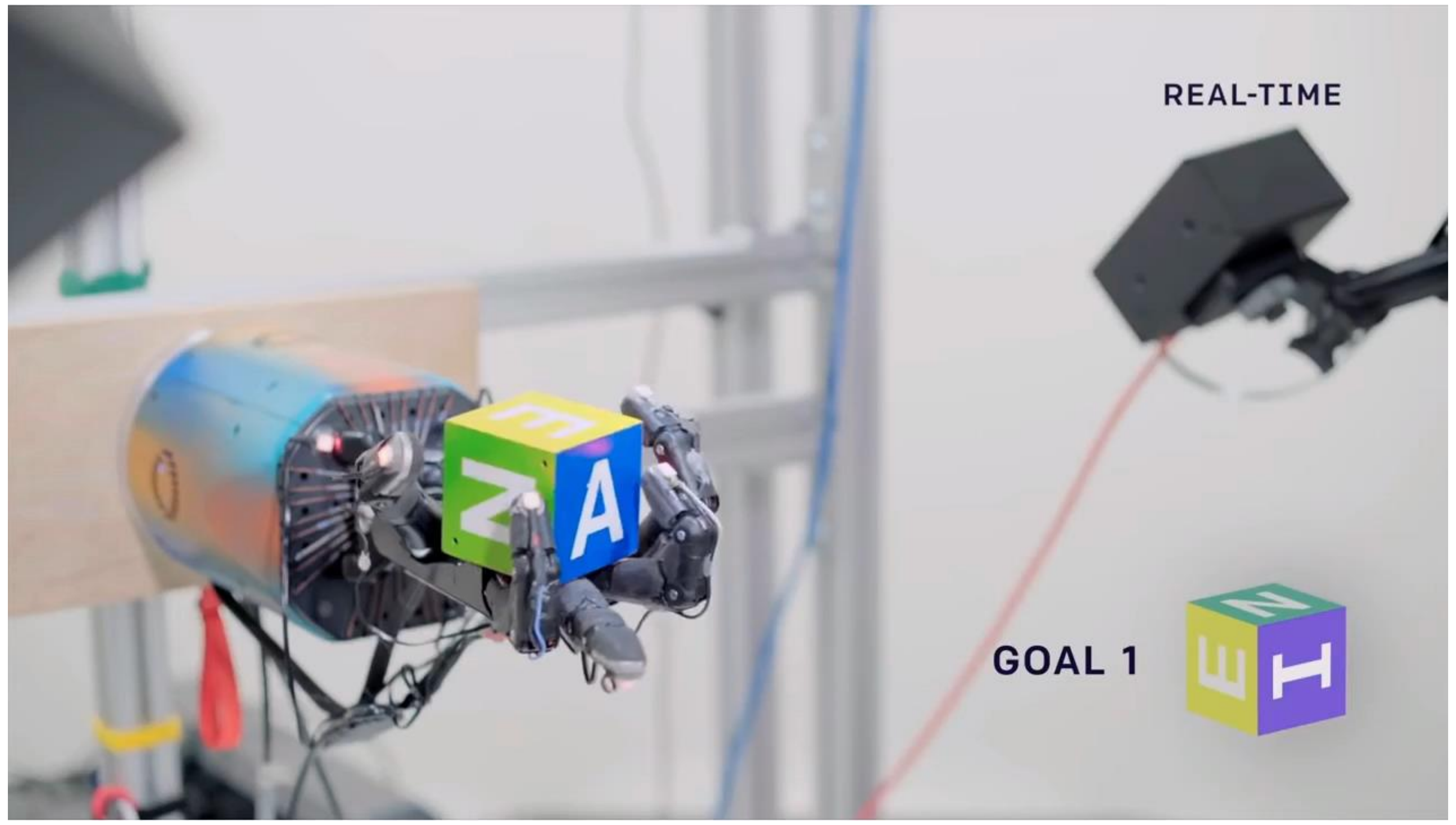
Apple Insider, June 2020

Data driven methods are promising methods to fill the gap between hardware capabilities and manipulation skills

ROBOT MANIPULATION METHODS

Object-agnostic model-free methods: From raw sensory data to actions

[Levine et al, 2016] [Kalashnikov et al, 2018]



[OpenAI et al, 2018]

- Advantages:
  - Minimal Assumptions about the world.
  - Minimal engineering required.
  - All you need is to collect data.

- Limitations:
  - Not sample efficient. Needs a lot of data.
  - Limited generalization to new environments.
  - Short horizon tasks.
  - No Compositionality.

ROBOT MANIPULATION METHODS

Model based methods: Assume known 3D models for everything, estimate the state of the world, do planning on that.

6D Pose Estimation
[Deng et al, RSS 2019]

Task: “Cook” Mustard and Tomoato Soup
[Garret et al, ICRA 2021]

Advantages:

- Handles long horizon tasks with theoretical guarantees.
- Modularity.
- Planning part does not need any training at all.

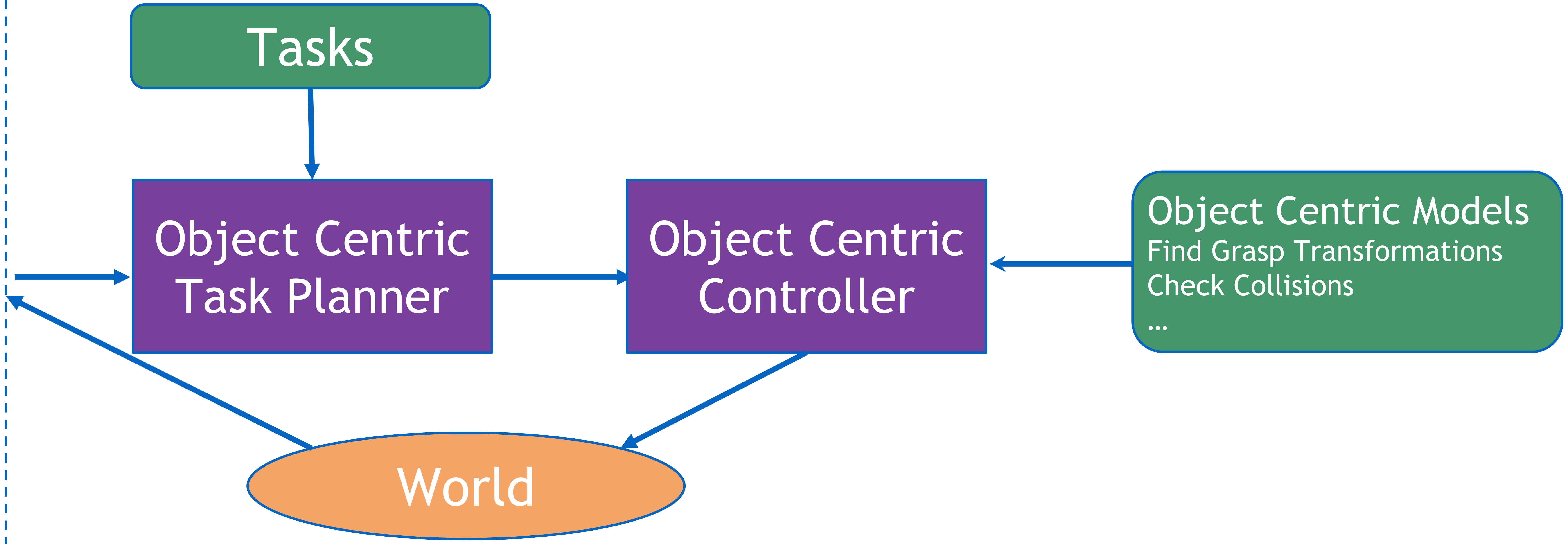
Limitations:

- Limited to only known objects.
- State estimation errors accumulate.
- Today’s pose estimation methods for large number of object are not scalable.
- Lot of implementation is needed for planning.

THIS TALK

Model-free Object Centric Models

Image + Instance Segmentation



OVERVIEW

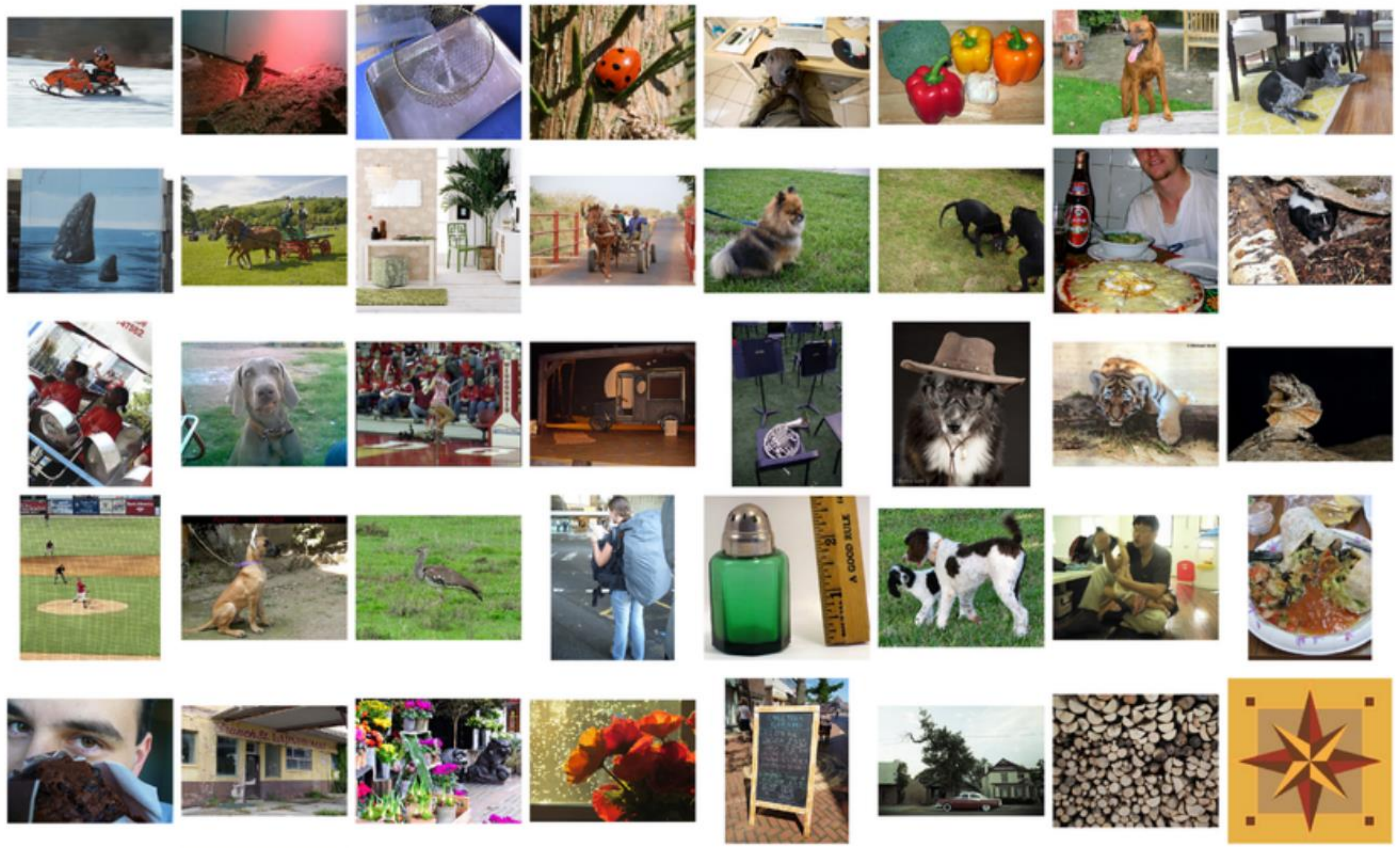
- Object Instance Segmentation
- Grasping Unknown Objects
- Collision checking and motion planning for unknown objects and scenes
- Image based object rearrangement

UNKNOWN OBJECT INSTANCE SEGMENTATION

LEARNING THE CONCEPT OF "OBJECTS"

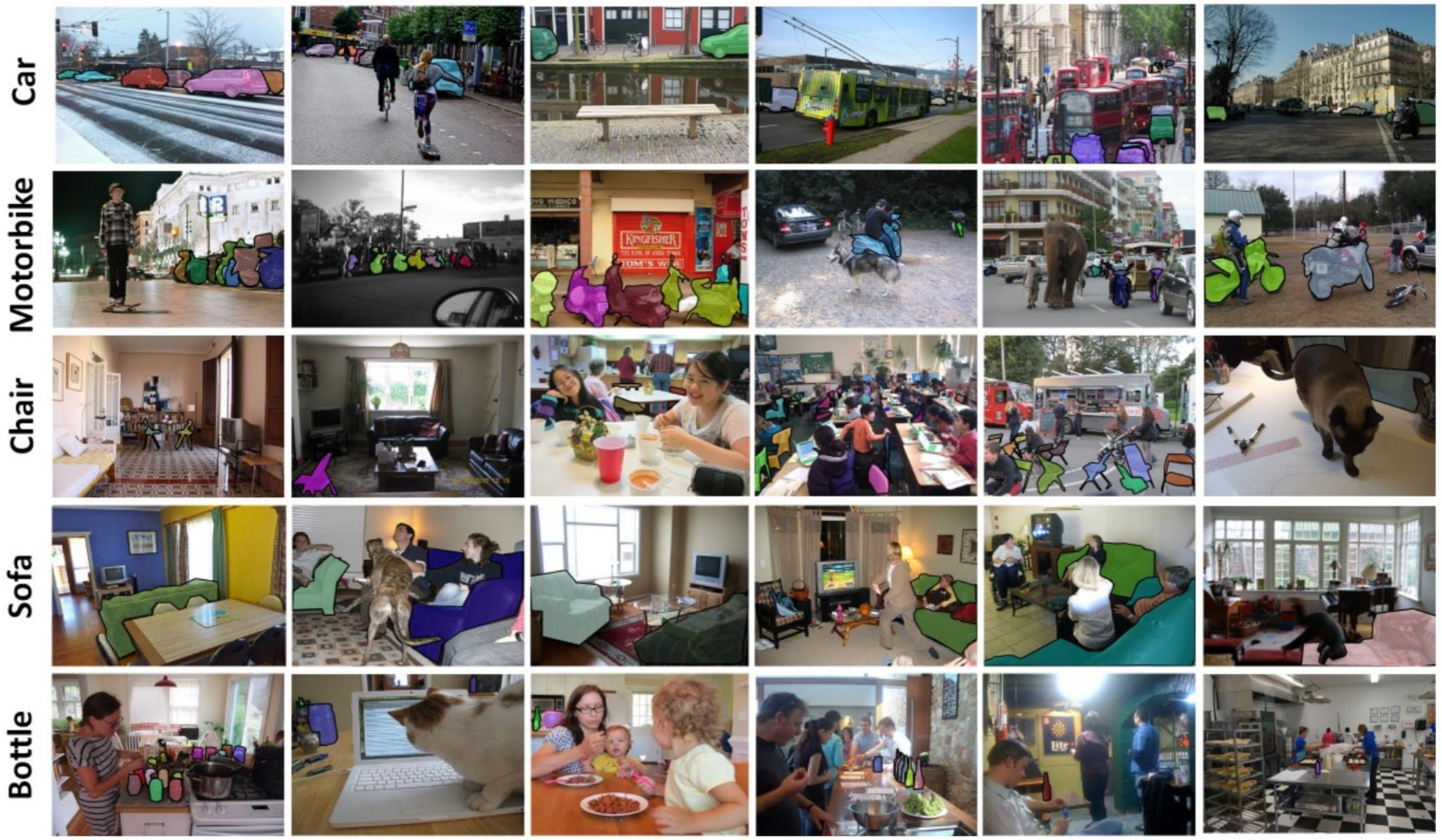
Learning from data

Internet Images



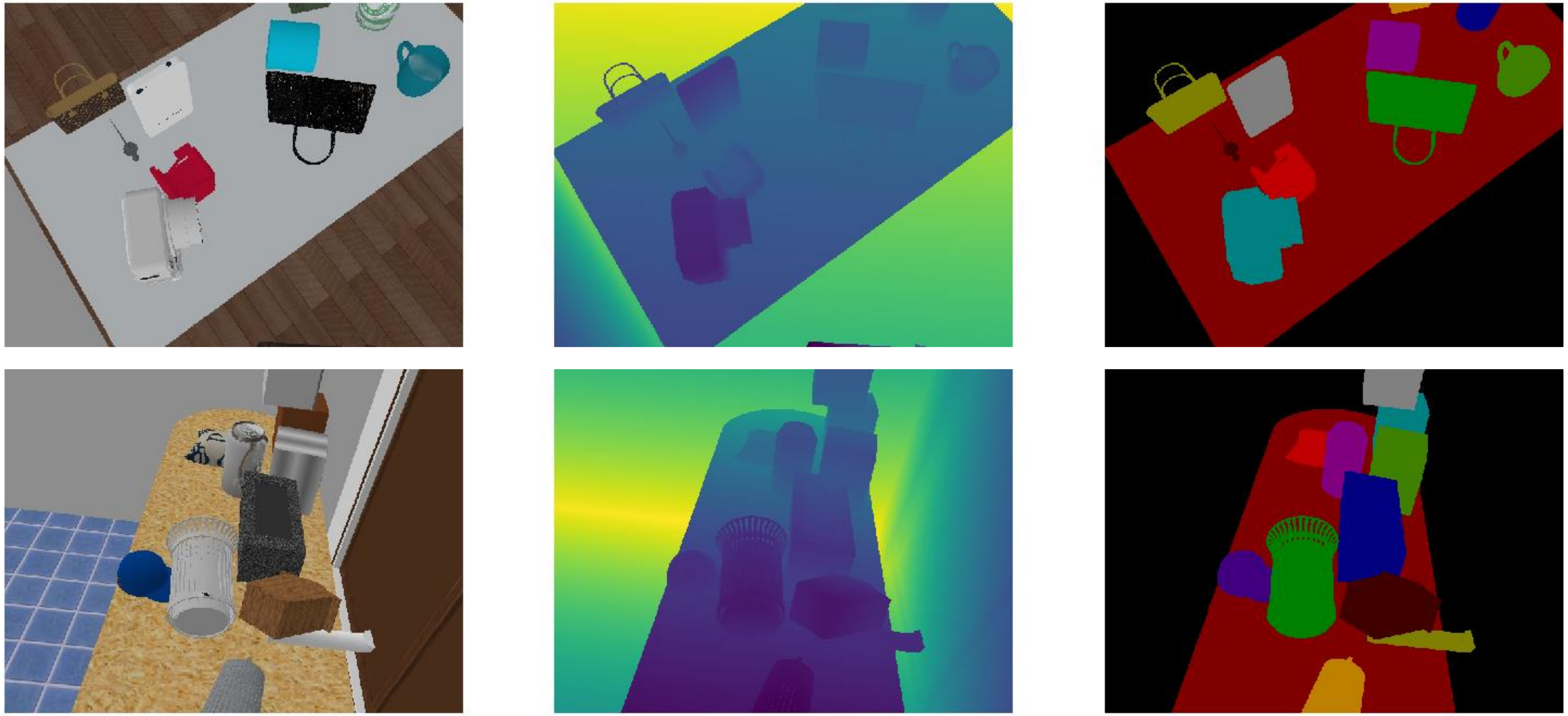
No Segmentation Labels

Vision Datasets



Not suitable for table-top

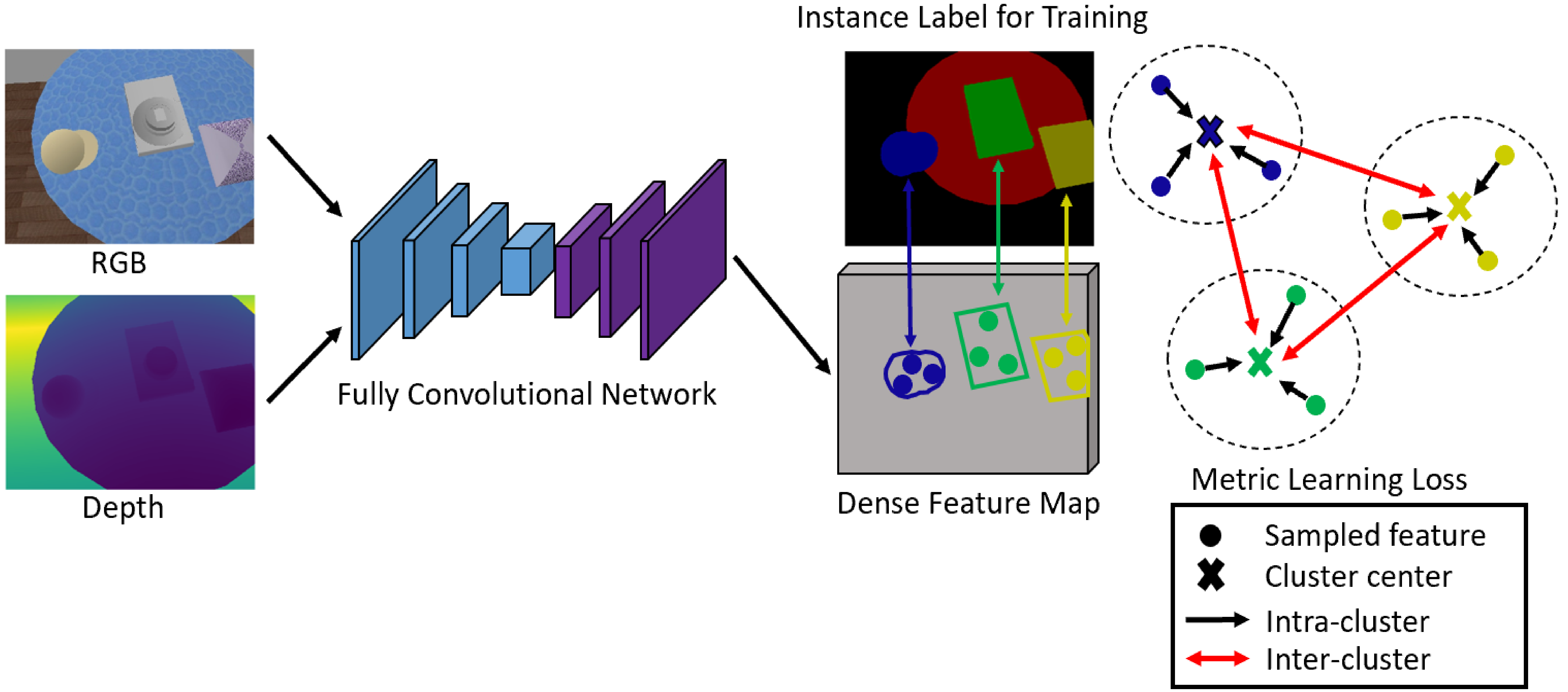
LEARNING FROM SYNTHETIC DATA



RGB | Depth | Instance Label

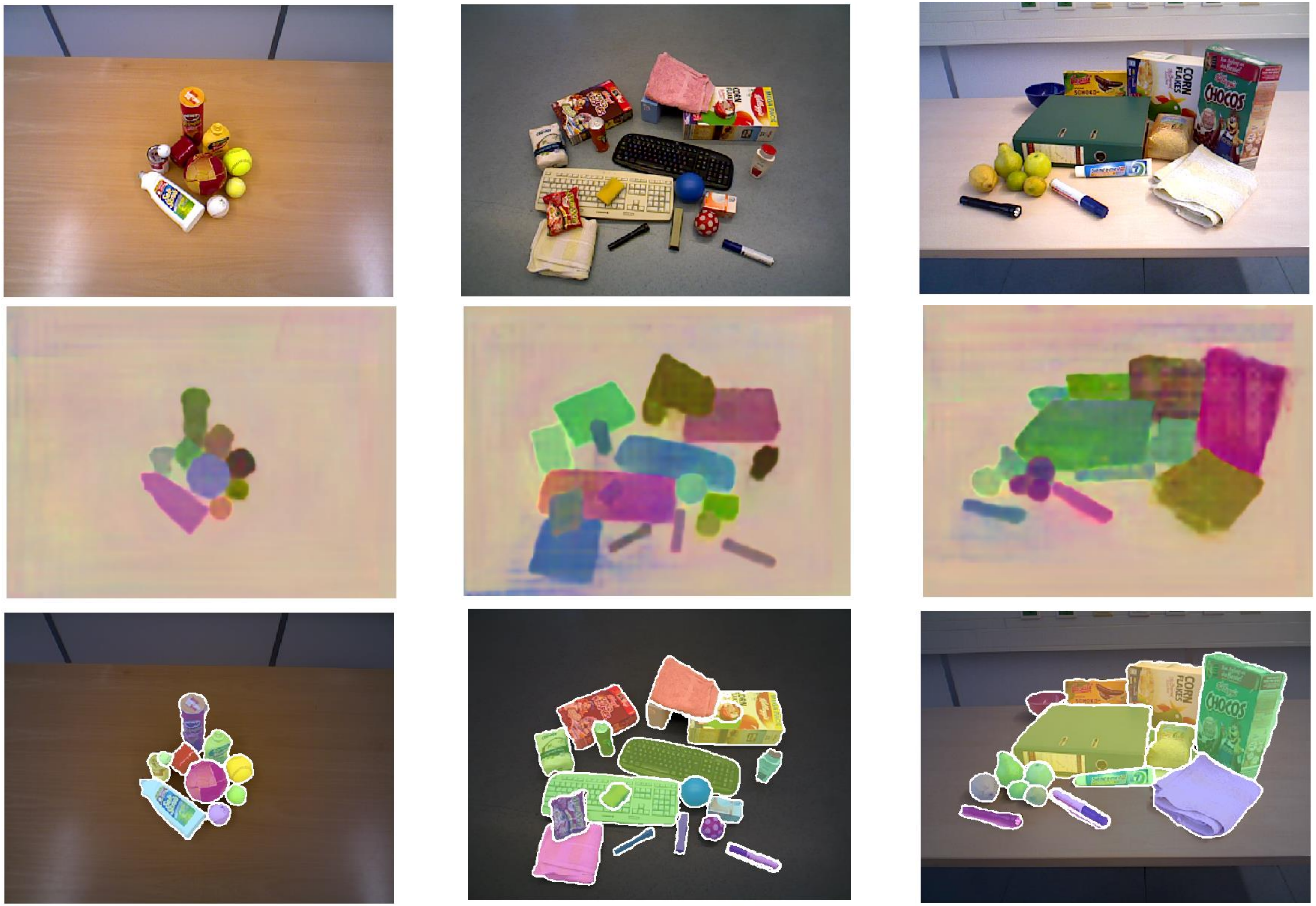
LEARNING RGB-D FEATURE EMBEDDINGS

The model predicts feature embeddings for each pixel in the RGB-D image



GitHub https://github.com/NVlabs/UnseenObjectClustering [Xiang-Xie-**Mousavian**-Fox, CoRL 2020]

NVIDIA


CORN FLAKES
CHOCOS

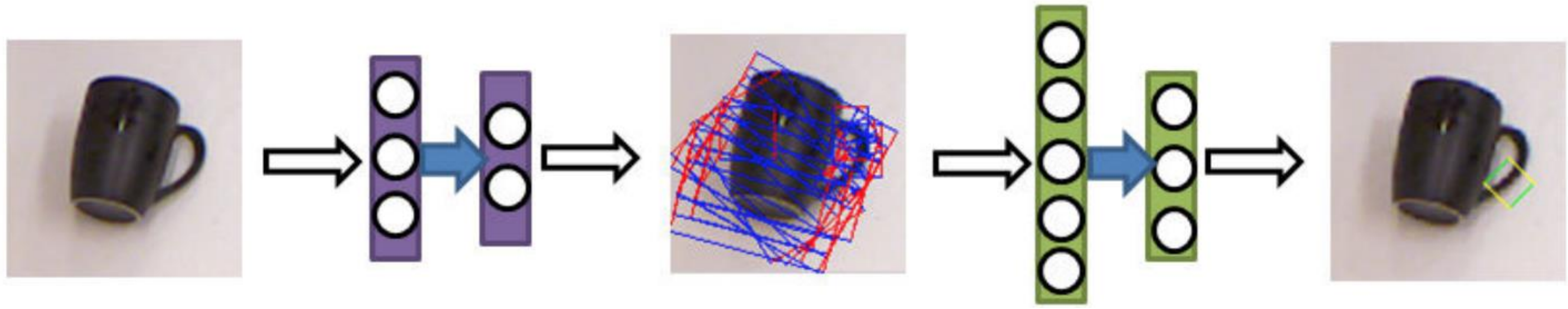
GRASPING UNKNOWN OBJECTS

PLANAR GRASPING

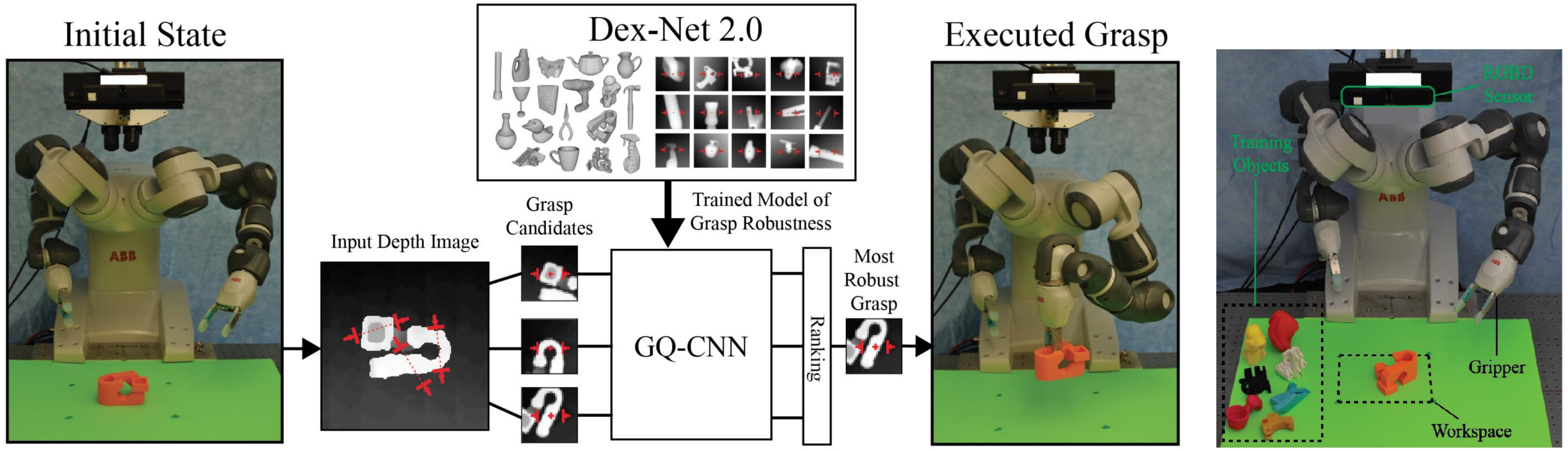
Representing grasps by oriented rectangles

- Camera needs to be roughly perpendicular to the scene.
- Effective workspace of grasping becomes limited.
- Does not handle grasping objects from enclosed areas such as shelves and cubbies.

Kalashnikov et al, CoRL 2018



Lenz et al, RSS 2013

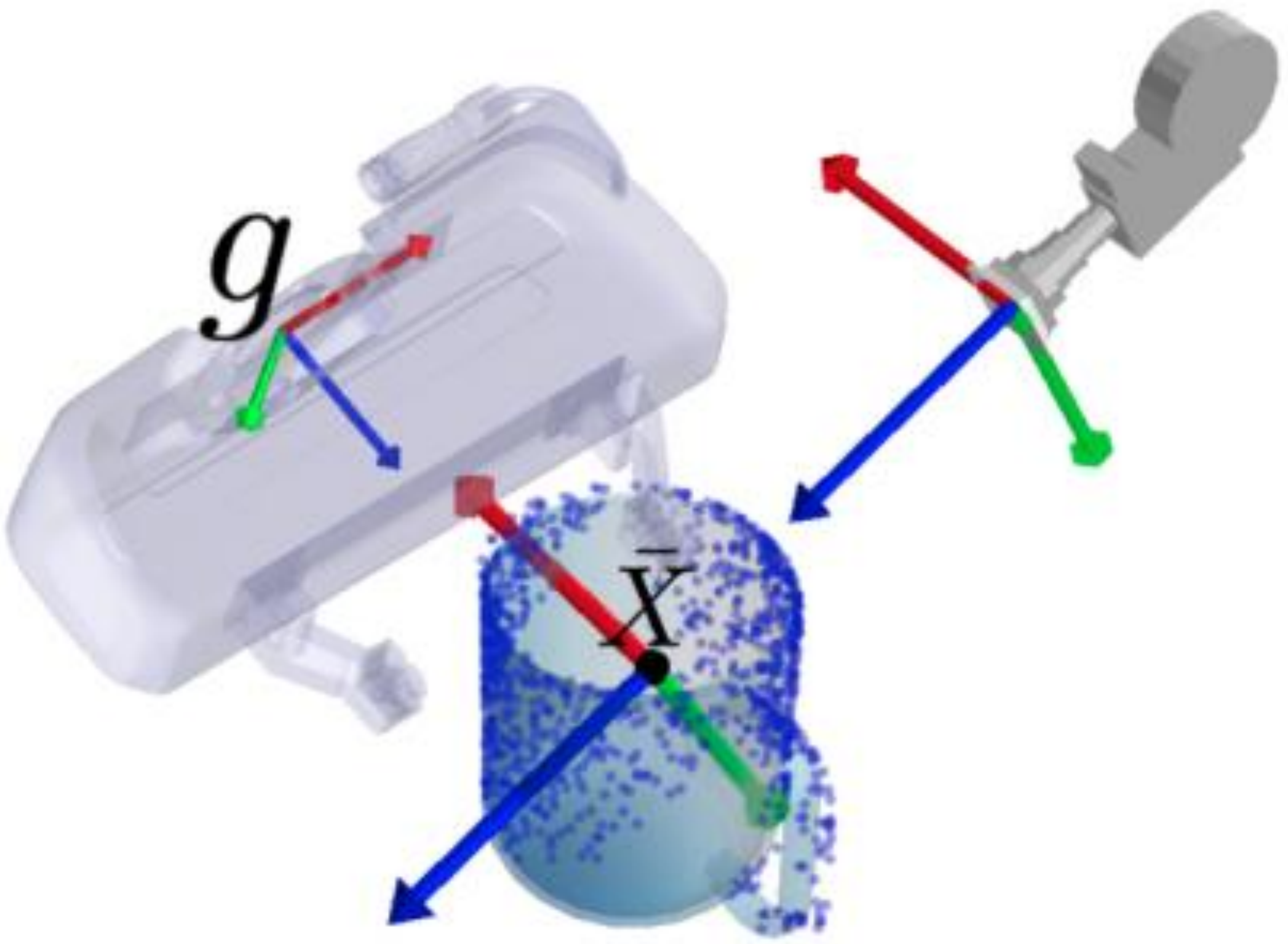


Mahler et al, RSS 2017

6-DOF GRASPING

Representing grasps by SE(3) transforms

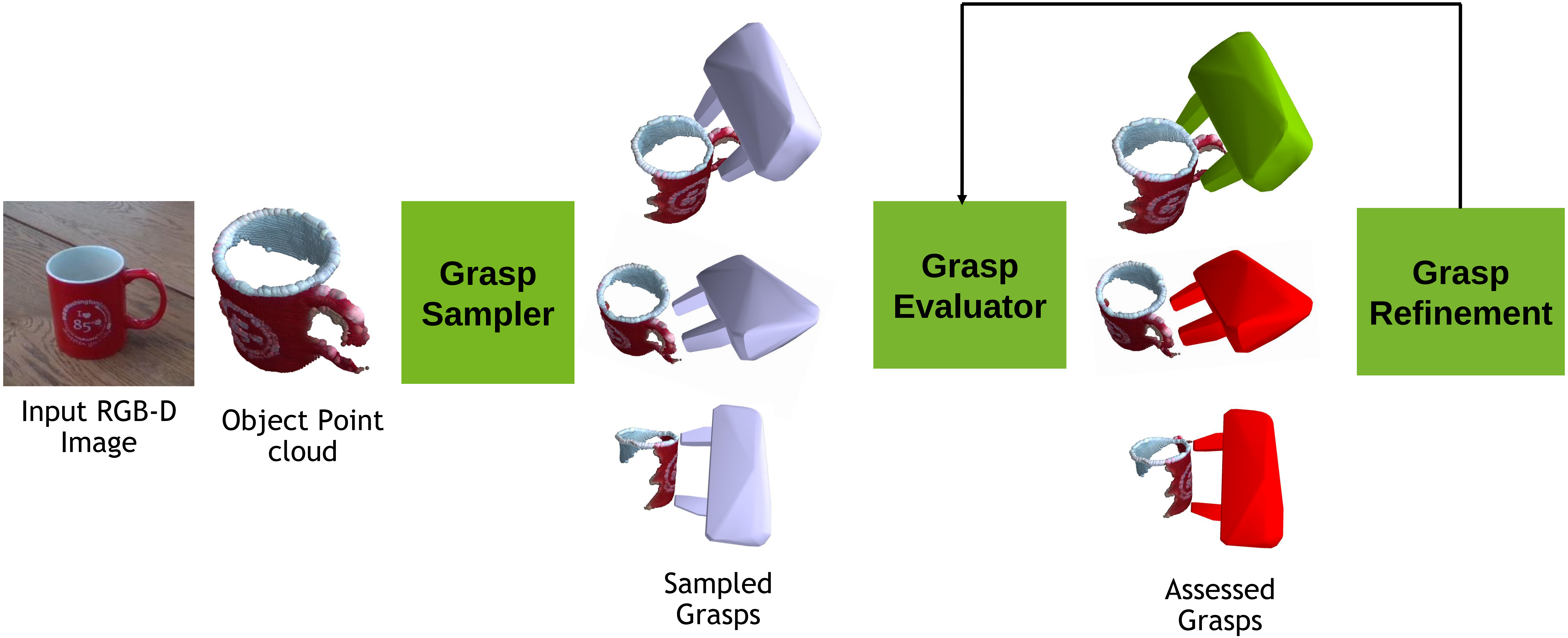
- Fully utilizes the rotation and translation space.
- Does not have any limitation on the camera angle.
- Grasps are represented in the object point cloud frame.
- Rotation axis is aligned with the camera
- Origin of the frame is placed at the center of mass for object point cloud.
- The coordinate frame is translation invariant.



[**Mousavian**-Eppner-Fox, ICCV 2019]

6-DOF GRASPNET

Generate 6D Grasp Poses from Input Point Cloud

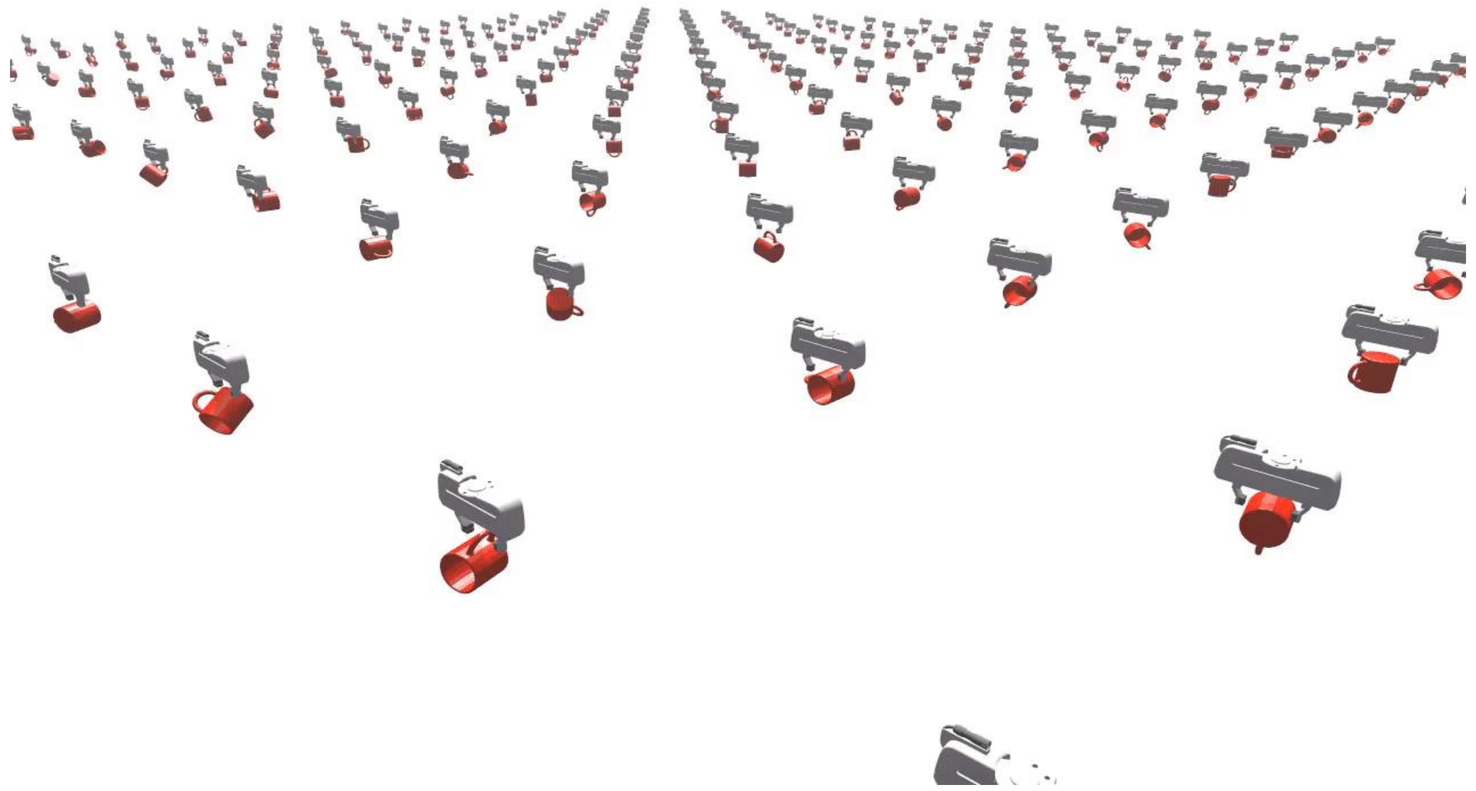


GitHub https://github.com/NVlabs/6dof-graspnet

[**Mousavian**-Eppner-Fox, ICCV 2019]

TRAINING

Training is done with synthetic data

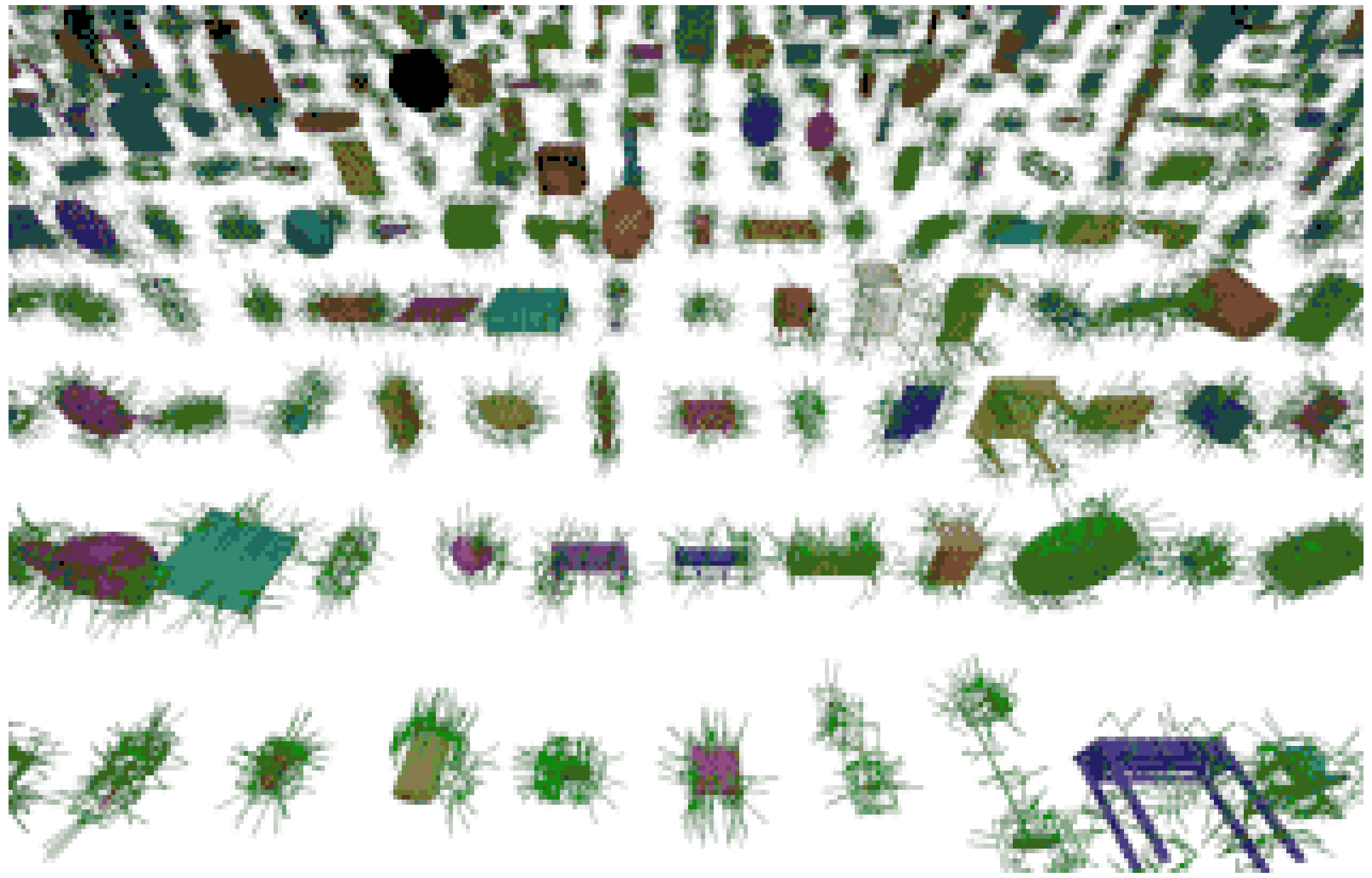


Trained on 126 random mugs, bowls, bottles, boxes, and cylinders.

Training grasps are evaluated in NVIDIA Flex.

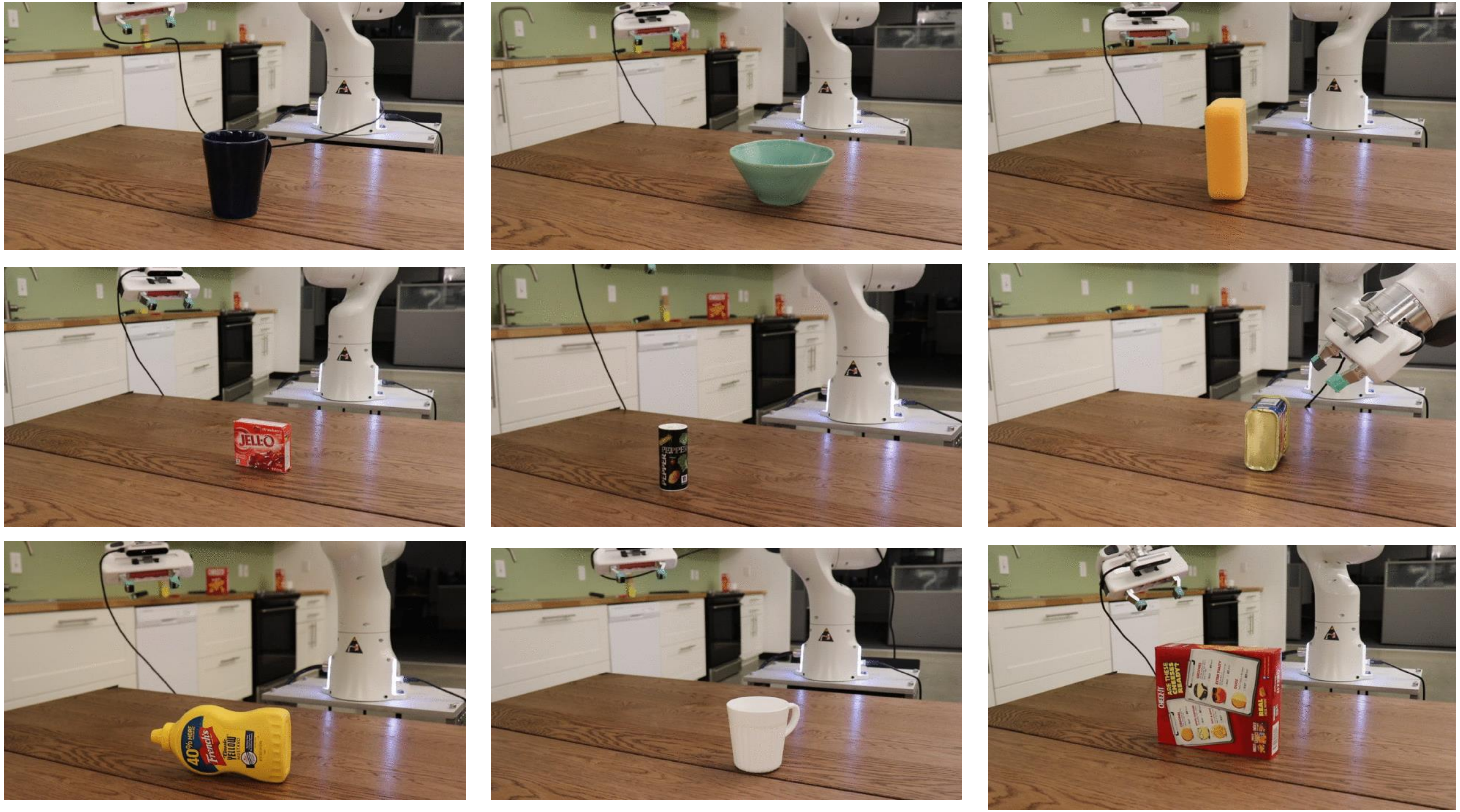
No Domain Adaptation is Needed

ACRONYM Dataset:[Eppner-**Mousavian**-Fox, ICRA 2021]
https://sites.google.com/nvidia.com/graspdataset



- 17.7M grasps
- 8872 meshes
- 262 categories
- Provided with Object scales
- Increases robustness and generalization across different methods.

QUALITATIVE RESULTS



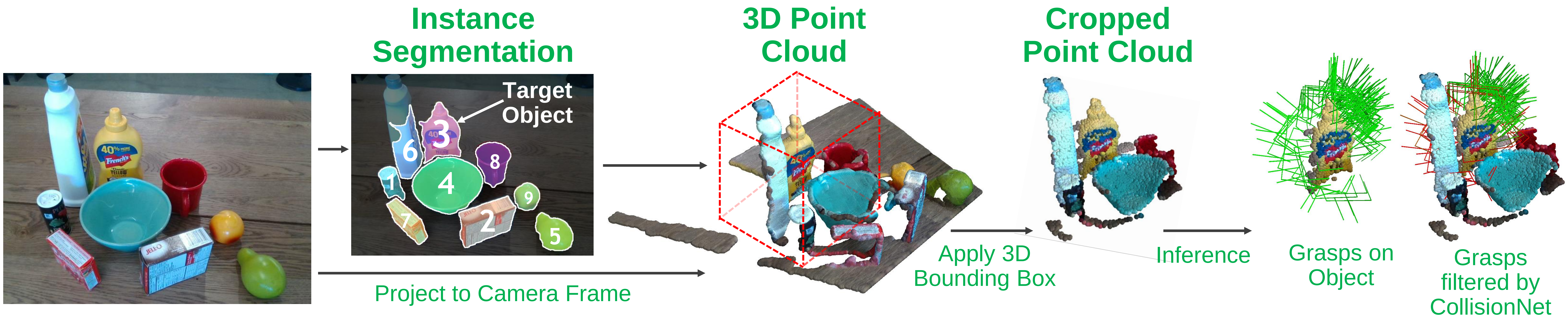
	Box	Cylinder	Bowl	Mug	Average Success Rate	Success Rate
6-DOF GraspNet	**83%**	**89%**	**100%**	**86%**	**90%**	**88%**
GPD [1]	50%	78%	78%	6%	52%	47%

ROBOT OBJECT HANDOVER

[Yang-Paxton-**Mousavian**-Chao-Cakmak-Fox, ICRA 2021]
Best Human Robot Interaction Award

6-DOF GRASPING FOR CLUTTERED SCENES

Extending single object grasping to clutter scenes



We only reason about gripper collision with the scene

[Murali-**Mousavian**-Eppner-Paxton-Fox, ICRA 2021]

Best Robot Manipulation Paper Finalist

ROBOT EXPERIMENTS

Removing blocker object

Target object is initially not reachable;
grasps will collide with surrounding clutter

Method	Success Rate
6-DOF GraspNet + Voxelization	62.7%
Cluttered 6-DOF Graspnet	80.3%

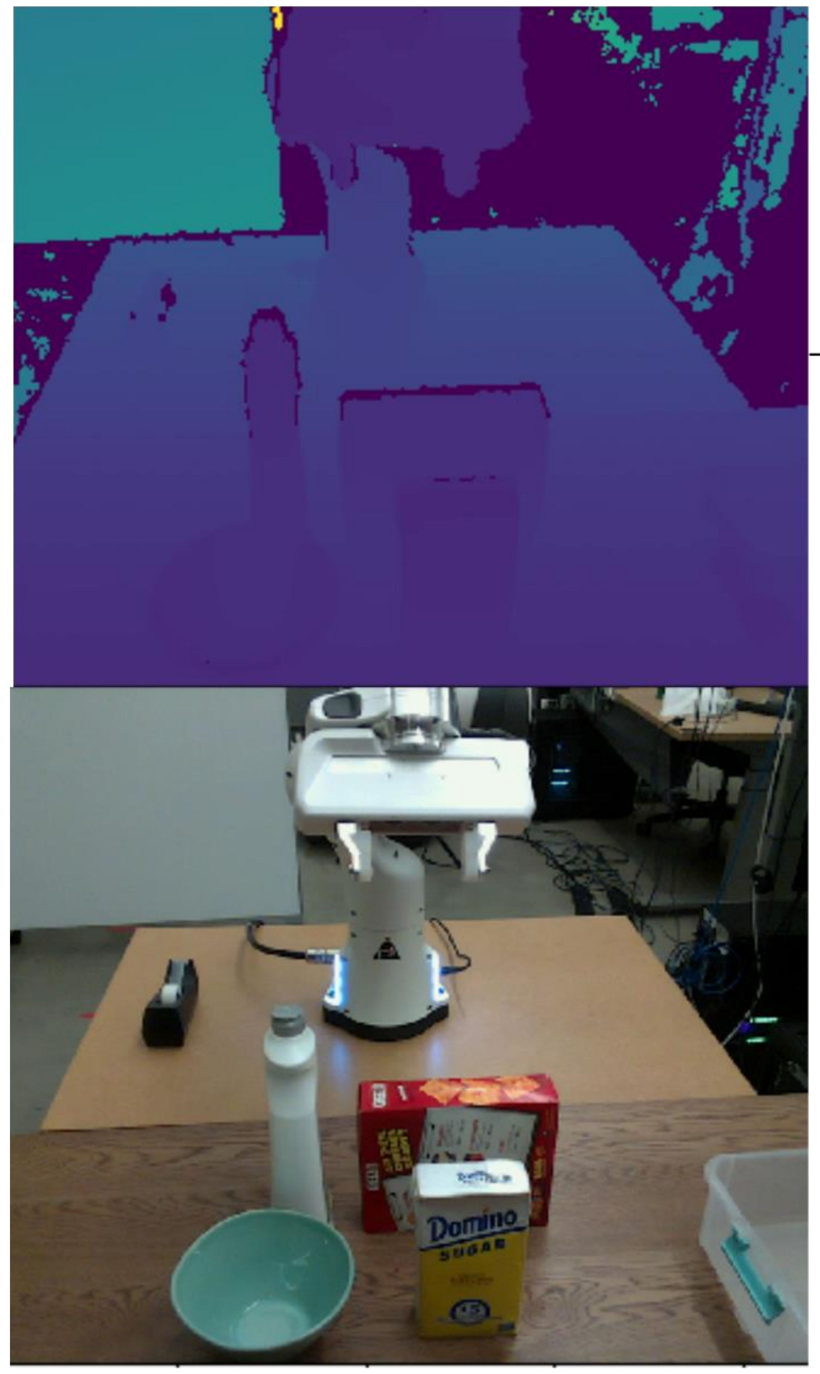
LIMITATIONS

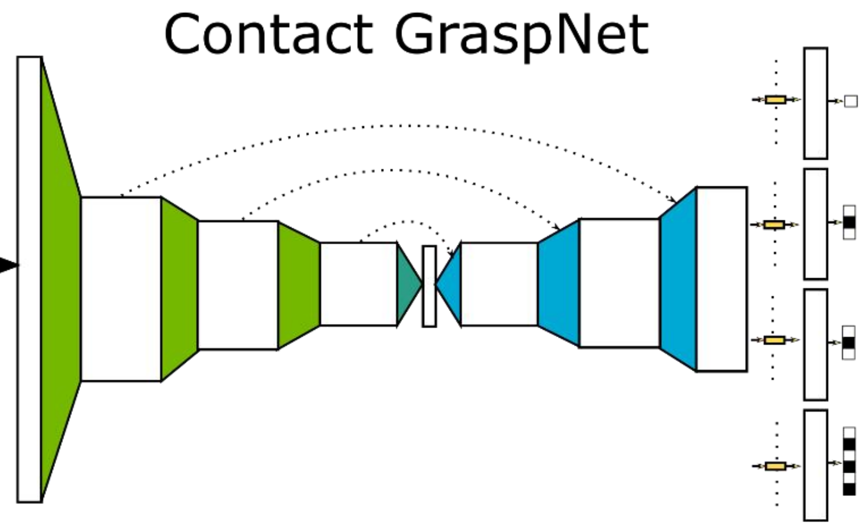
- Sensitivity to instance segmentation

- Grasps are generated for the target object in isolation of surrounding and then filtering is done.

CONTACT GRASPNET

Contact-GraspNet generates a dense collision-free 6-DoF grasp distribution from raw point clouds.

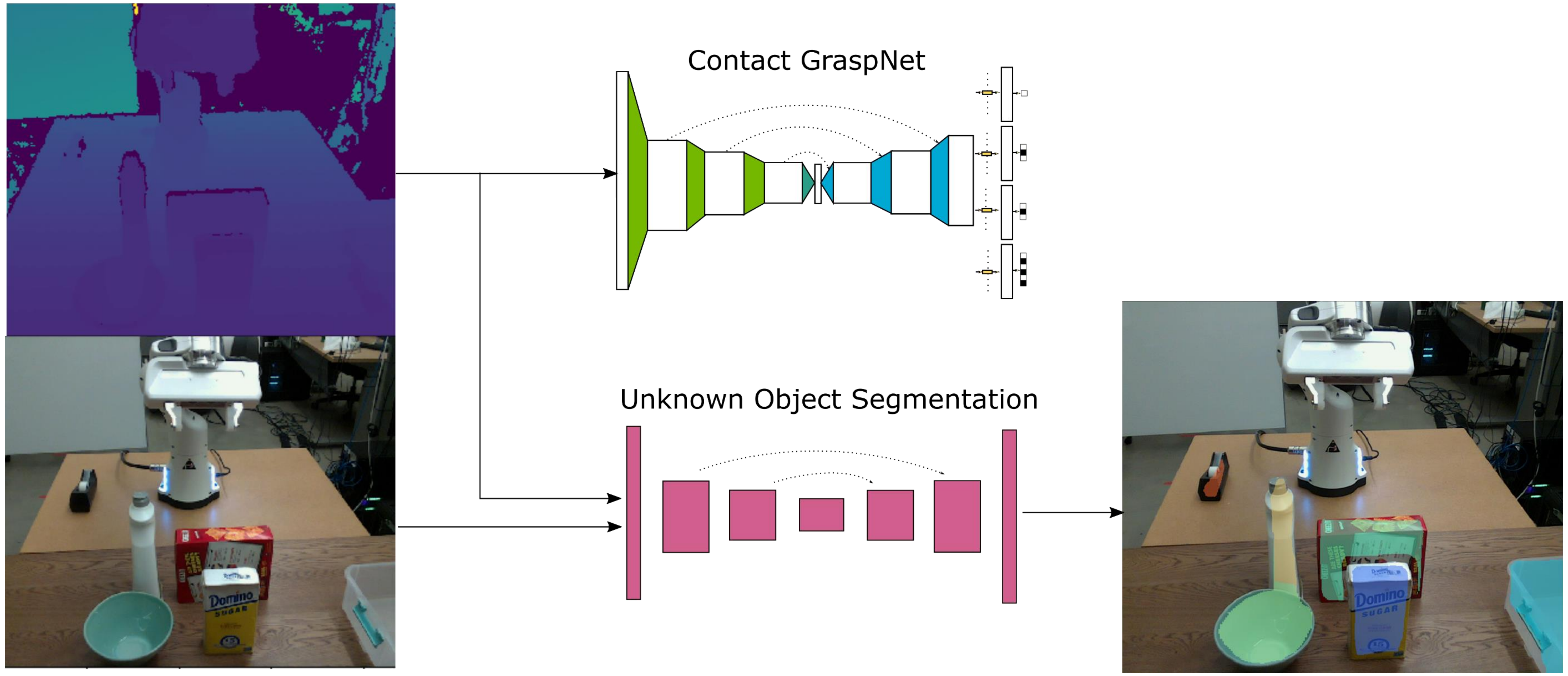




[Sundermeyer-**Mousavian**-Triebel-Fox, ICRA 2021]

OUR APPROACH

Contact-GraspNet generates a dense collision-free 6-DoF grasp distribution from raw point clouds.



[Sundermeyer-**Mousavian**-Triebel-Fox, ICRA 2021]

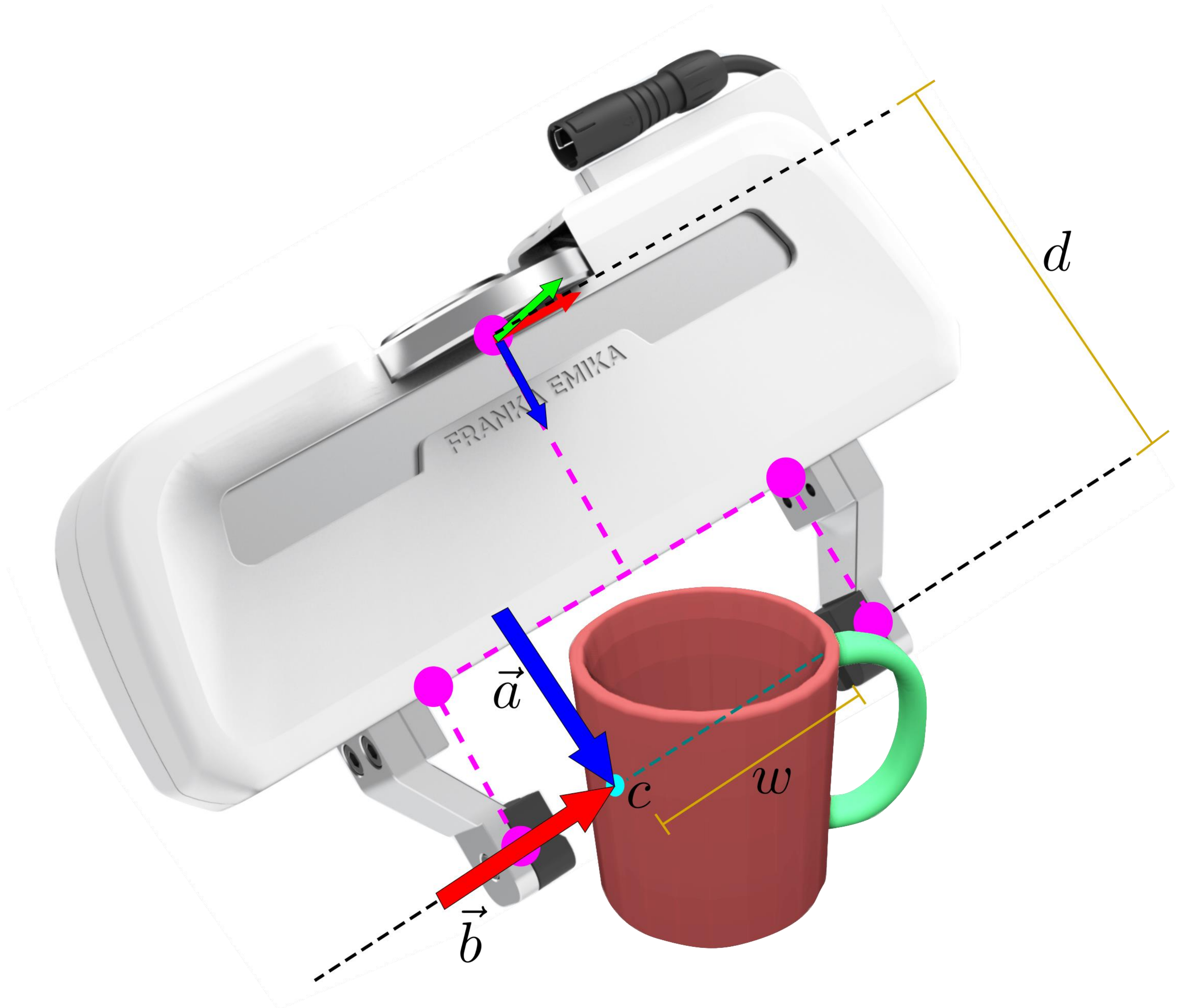
OUR APPROACH

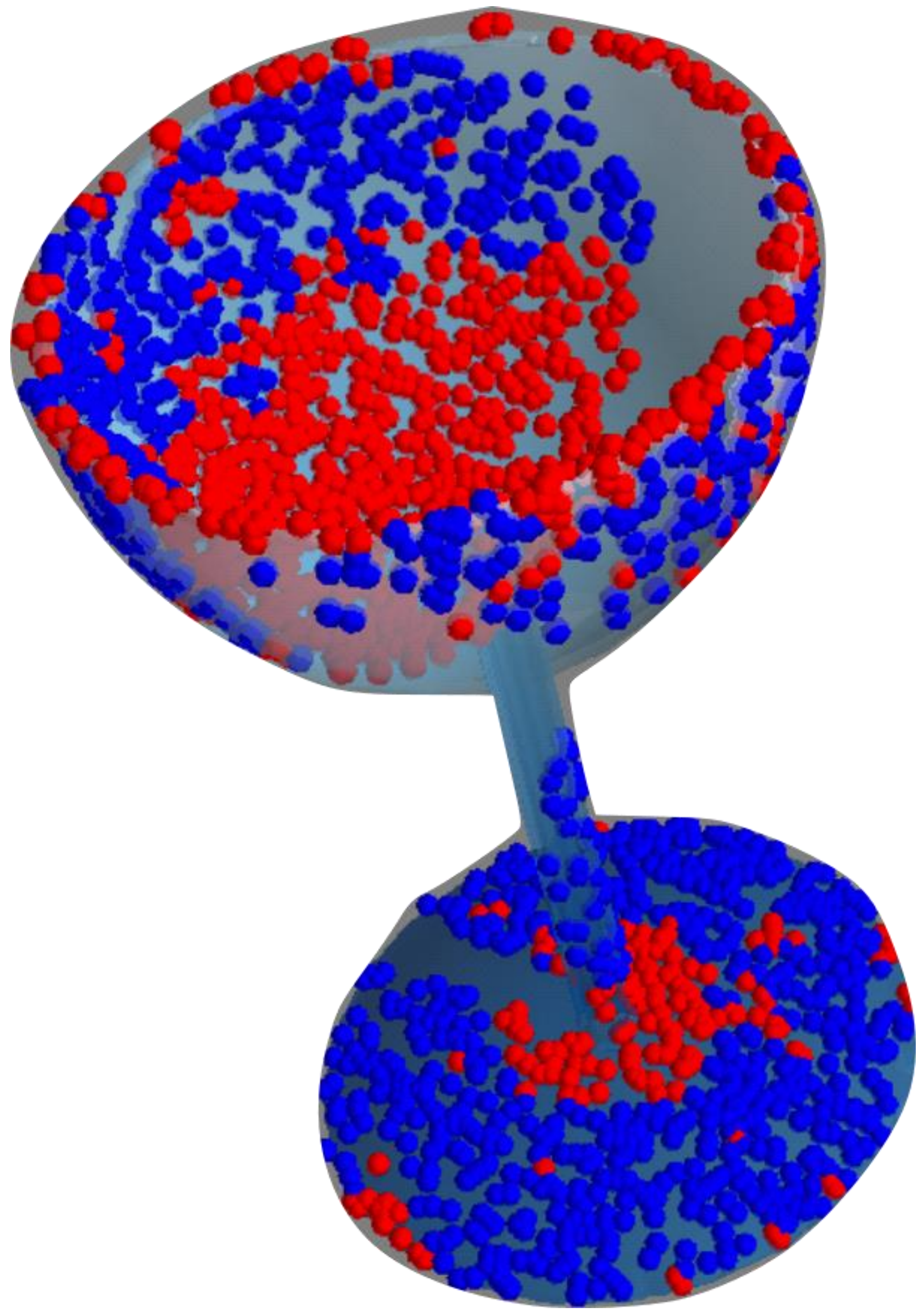
Contact-GraspNet generates a dense collision-free 6-DoF grasp distribution from raw point clouds.

[Sundermeyer-**Mousavian**-Triebel-Fox, ICRA 2021]

GRASP REPRESENATION

Assumption: Most of the suitable grasps have at least one visible contact point





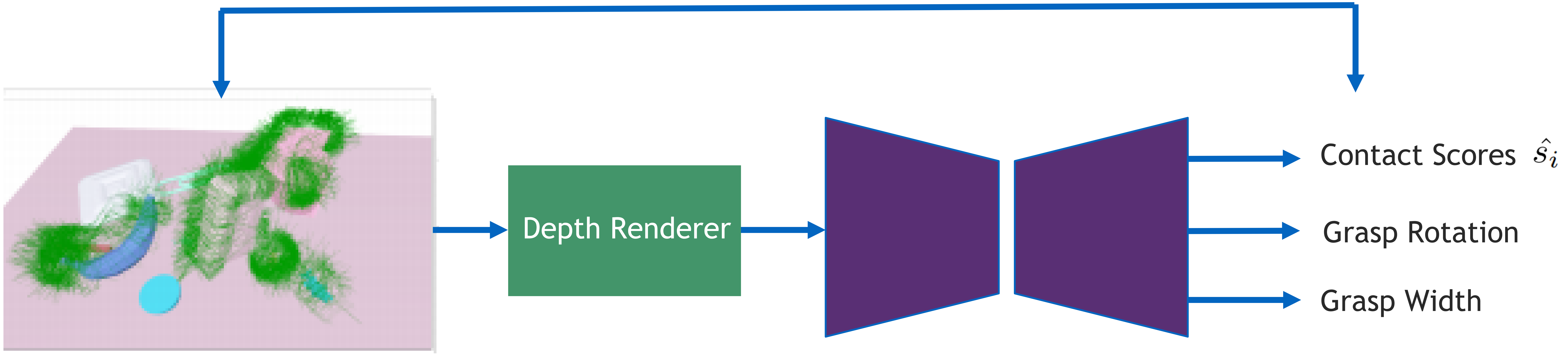
- Suitable Contact Point
- Unsuitable Contact Point

6-DoF grasps are **mapped** to their **contacts** in point cloud

Predict **Contact Points** + **3-DoF** Rotation + **1-DoF** Width

Learnability + **Efficiency + Grasp Coverage**

NETWORK ARCHITECTURE AND TRAINING LOSSES



Scene Generation

$$l_{add-s} = \frac{1}{n^+} \sum_{i}^{n^+} \hat{s}_i \min_{u} ||\mathbf{v}_i^{pred} - \mathbf{v}_u^{gt}||_2$$

Suitable Contact points

Weighted Average distance of gripper points in predicted + closest gt pose

$$l = \alpha l_{bce,k} + \beta l_{add-s} + \gamma l_{width}$$

$$\alpha = 1, \beta = 10, \gamma = 1$$

GRASPING UNKNOWN OBJECTS IN REAL WORLD

90.2% success rate in real world. Outperforms cluttered graspnet by 10% on the same scenes.

[Sundermeyer-**Mousavian**-Triebel-Fox, ICRA 2021]

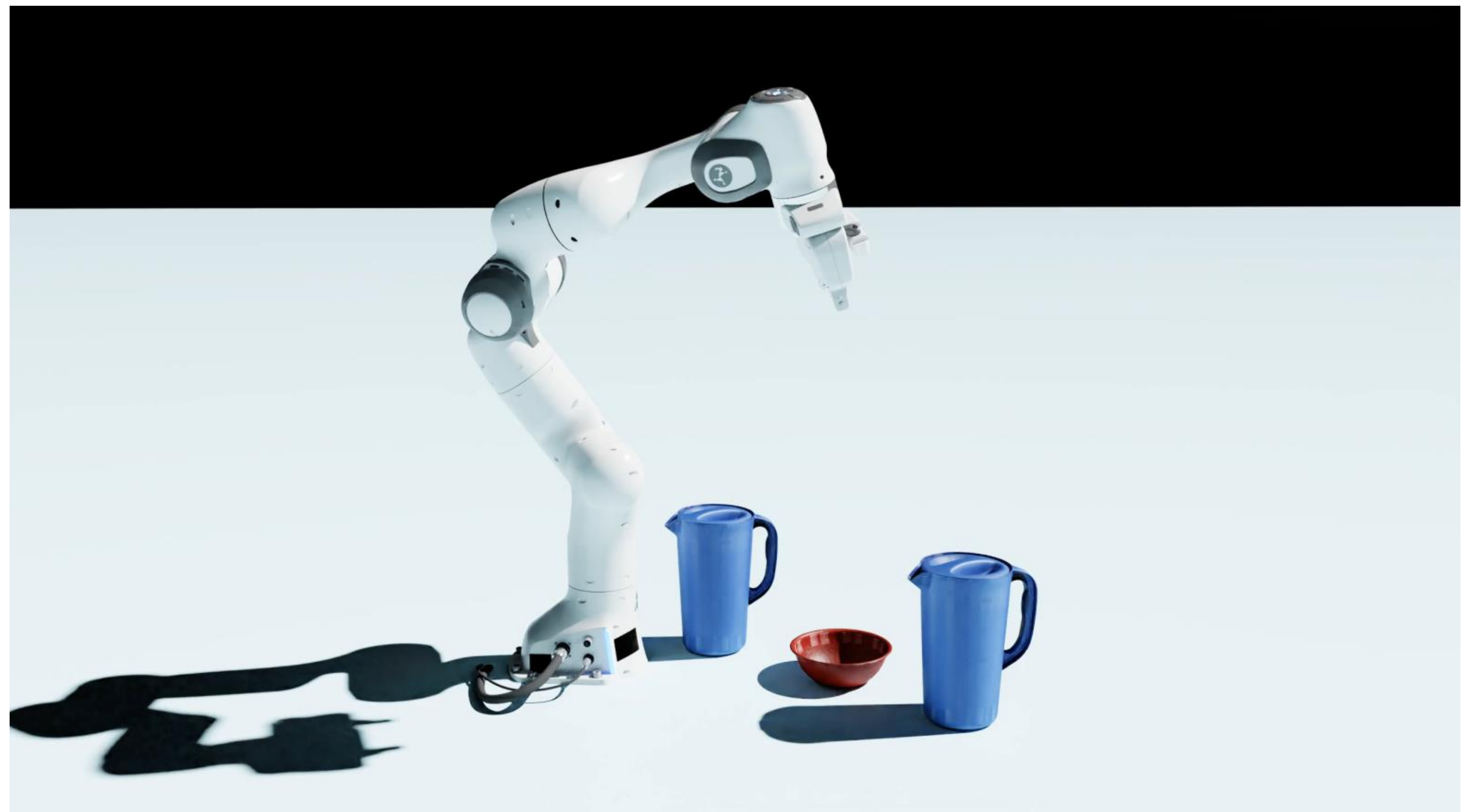
NVIDIA

MODEL-FREE MOTION PLANNING

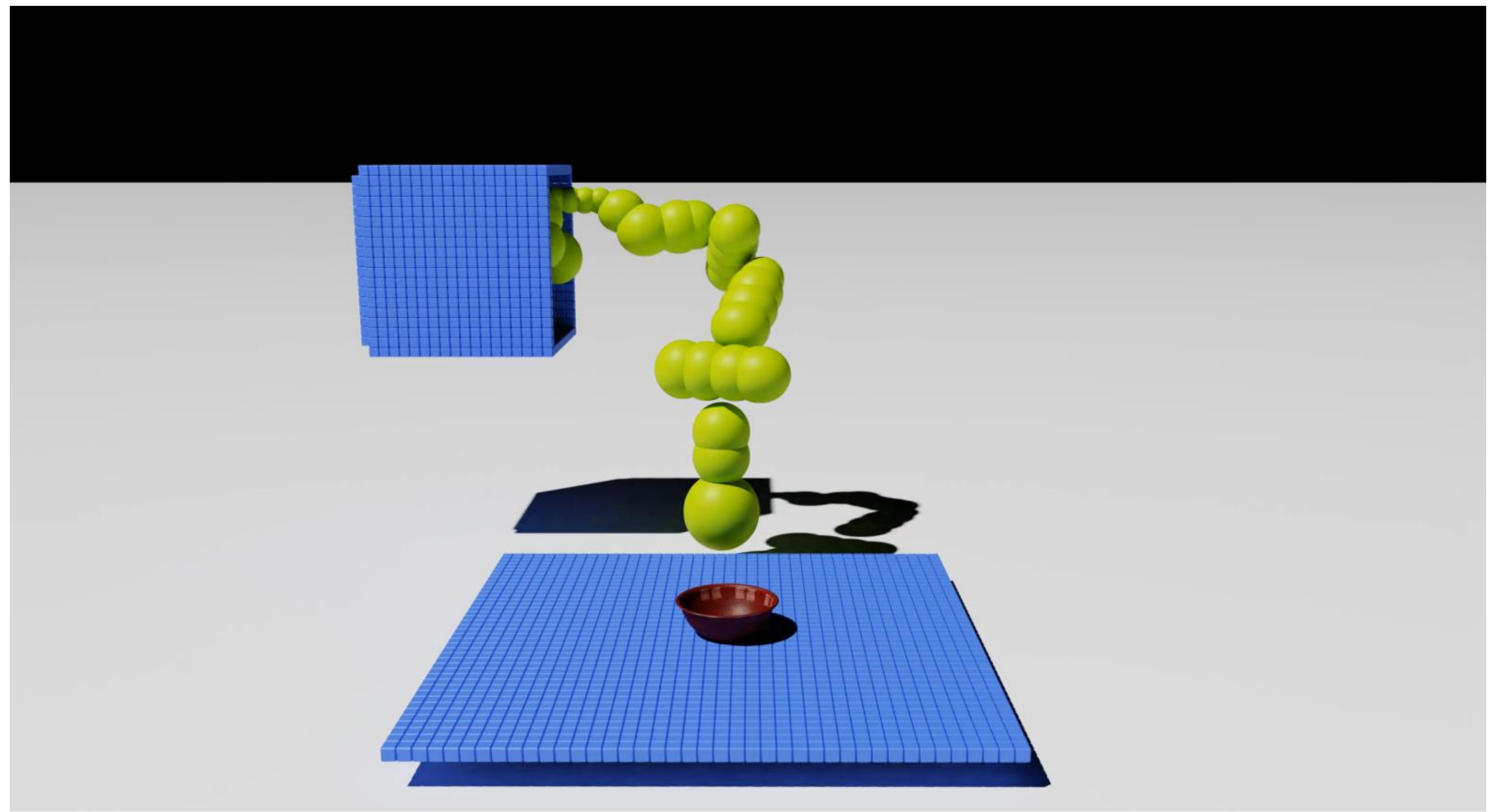
MOTION PLANNING

- Motion Planning: Find a sequence of valid and collision-free configurations that gets the robot from configuration A to configuration B.

- Collision Checking:
    - Model based
    - Approximate voxelization
    - Meshify Scene Pointcloud and compute SDF



Model Based

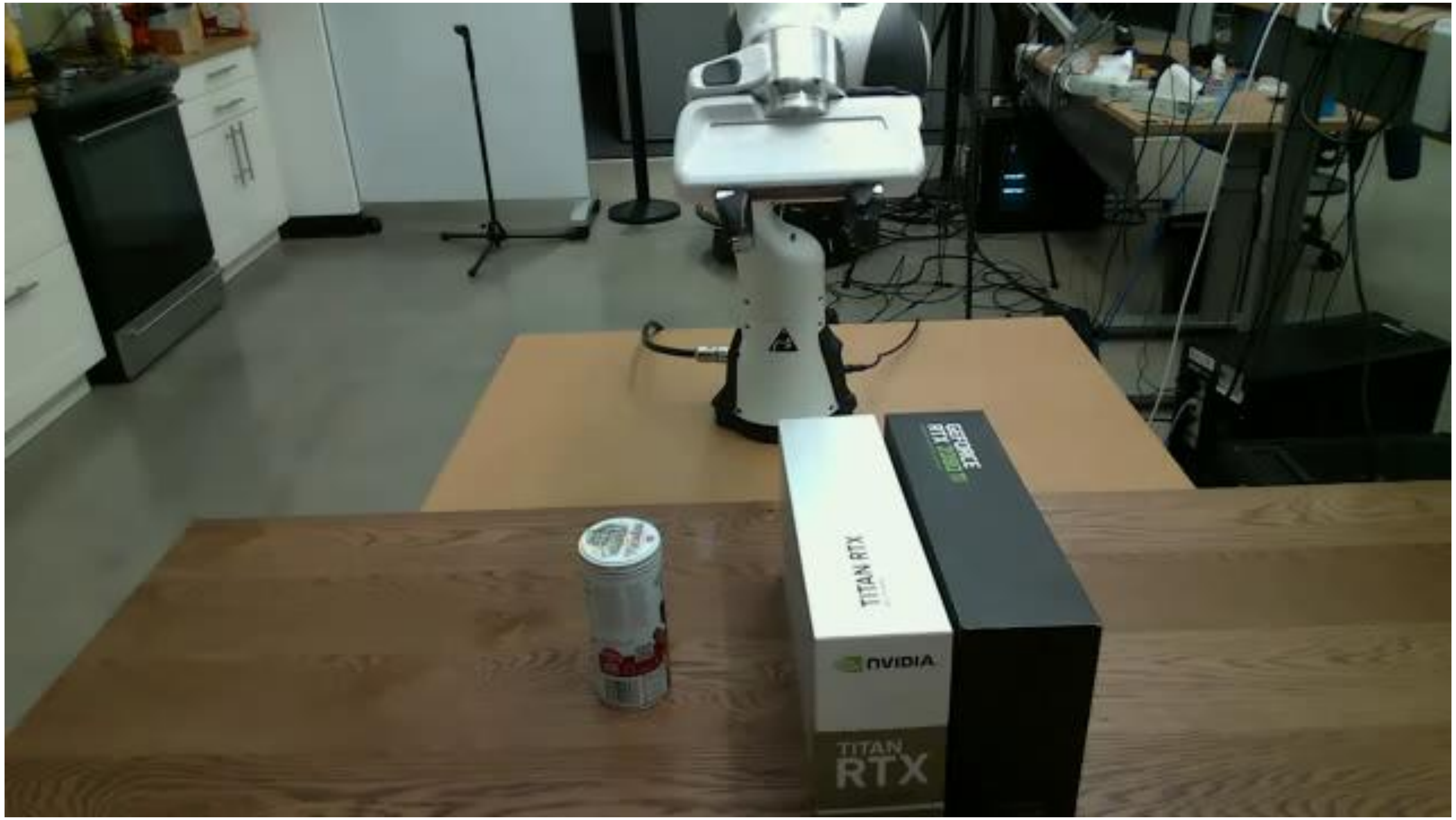


Scene Approximation

Videos Credit: [Islam-Paxton-Eppner-Peele-Likachev-Fox, ICRA 2021]

MOTIVATION

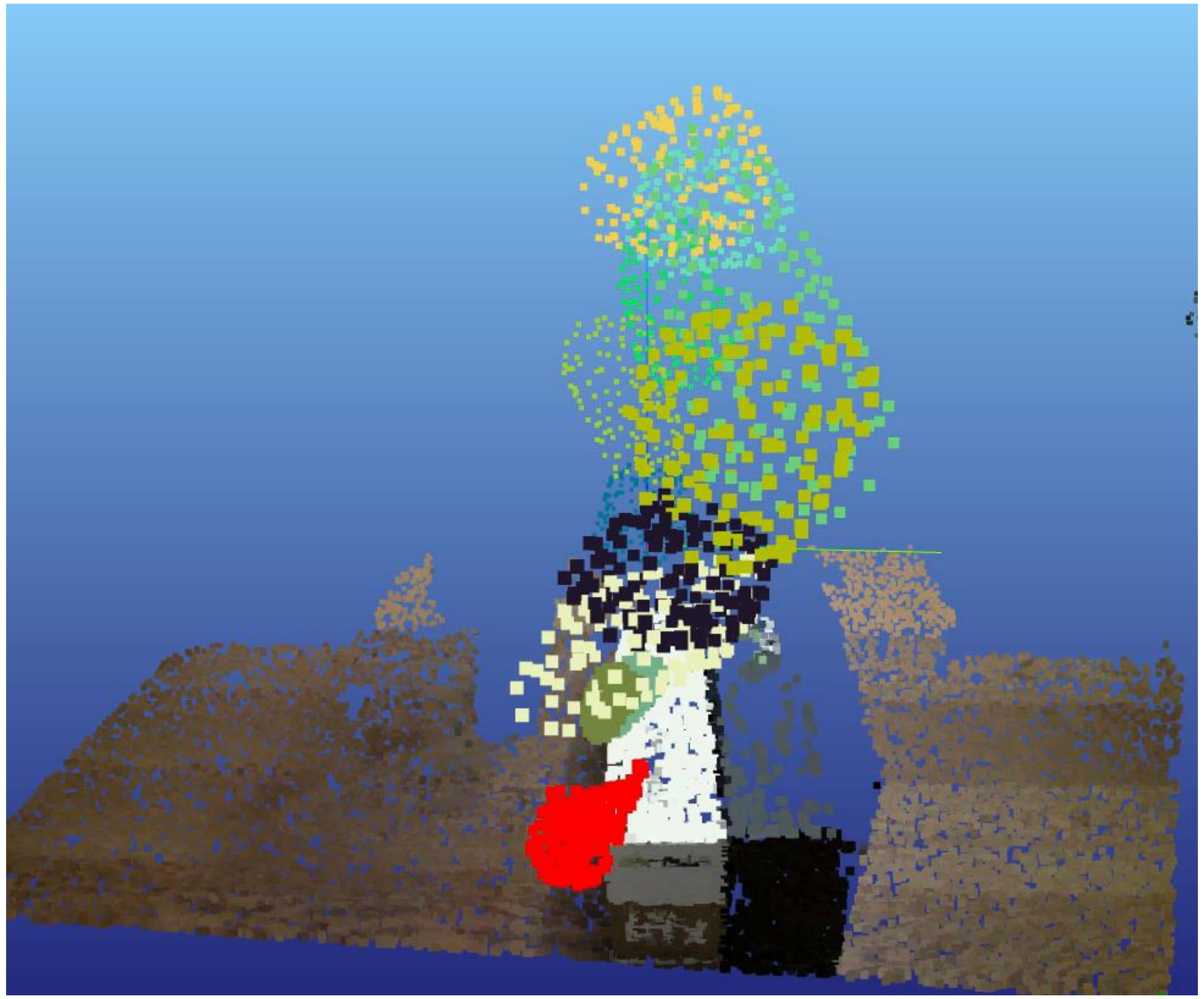
Motion planning for unknown objects and novel scenes from scene point cloud.



[Danielczuk*-**Mousavian***-Eppner-Fox, ICRA 2021]

COLLISION CHECKING

How do we check collisions?

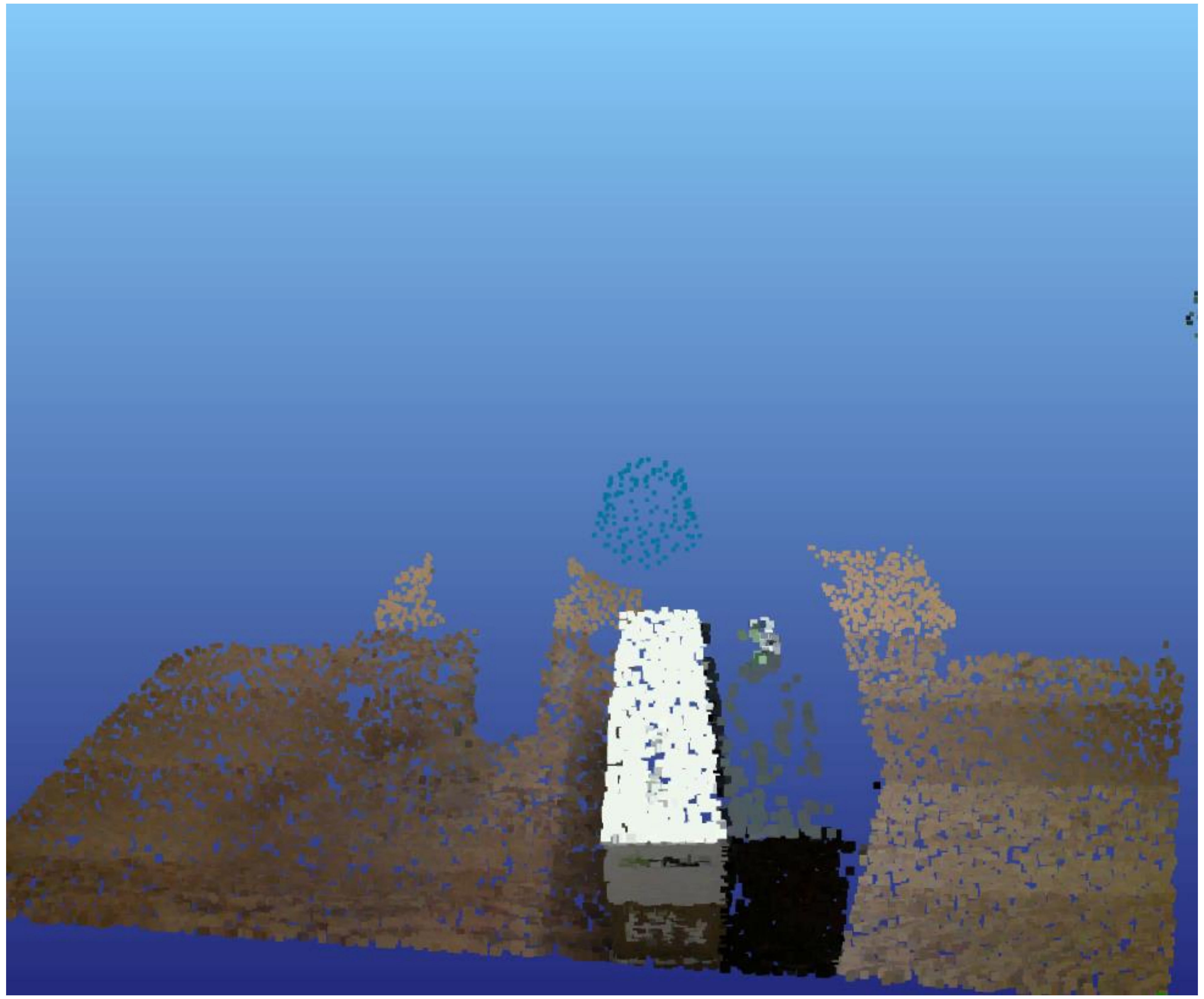


[Danielczuk*-**Mousavian***-Eppner-Fox, ICRA 2021]

COLLISION CHECKING

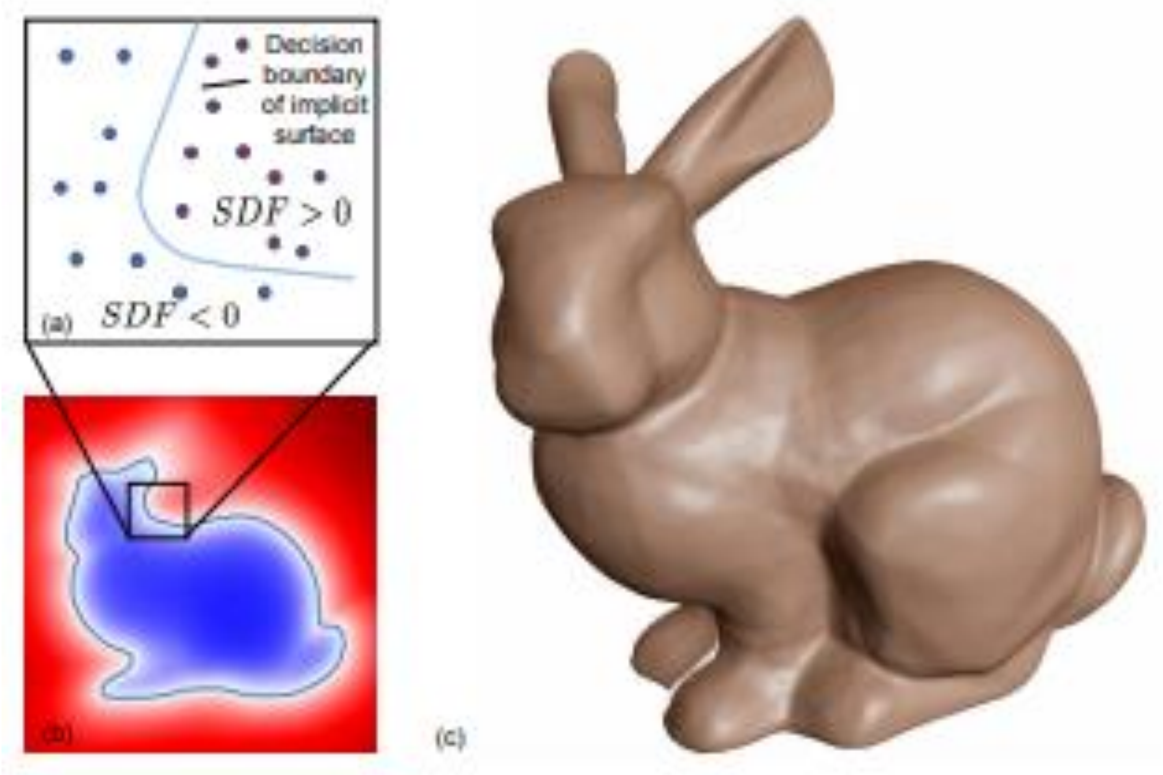
Check each query against the scene

Collision Query: Query Object point cloud + Transform

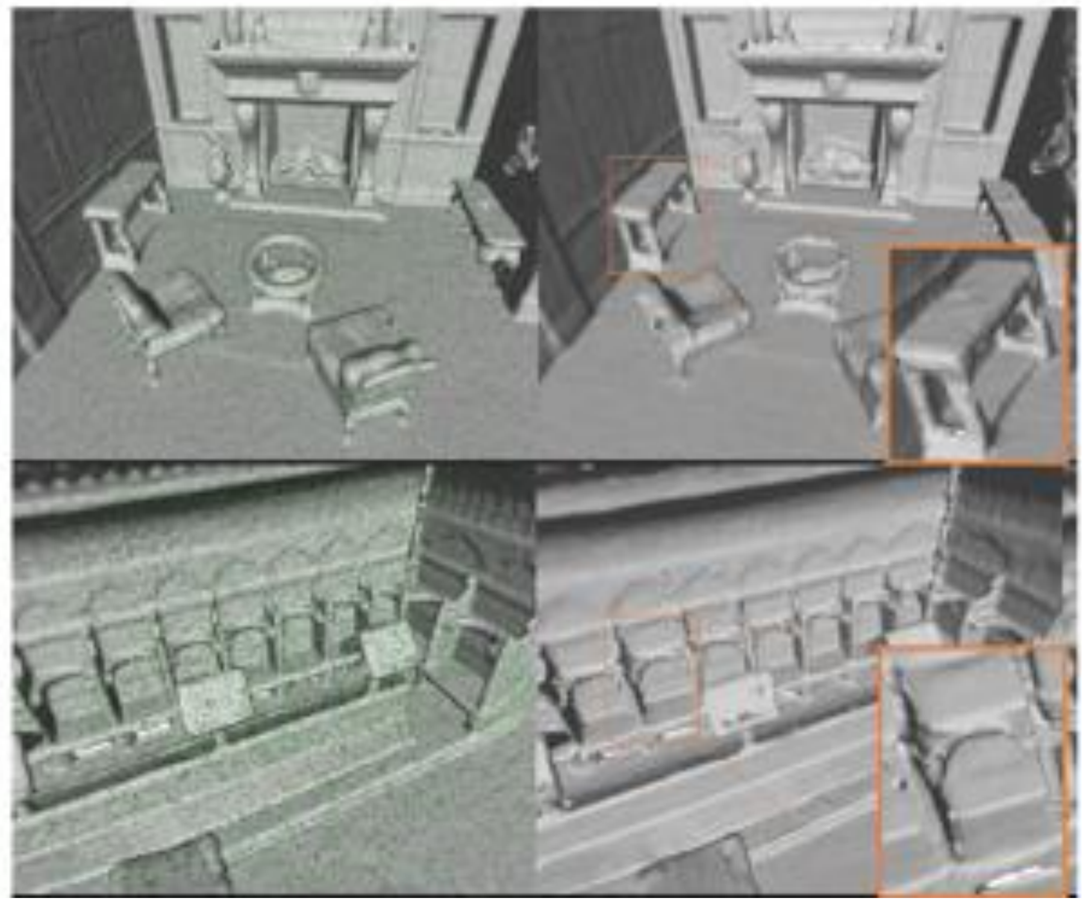


[Danielczuk*-**Mousavian***-Eppner-Fox, ICRA 2021]

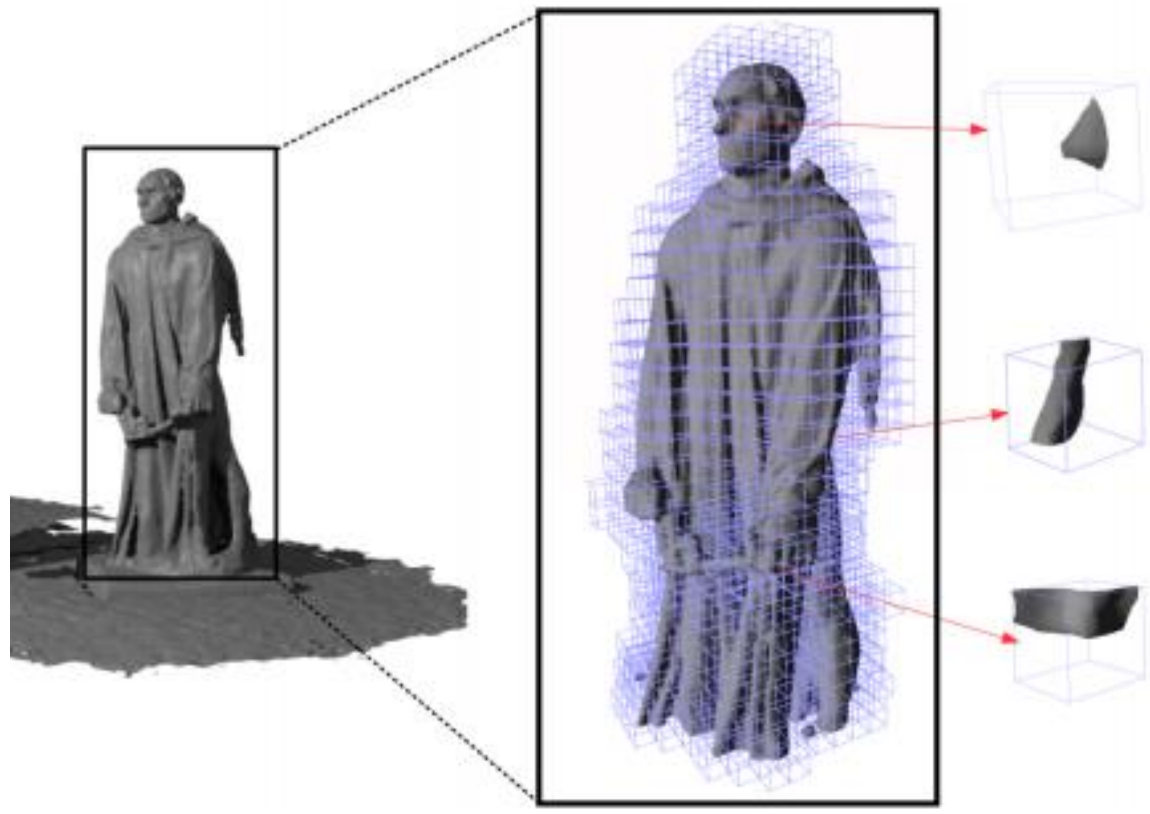
PREVIOUS WORKS



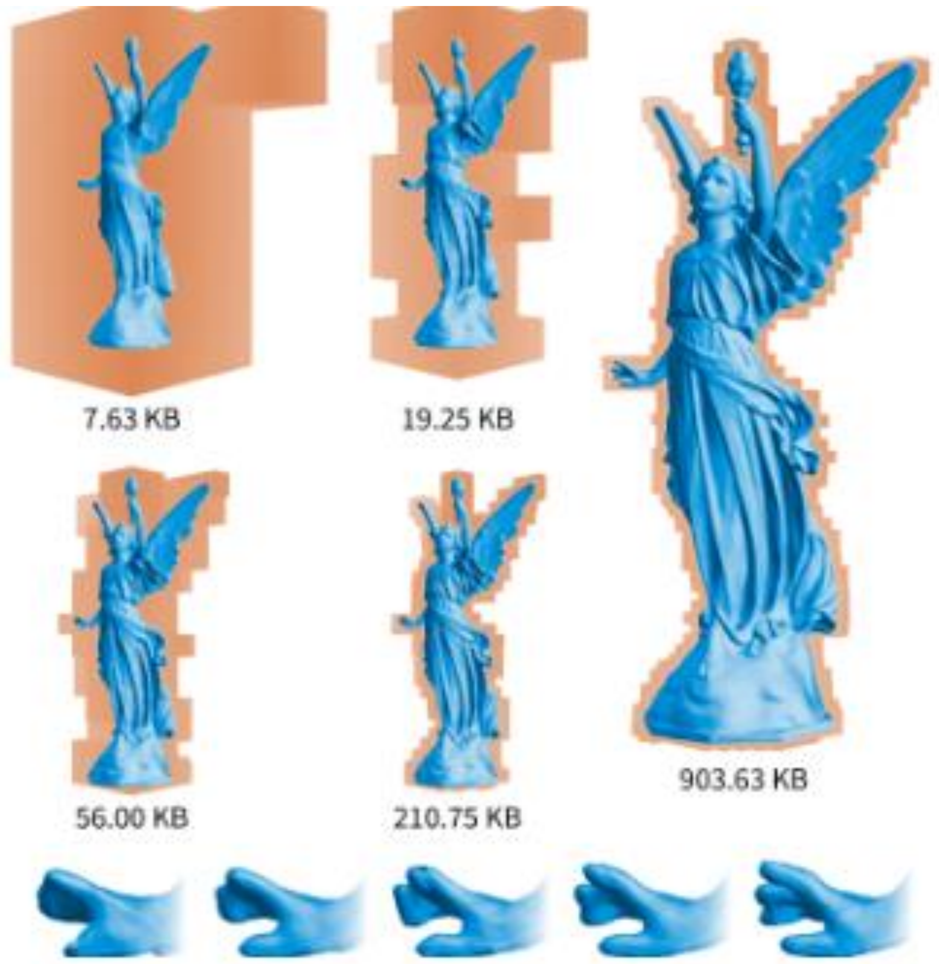
DeepSDF, Park et al, CVPR 2019



Local Implicit Grid, Jiang et al, CVPR 2020



Deep LS, Chabra et al, ECCV 2020



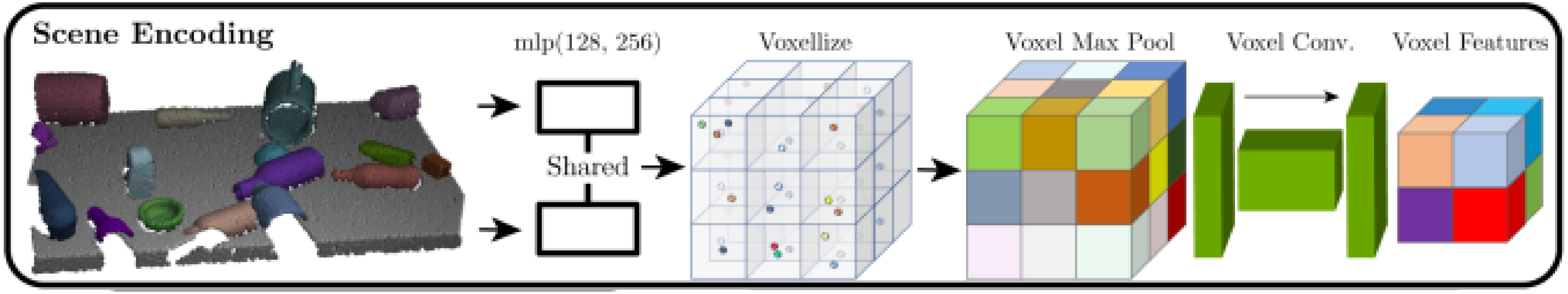
Neural LOD, Takikawa et al, CVPR 2021

- Downsides of the previous works for robotics applications:
    - Slow Inference Time
    - Poor generalization: One model per scene/category
- In robotics, we need fast inference and high generalization

SCENE COLLISIONNET

Instead of reconstructing the scene and querying it, train for solving collision queries.

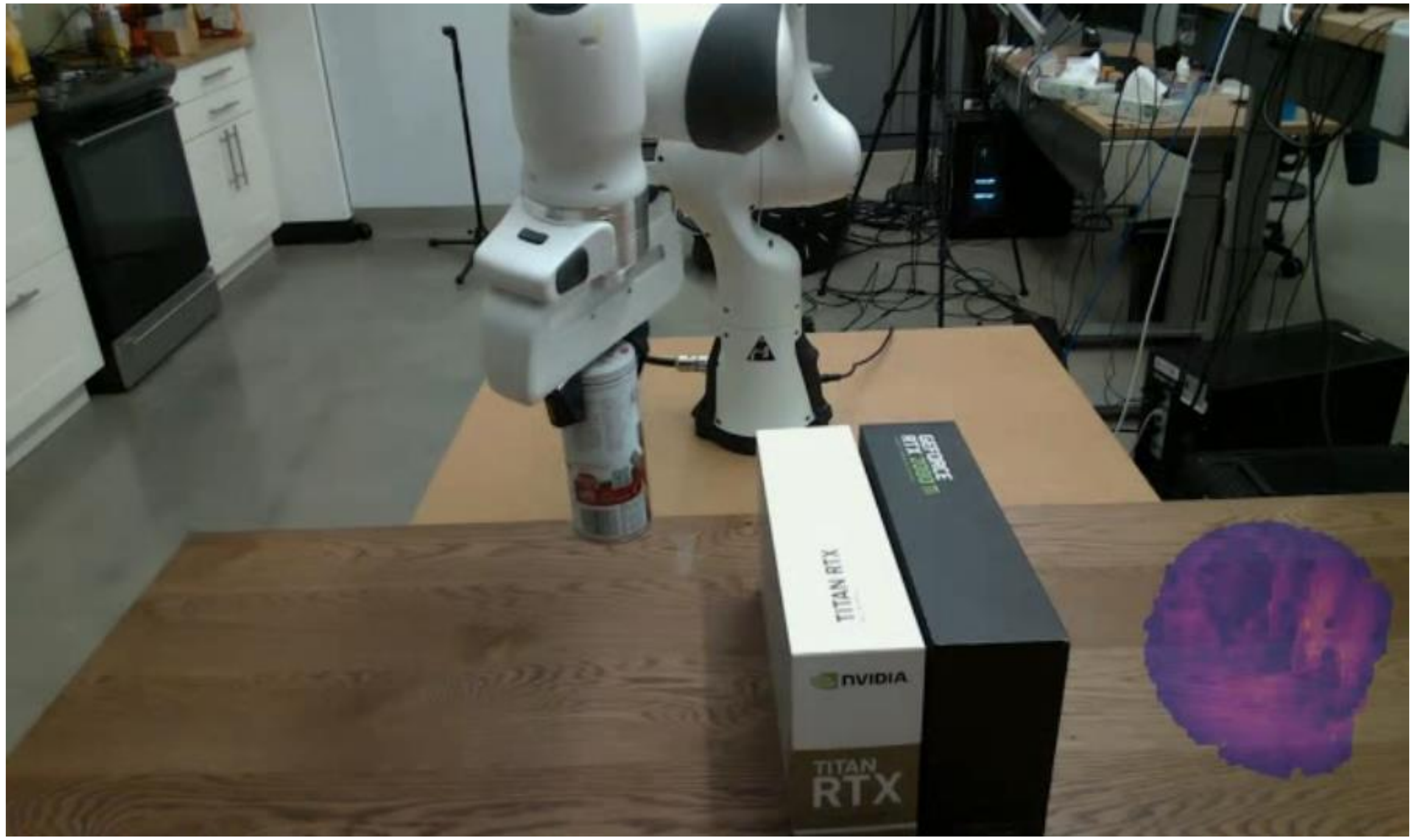
- The model can process 500K collision queries per scene in one forward pass on a 2080ti GPU in 0.1 seconds
- Training is done by generating scenes and having collision queries at different locations in the scene.



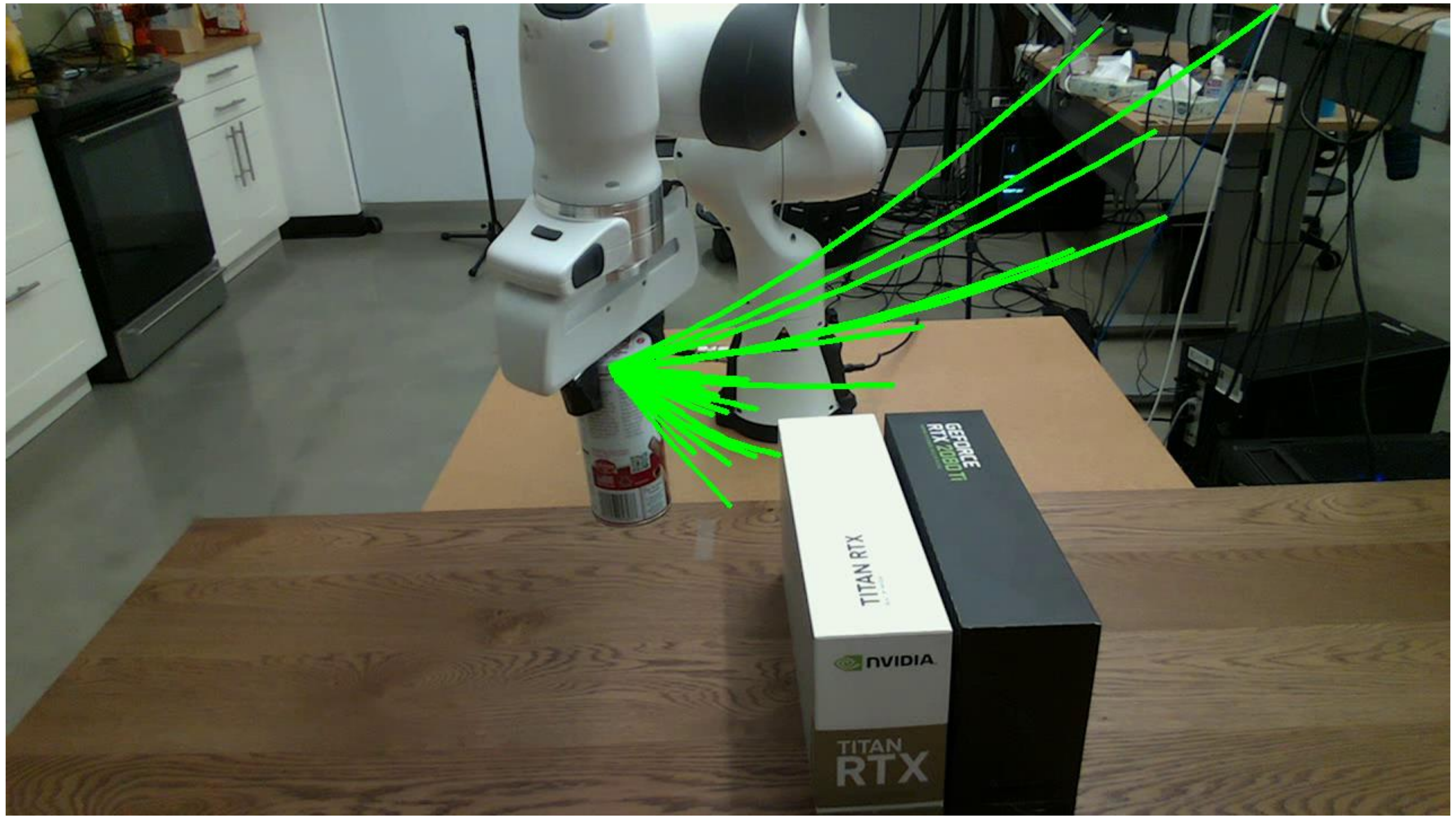
[Danielczuk*-**Mousavian***-Eppner-Fox, ICRA 2021]

MPPI FOR PLANNING TRAJECTORIES

Sample large number of paths in future and validate them.



Target Placements



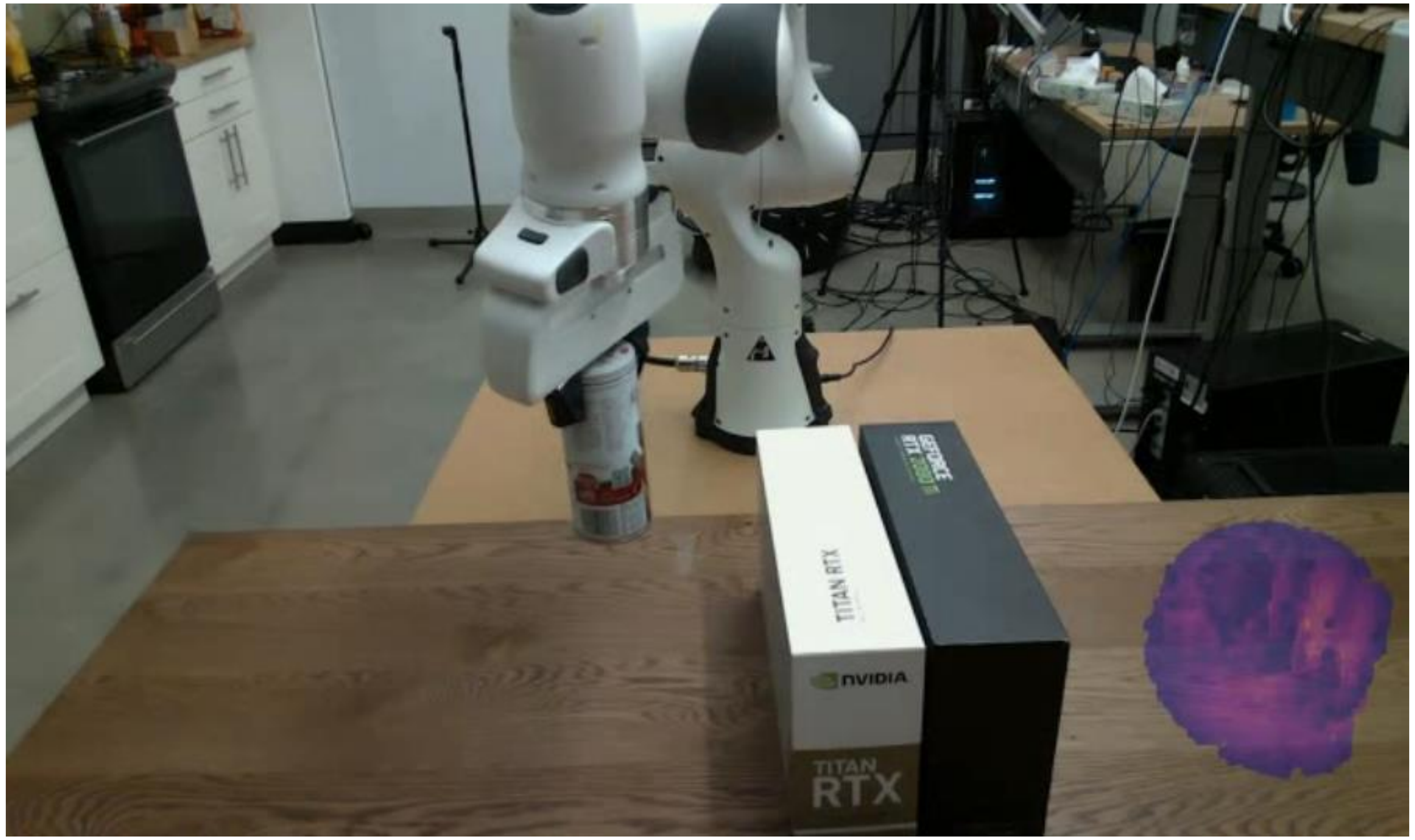
Rollouts

[Danielczuk*-**Mousavian***-Eppner-Fox, ICRA 2021]

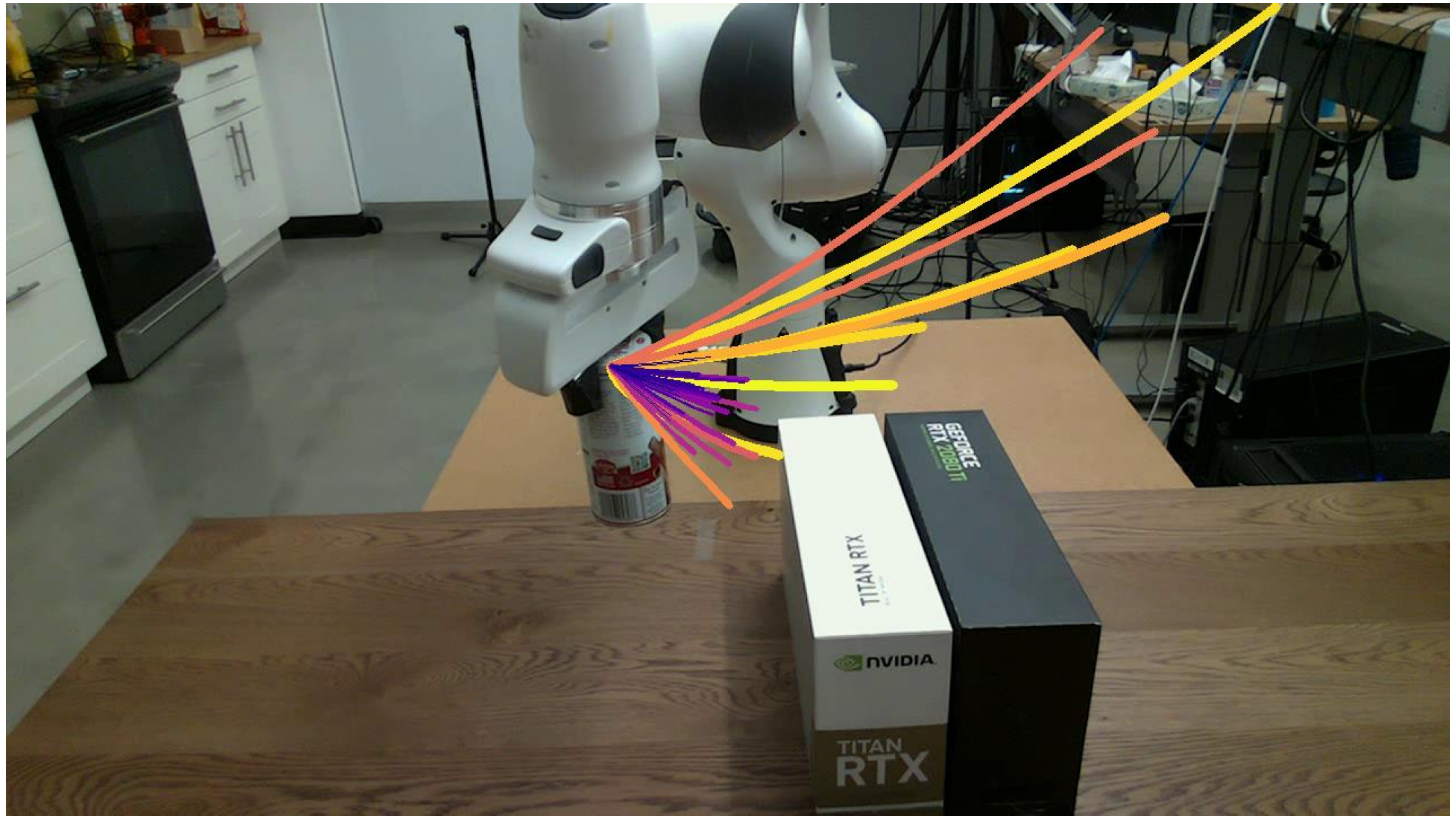
NVIDIA

SCORE ALL THE ROLLOUTS

Based on distance to goal configuration space and execute the best rollout



Target Placements



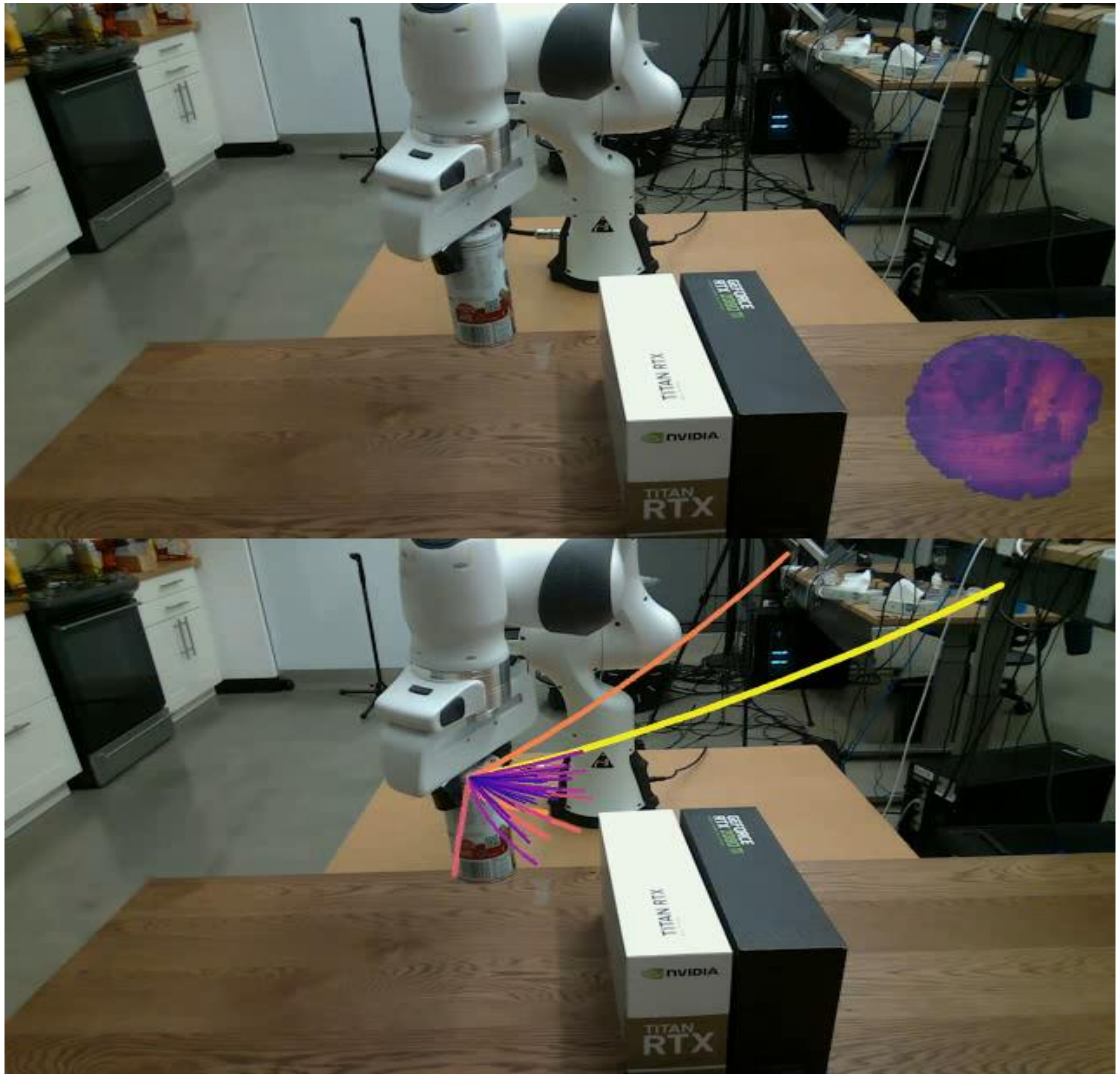
Rollout Scores

[Danielczuk*-**Mousavian***-Eppner-Fox, ICRA 2021]

NVIDIA

EXECUTE THE BEST ROLLOUT

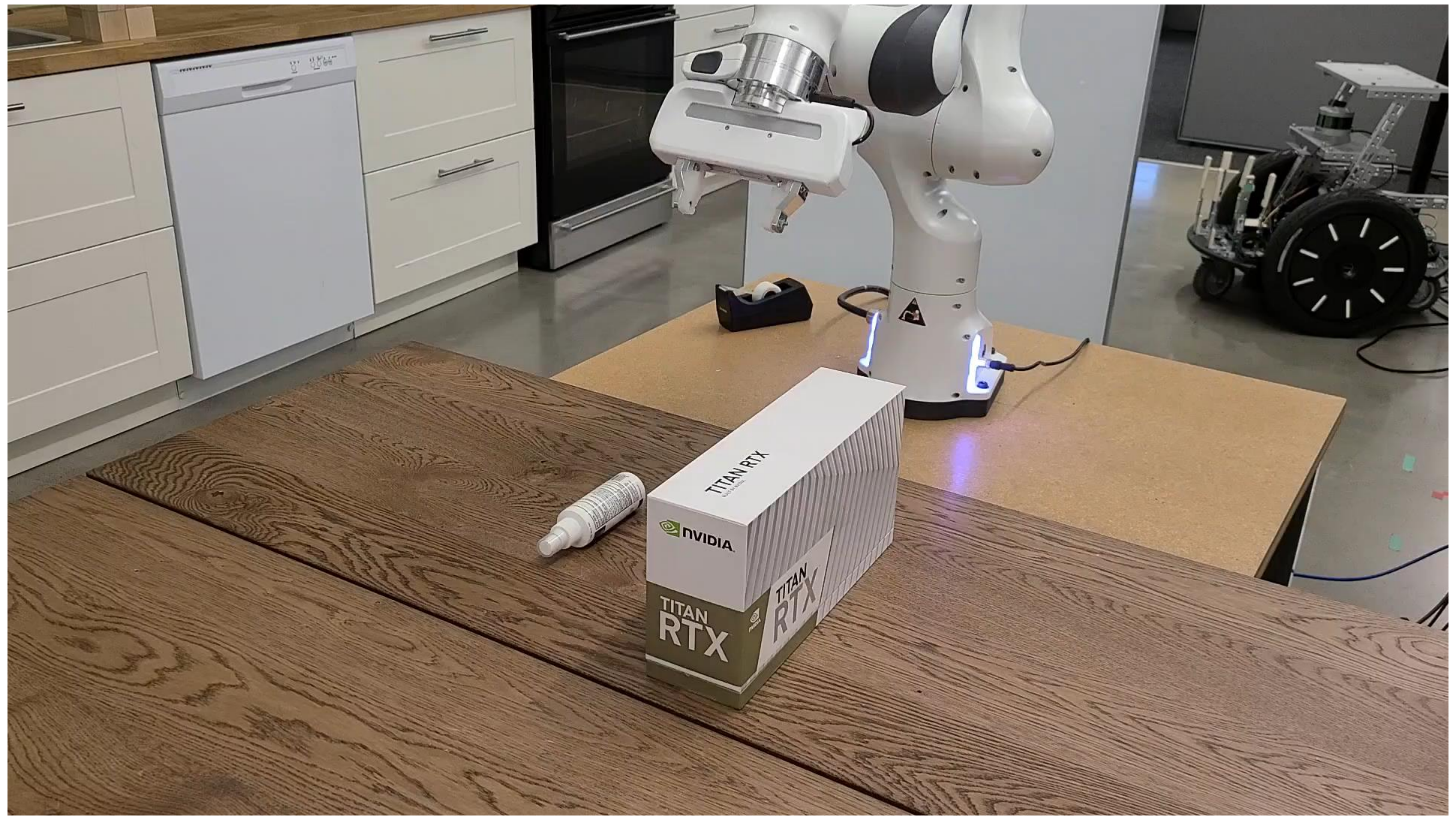
Execute the best rollout and repeat the process



nVIDIA.

REACTIVITY

Planner can react to the changes in the environment



Reactive Placing

Reactive Grasping

REAL ROBOT EXECUTION

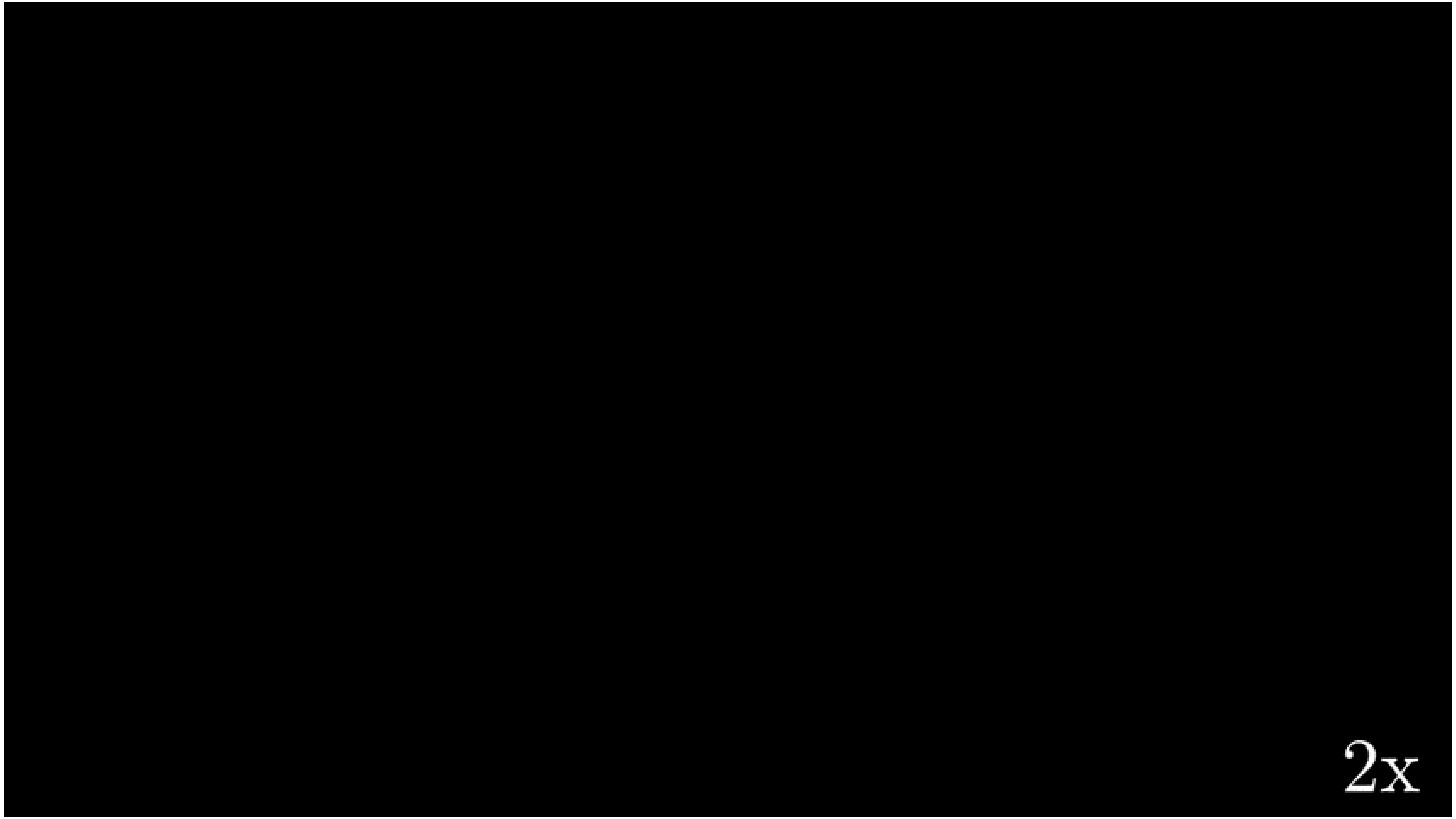


IMAGE BASED OBJECT REARRANGEMENT

IMAGE BASED OBJECT REARRANGEMENT

Given the target image, the robot has to rearrange the objects to get similar configuration to target image.
Novel objects and configuration. No additional training

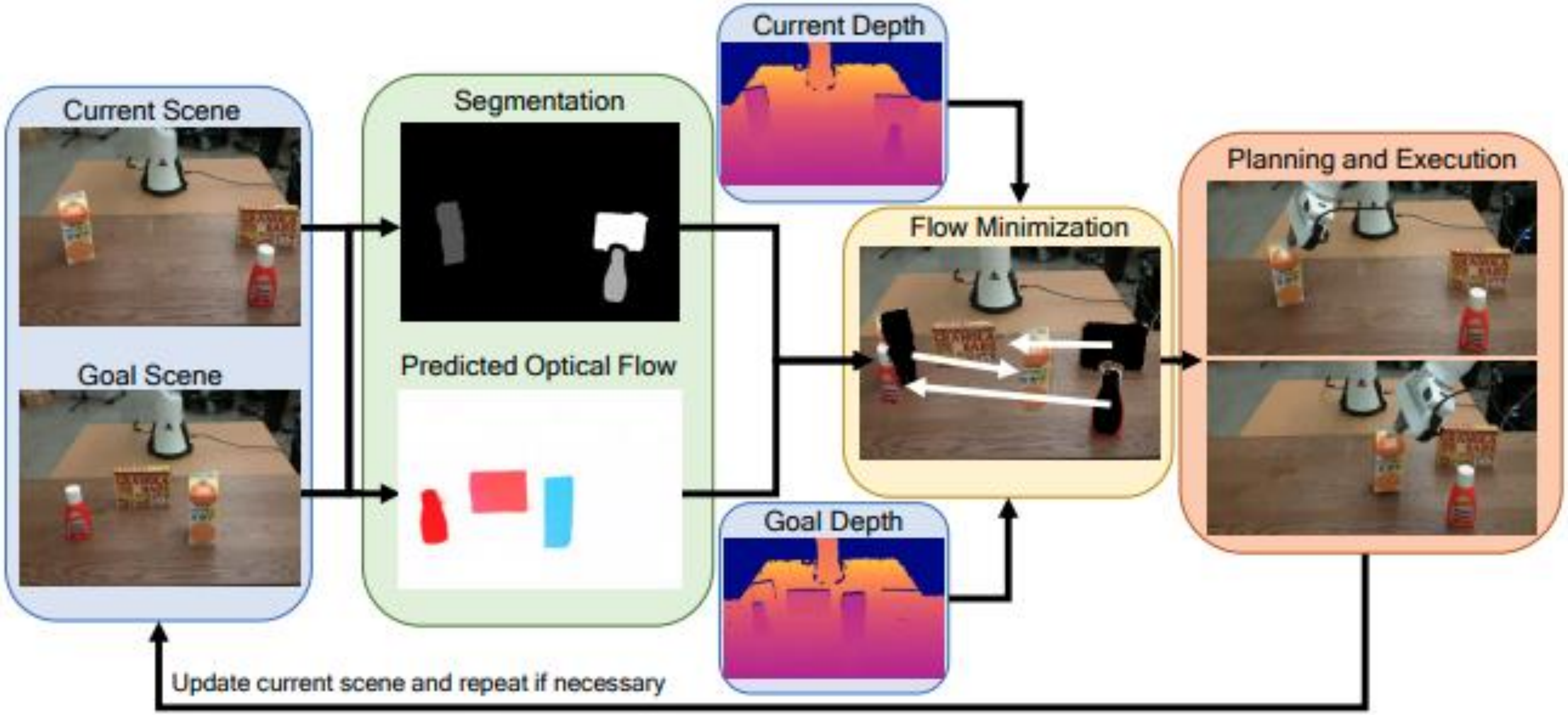
Target Image

Robot Execution

[Goyal-**Mousavian**-Paxton-Chao-Fox, CVPR 2022]

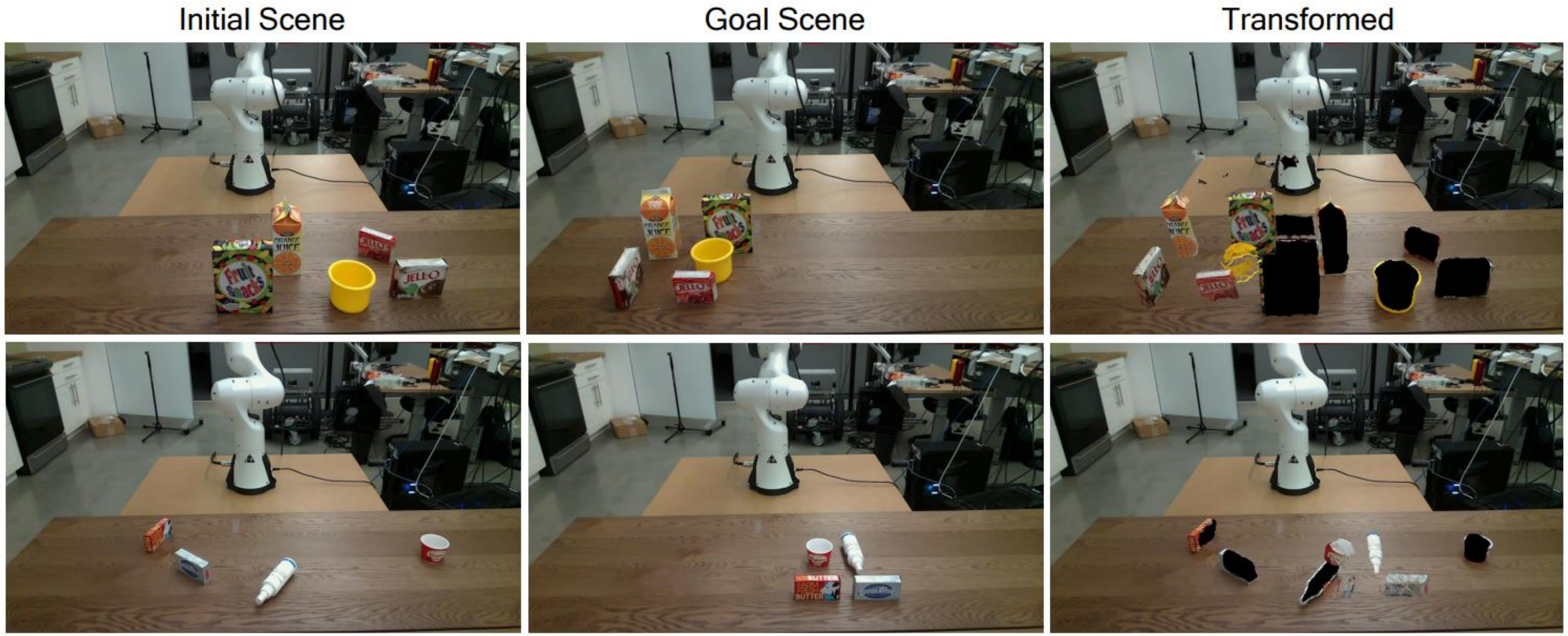
METHOD

Using optical flow to predict relative object transforms that aligns the current image and target image.



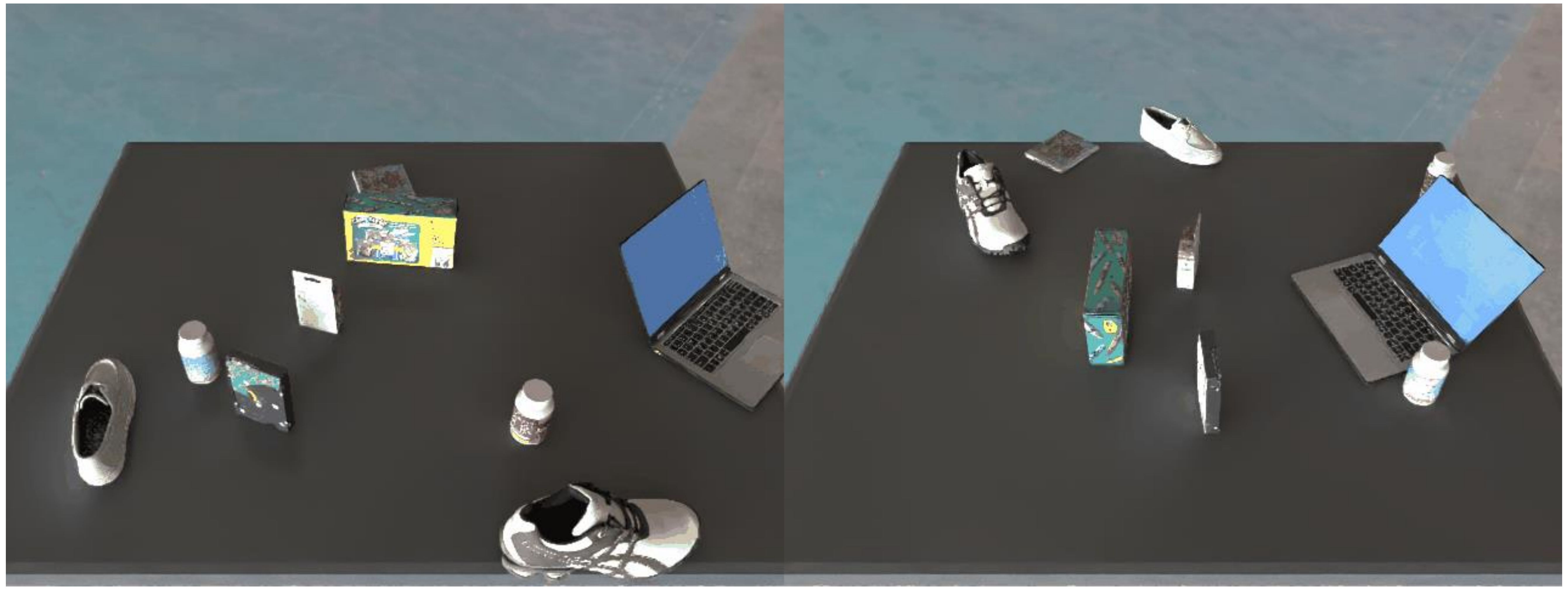
COMPUTING RELATIVE OBJECT TRANSFORMS

Use dense correspondence from optical flow and estimate transforms from 3D points



TRAINING DATA

Retraining RAFT on 54000 synthetic scenes where objects are moved around



PLANNING METHOD

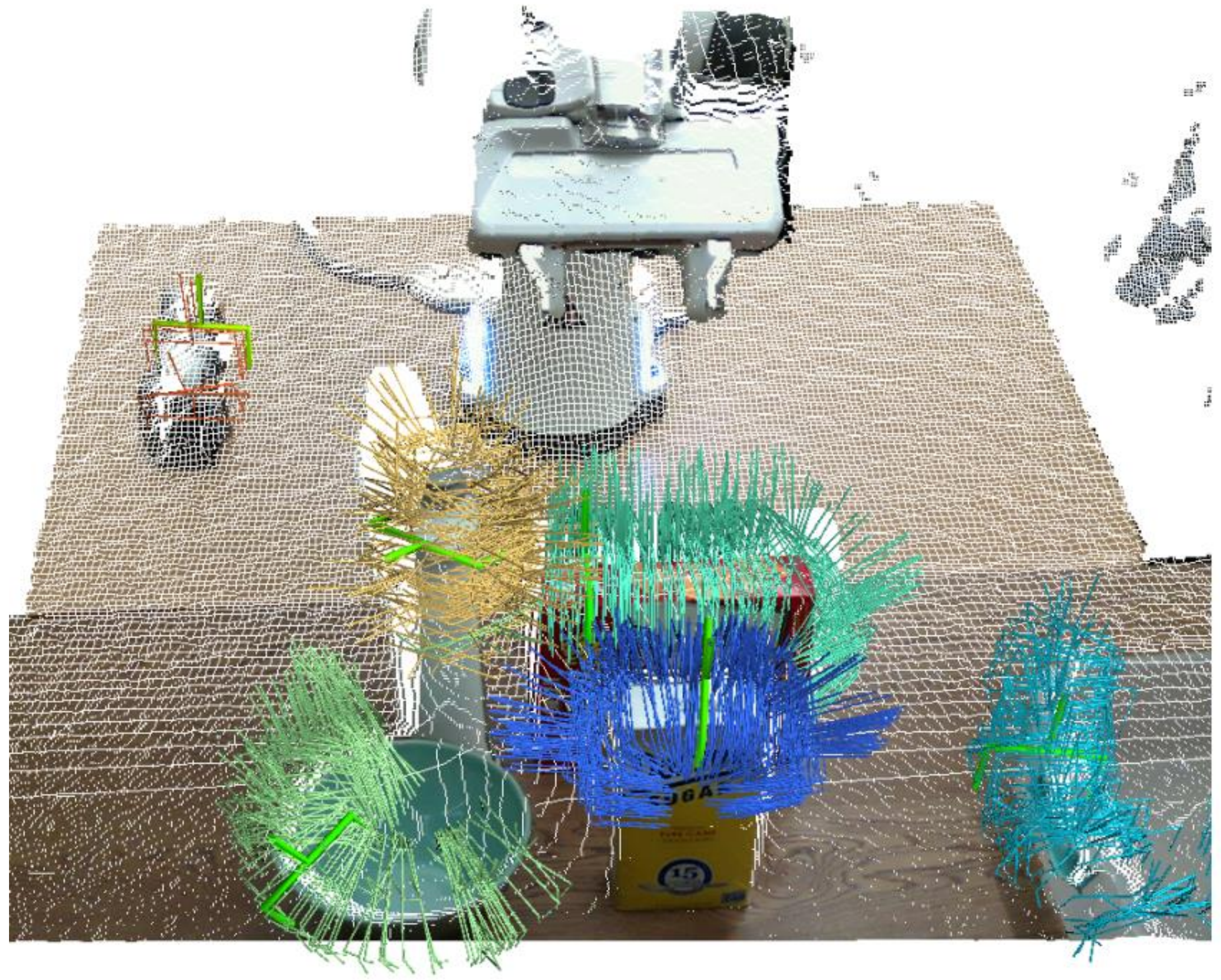
Simple algorithms work!

- Compute dense flow between current image and goal image.
- Estimate the relative rotation translation of each object.
- Pick the object with largest displacement that can be placed directly to the final goal.
- If none of the objects can be moved, pick a random object and place it in a random collision free placement.
- Repeat until all the predicted transforms are small.

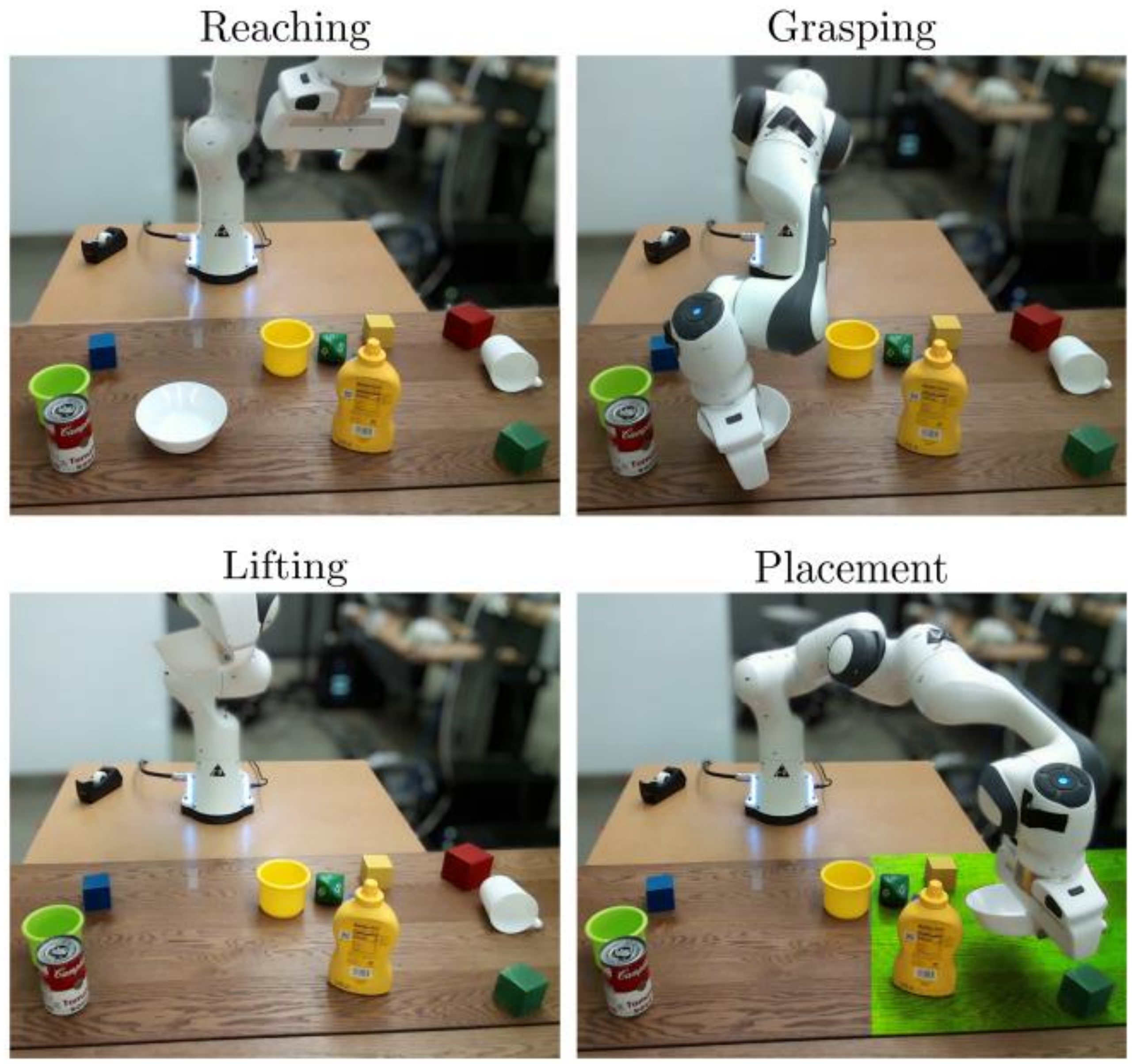
PERCEPTION COMPONENTS FOR OBJECT REARRANGEMENT

We build prior works as building blocks for rearrangement stack

Instance Segmentation



Grasp Generation



Collision Checking and Planning

CONCLUSIONS

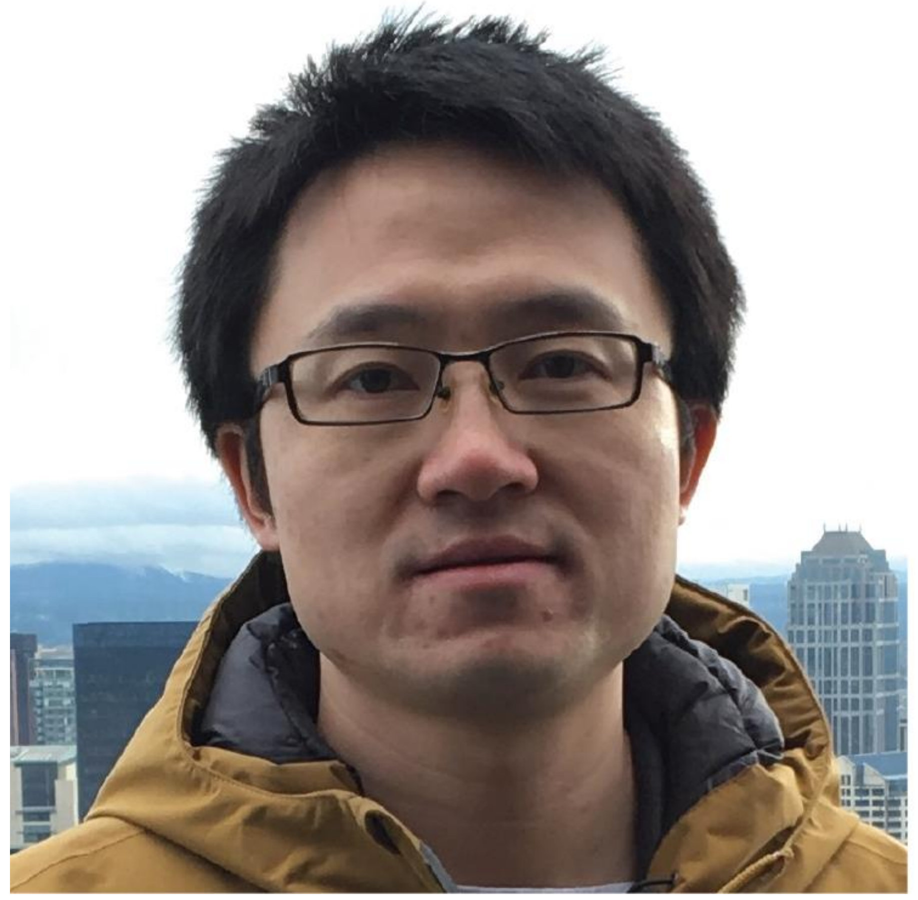
- Object-centric models are useful building blocks that can be used in different tasks.
- Modularity is the key to have scalable manipulation system.
- Depth and point cloud has small gap between simulation and real world.
- Training directly for the final task achieves better performance than training for implicit objectives.

THANK YOU

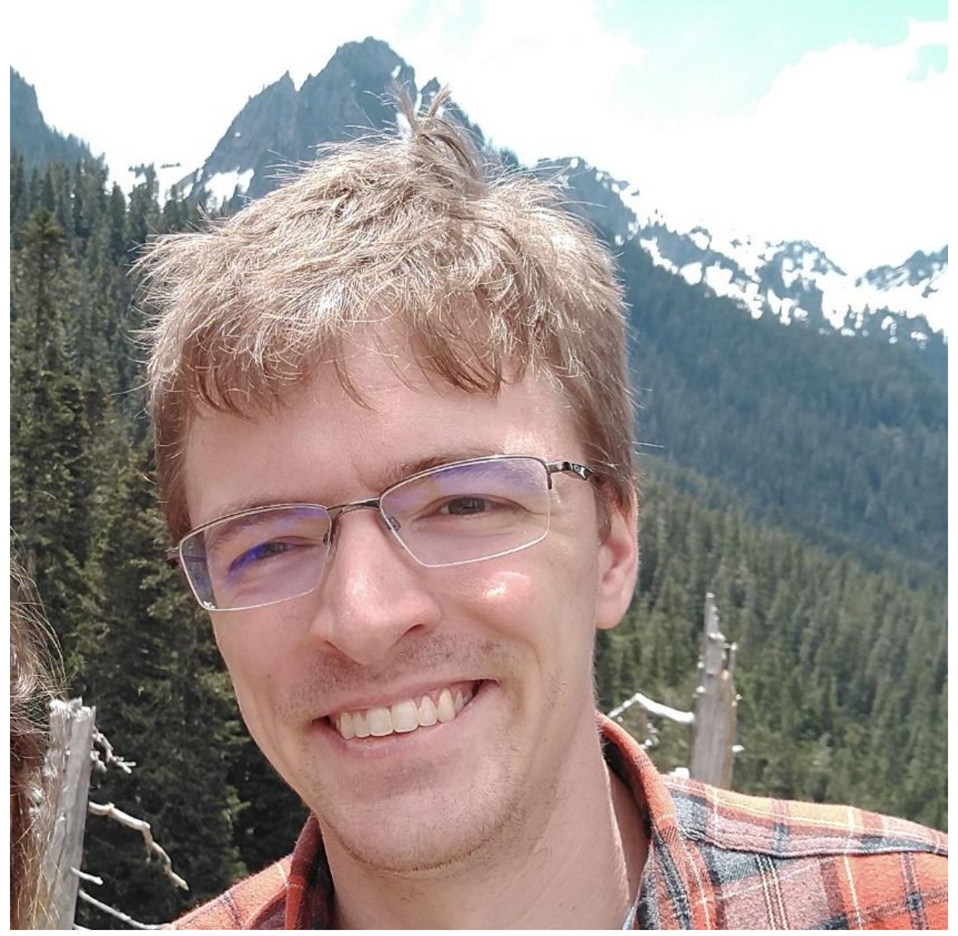
Questions?

Dieter Fox

Clemens Eppner



Yu Xiang



Chris Paxton

Adithya Murali

Wei Yang

Chris Xie

Mike Danielczuk

Martin Sundermeyer

Ankit Goyal