Ph. D. Thesis Defense

Learning for Vision-Based Object Manipulation: A Shape Recognition-Based Approach

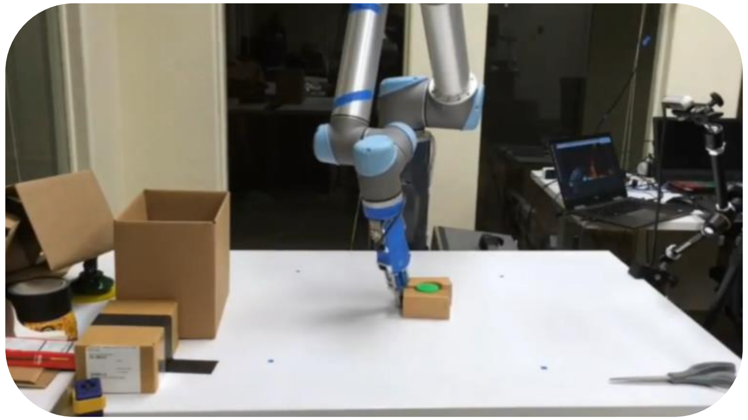
Seungyeon Kim
ksy@robotics.snu.ac.kr
https://seungyeon-k.github.io/

Supervisor: Prof. Frank C. Park

2023. 10. 27

Vision-based Object Manipulation

Grasping



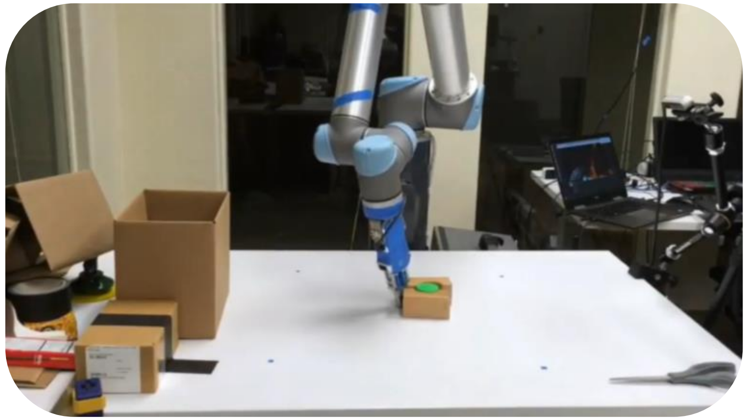
Pushing

Tossing

Sundermeyer, Martin, et al. "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes." 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021.
Zeng, Andy, et al. "Learning synergies between pushing and grasping with self-supervised deep reinforcement learning." 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018.
Zeng, Andy, et al. "Tossingbot: Learning to throw arbitrary objects with residual physics." IEEE Transactions on Robotics 36.4 (2020): 1307-1319.

Vision-based Object Manipulation

Grasping



Pushing

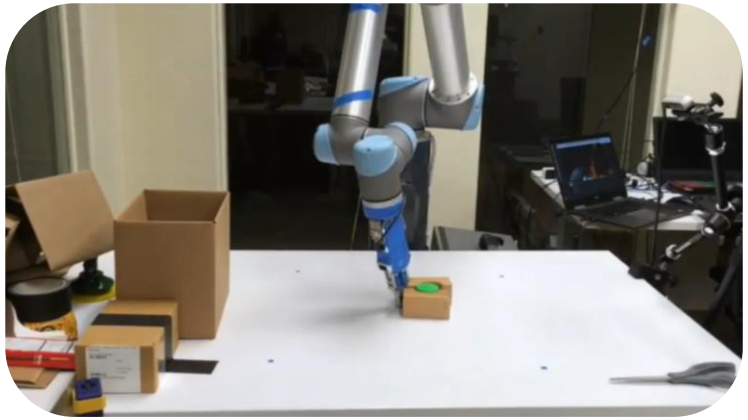
Tossing

Current challenges lie on manipulating **unknown object** with **only vision sensor data**.

Sundermeyer, Martin, et al. "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes." 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021.
Zeng, Andy, et al. "Learning synergies between pushing and grasping with self-supervised deep reinforcement learning." 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018.
Zeng, Andy, et al. "Tossingbot: Learning to throw arbitrary objects with residual physics." IEEE Transactions on Robotics 36.4 (2020): 1307-1319.

Vision-based Object Manipulation

Grasping



Pushing

Tossing

Current challenges lie on manipulating **unknown object** with **only vision sensor data**.

Sundermeyer, Martin, et al. "Contact-graspnet: Efficient 6-dof grasp generation in cluttered scenes." 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021.
Zeng, Andy, et al. "Learning synergies between pushing and grasping with self-supervised deep reinforcement learning." 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018.
Zeng, Andy, et al. "Tossingbot: Learning to throw arbitrary objects with residual physics." IEEE Transactions on Robotics 36.4 (2020): 1307-1319.

End-to-end Approaches

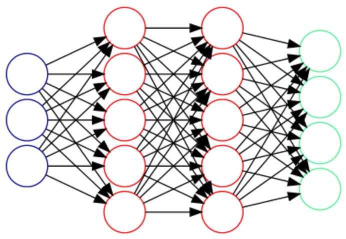
Grasping

Pushing

End-to-end Approaches

Grasping

Vision data

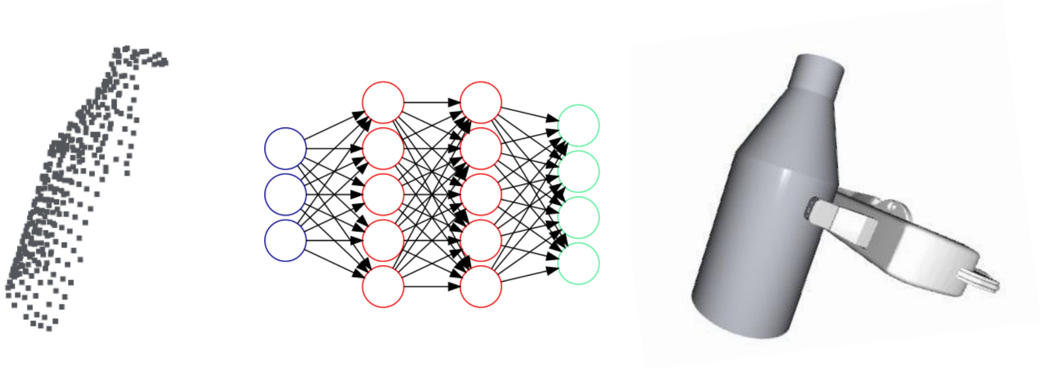


Grasp pose

Pushing

End-to-end Approaches

Grasping



Vision data

Grasp pose

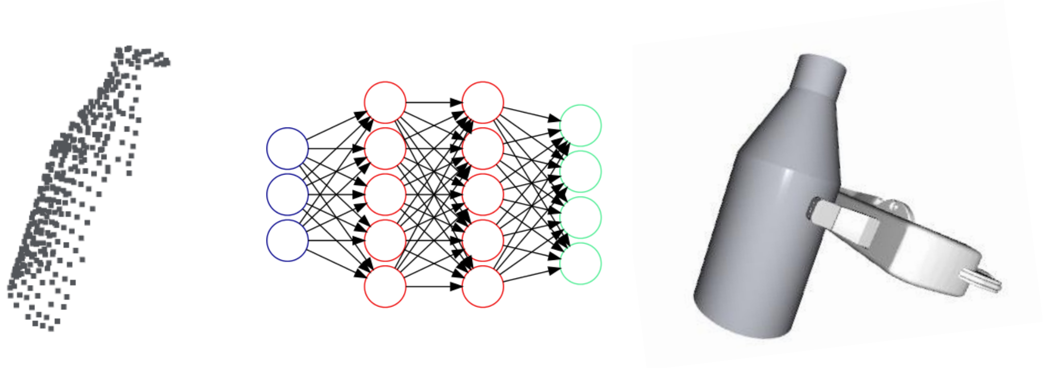
(~10,000,000 pairs)

- Require large amounts of training data.

Pushing

End-to-end Approaches

Grasping



Vision data

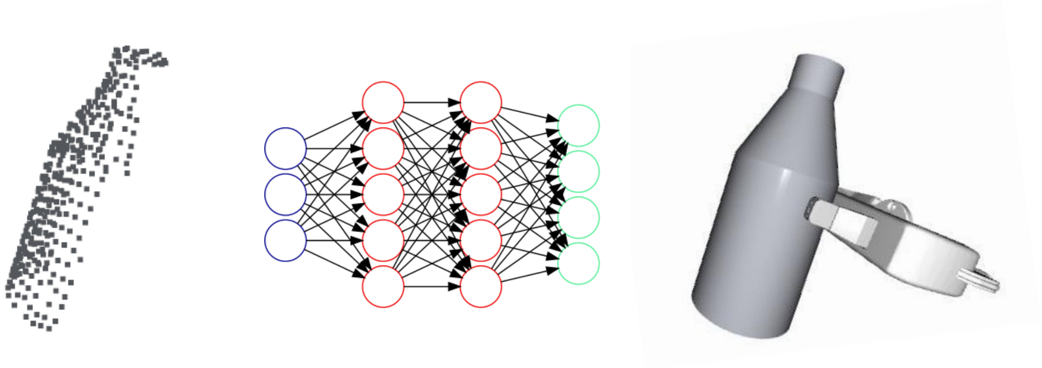
Grasp pose

- Require large amounts of training data.
- The trained network will only work reliably for the gripper used to collect the training data.

Pushing

End-to-end Approaches

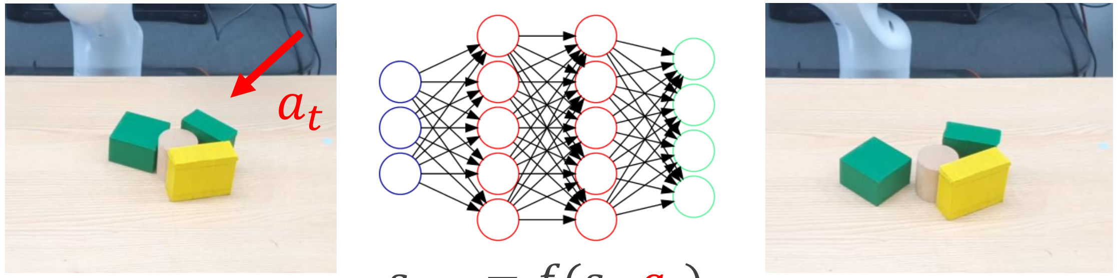
Grasping



Vision data Grasp pose

- Require large amounts of training data.
- The trained network will only work reliably for the gripper used to collect the training data.

Pushing

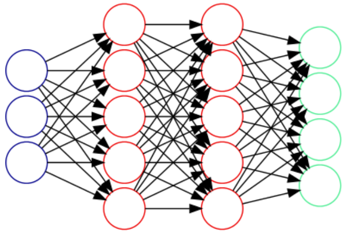


$s_{t+1} = f(s_t, a_t)$

Vision data s_t Next vision data s_{t+1}

End-to-end Approaches

Grasping

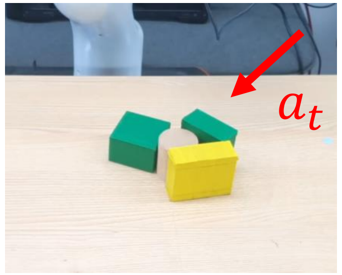


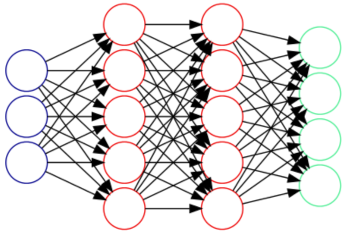
Vision data

Grasp pose

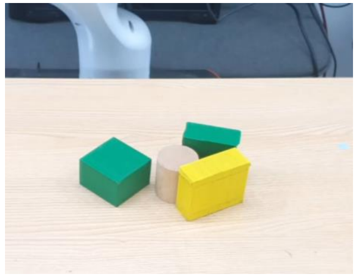
- Require large amounts of training data.
- The trained network will only work reliably for the gripper used to collect the training data.

Pushing





$s_{t+1} = f(s_t, a_t)$



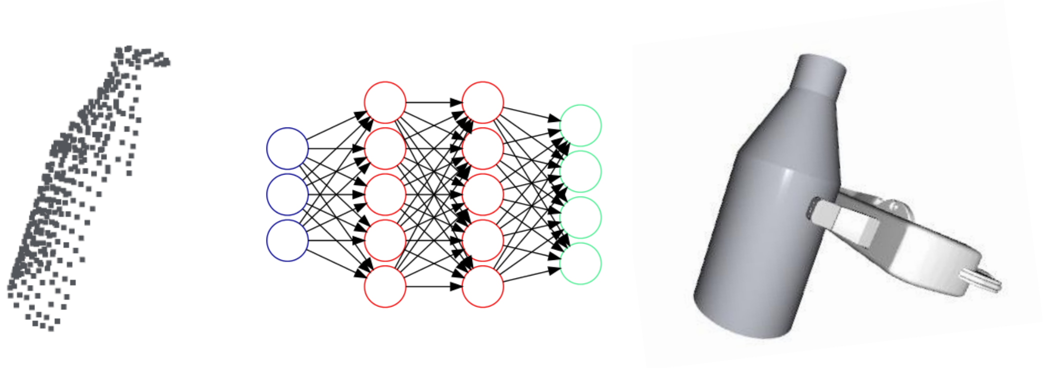
Vision data s_t

Next vision data s_{t+1}

- Require large amounts of training data.

End-to-end Approaches

Grasping

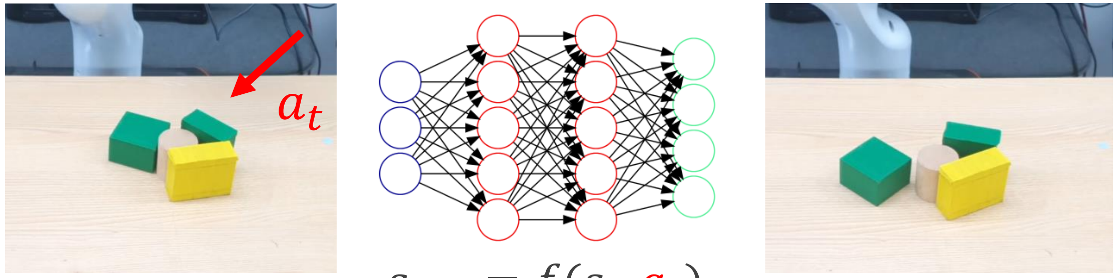


Vision data

Grasp pose

- Require large amounts of training data.
- The trained network will only work reliably for the gripper used to collect the training data.

Pushing



$s_{t+1} = f(s_t, a_t)$

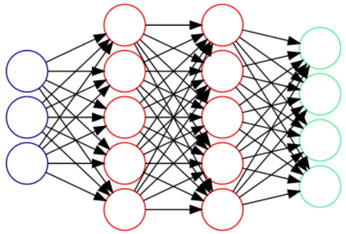
Vision data s_t

Next vision data s_{t+1}

- Require large amounts of training data.
- Generalization performance is less-than-satisfying.

End-to-end Approaches

Grasping

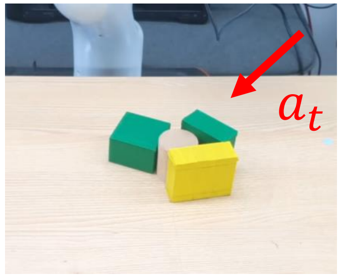


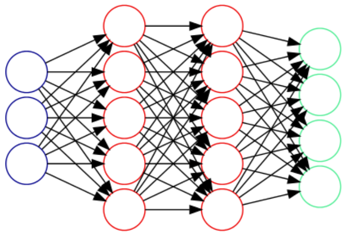
Vision data

Grasp pose

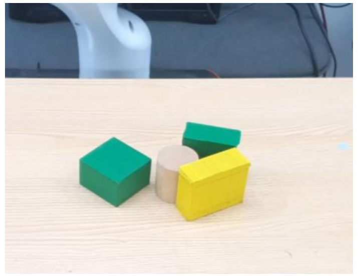
- Require large amounts of training data.
- The trained network will only work reliably for the gripper used to collect the training data.

Pushing





$s_{t+1} = f(s_t, a_t)$



Vision data s_t

Next vision data s_{t+1}

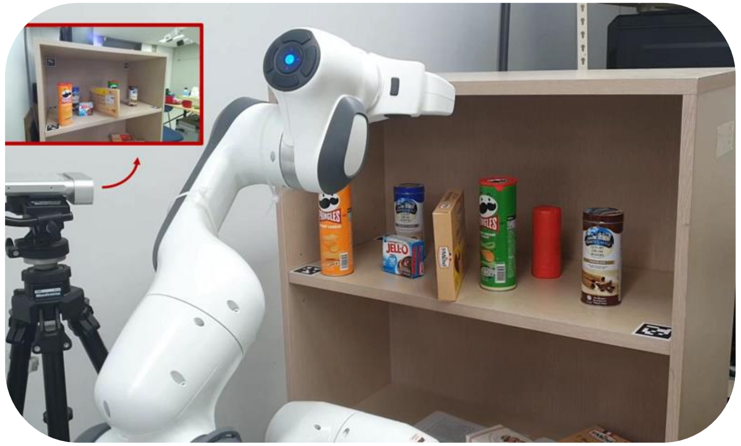
- Require large amounts of training data.
- Generalization performance is less-than-satisfying.

The primary contribution lies in employing **shape recognition** to address the challenges!

Shape Recognition-based Approaches

DSQNet
(S. Kim, et al., T-ASE’22)

SQPDNet
(S. Kim, et al., CoRL’22)



Search-for-Grasp
(S. Kim, et al. CoRL’23)

Shape Recognition-based Approaches

DSQNet
(S. Kim, et al., T-ASE'22)

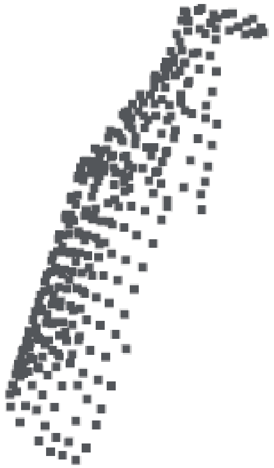
SQPDNet
(S. Kim, et al., CoRL'22)

Search-for-Grasp
(S. Kim, et al. CoRL'23)

Vision-based Grasping

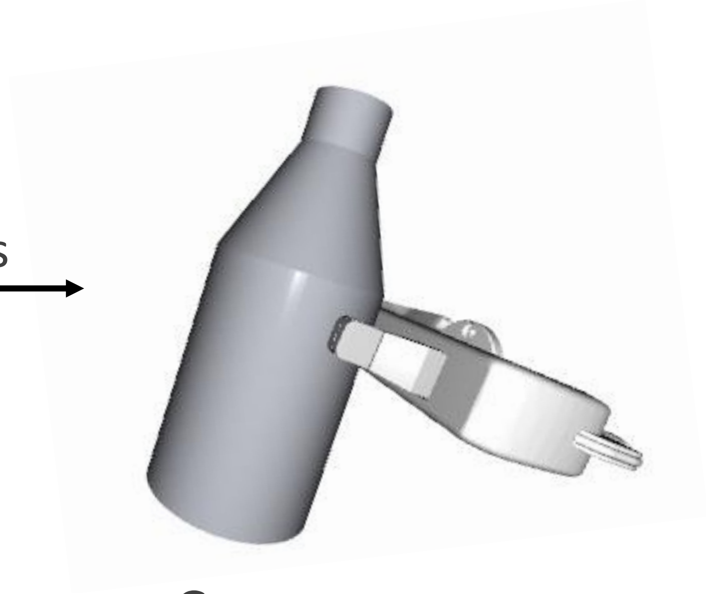
RGB-D camera

Target object



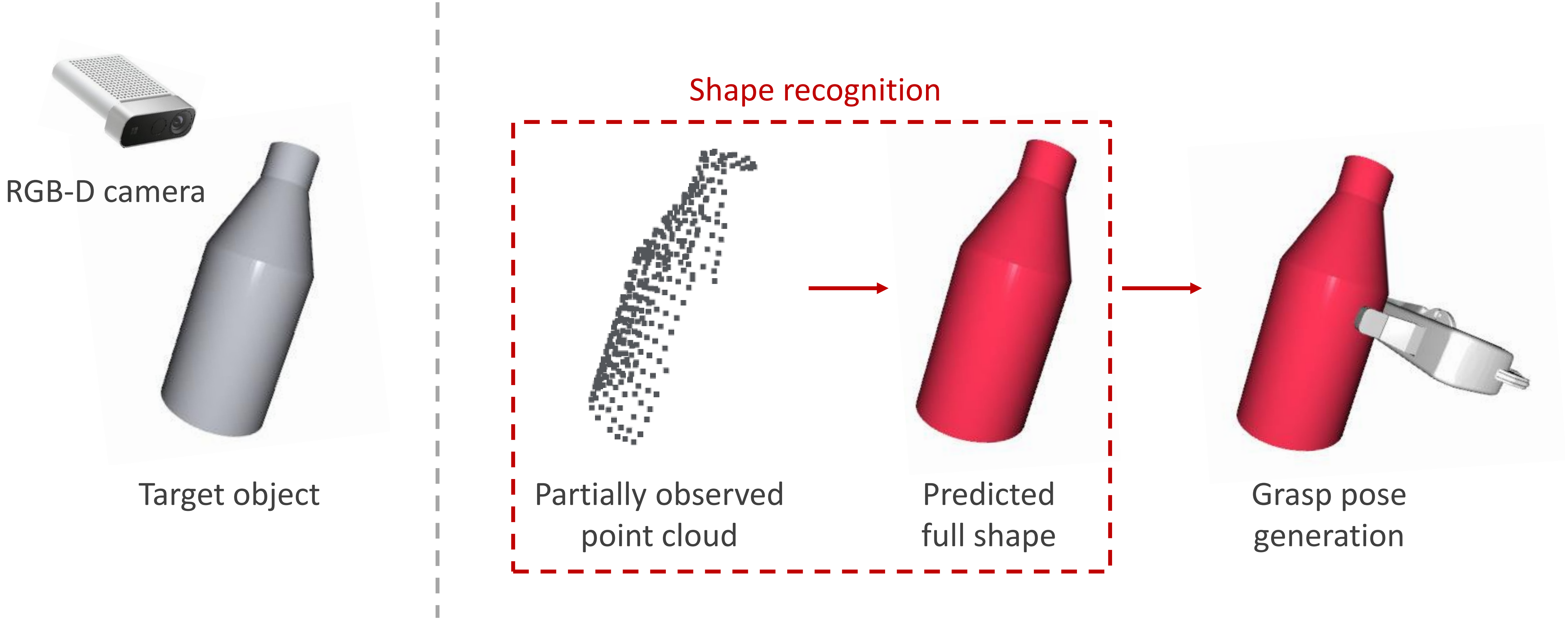
Partially observed point cloud

End-to-end methods



Grasp pose generation

Shape Recognition-based Grasping



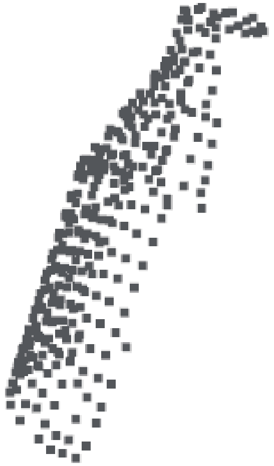
Shape Recognition-based Grasping

- Shape recognition requires considerably less data.

RGB-D camera

Target object

Shape recognition



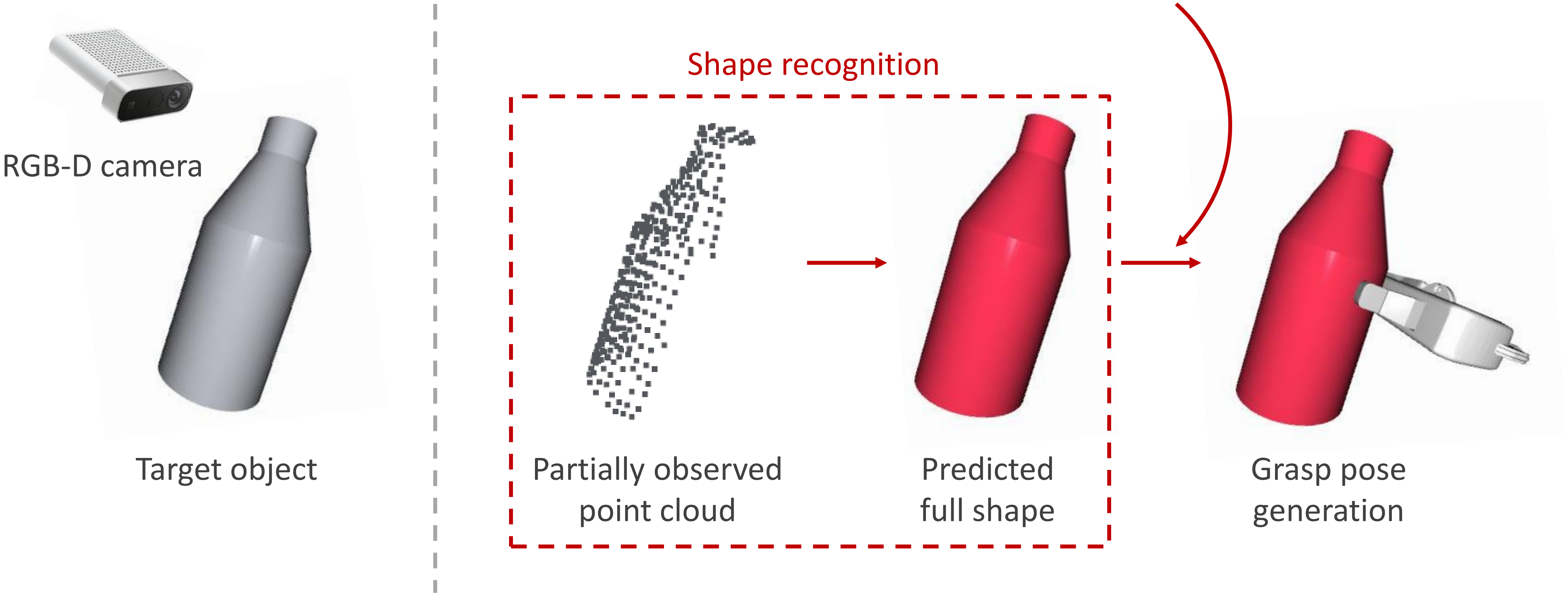
Partially observed point cloud

Predicted full shape

Grasp pose generation

Shape Recognition-based Grasping

- Shape recognition requires considerably less data.
- Grasp pose generation can be simply modified.



Shape Recognition-based Grasping

Shape expressiveness

Shape Recognition-based Grasping

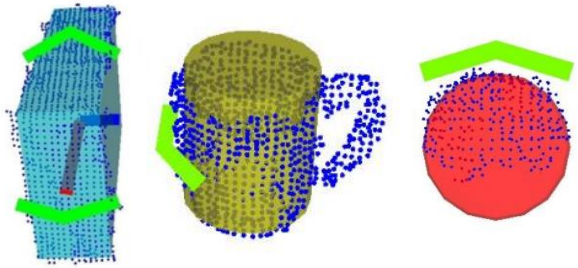
Shape expressiveness

Bounding box
(K. Huebner, et al., ICRA’08)

Shape Recognition-based Grasping

Shape expressiveness

Bounding box
(K. Huebner, et al., ICRA'08)



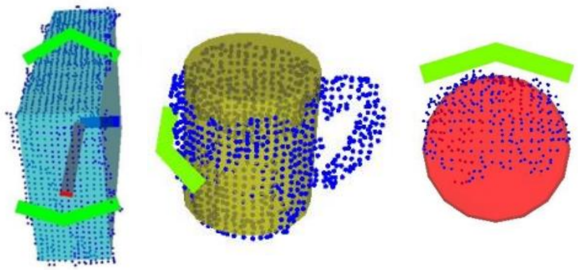
Box, cylinder, sphere
(S. Jain, et al., ICRA'16)

Shape Recognition-based Grasping

Shape expressiveness

Bounding box
(K. Huebner, et al., ICRA'08)

Superquadrics
(G. Vezzani, ICRA'17)



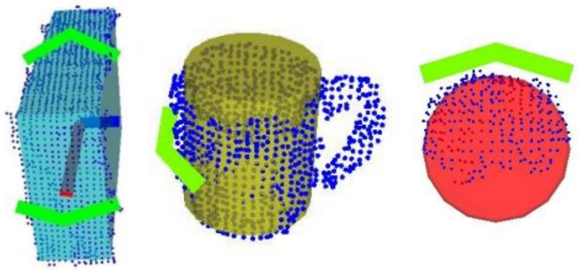
Box, cylinder, sphere
(S. Jain, et al., ICRA'16)

Shape Recognition-based Grasping

Shape expressiveness

Bounding box
(K. Huebner, et al., ICRA'08)

Superquadrics
(G. Vezzani, ICRA'17)



Box, cylinder, sphere
(S. Jain, et al., ICRA'16)

Custom templates
(Y. Lin, et al., ICRA'20)

Shape Recognition-based Grasping

Shape expressiveness

Bounding box

(K. Huebner, et al., ICRA'08)

Superquadrics

(G. Vezzani, ICRA'17)

Even simple everyday objects such as bottles and mugs often cannot be easily represented.

Box, cylinder, sphere

(S. Jain, et al., ICRA'16)

Custom primitives

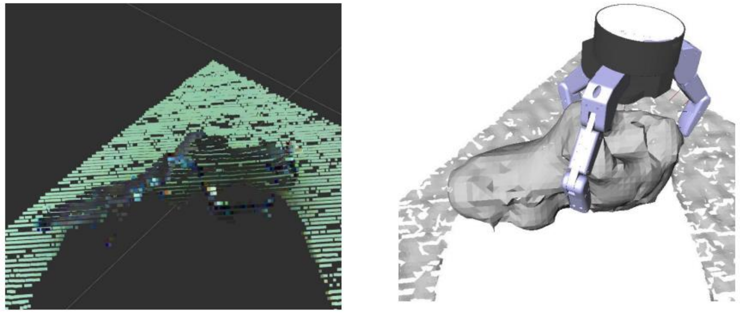
(Y. Lin, et al., ICRA'20)

Shape Recognition-based Grasping

Shape expressiveness

Bounding box
(K. Huebner, et al., ICRA'08)

Superquadrics
(G. Vezzani, ICRA'17)



Mesh reconstruction
(J. Varley, et al., IROS'17)

Even simple everyday objects such as bottles and mugs often cannot be easily represented.

Box, cylinder, sphere
(S. Jain, et al., ICRA'16)

Custom primitives
(Y. Lin, et al., ICRA'20)

Shape Recognition-based Grasping

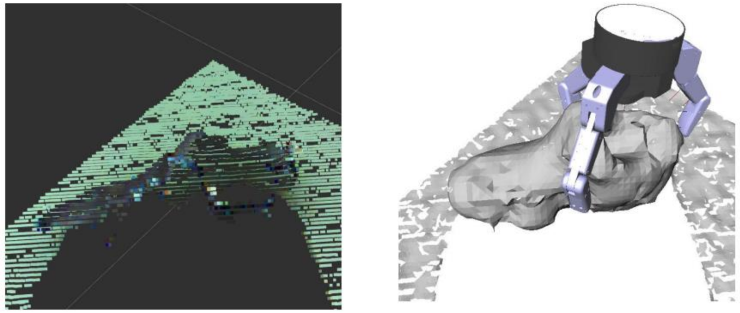
Shape expressiveness

Bounding box

(K. Huebner, et al., ICRA'08)

Superquadrics

(G. Vezzani, ICRA'17)



Mesh reconstruction

(J. Varley, et al., IROS'17)

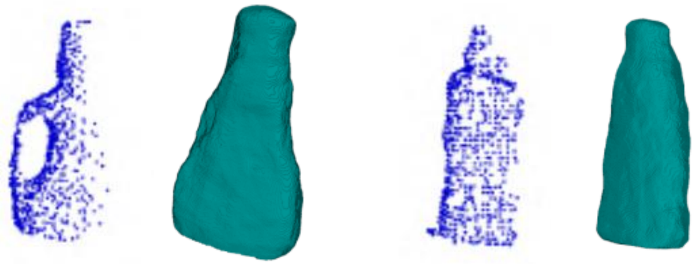
Even simple everyday objects such as bottles and mugs often cannot be easily represented.

Box, cylinder, sphere

(S. Jain, et al., ICRA'16)

Custom primitives

(Y. Lin, et al., ICRA'20)



Implicit function

(M. Van der Merwe, et al., ICRA'20)

Shape Recognition-based Grasping

Shape expressiveness

Bounding box

(K. Huebner, et al., ICRA'08)

Superquadrics

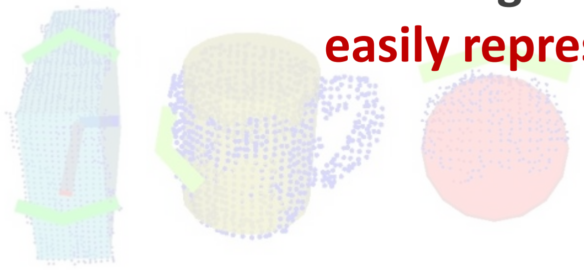
(G. Vezzani, ICRA'17)

Mesh reconstruction

(J. Varley, et al., IROS'17)

Even simple everyday objects such as bottles and mugs often cannot be easily represented.

Require a time-consuming planning stage to generate feasible grasp poses.



Box, cylinder, sphere

(S. Jain, et al., ICRA'16)

Custom primitives

(Y. Lin, et al., ICRA'20)

Implicit function

(M. Van der Merwe, et al., ICRA'20)

Shape Recognition-based Grasping

Shape expressiveness

Bounding box

(K. Huebner, et al., ICRA'08)

Superquadrics

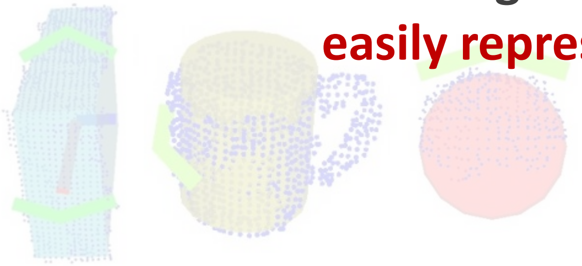
(G. Vezzani, ICRA'17)

Mesh reconstruction

(J. Varley, et al., IROS'17)

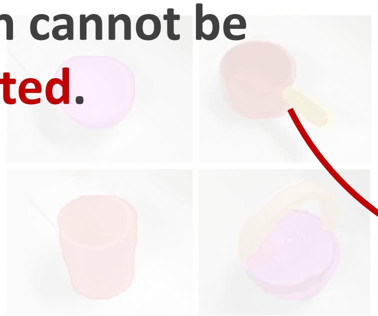
Even simple everyday objects such as bottles and mugs often cannot be **easily represented**.

Require a **time-consuming** planning stage to generate feasible grasp poses.



Box, cylinder, sphere

(S. Jain, et al., ICRA'16)



Custom primitives

(Y. Lin, et al., ICRA'20)

?

Implicit function

(M. Van der Merwe, et al., ICRA'20)

Deformable Superquadrics

Superquadrics

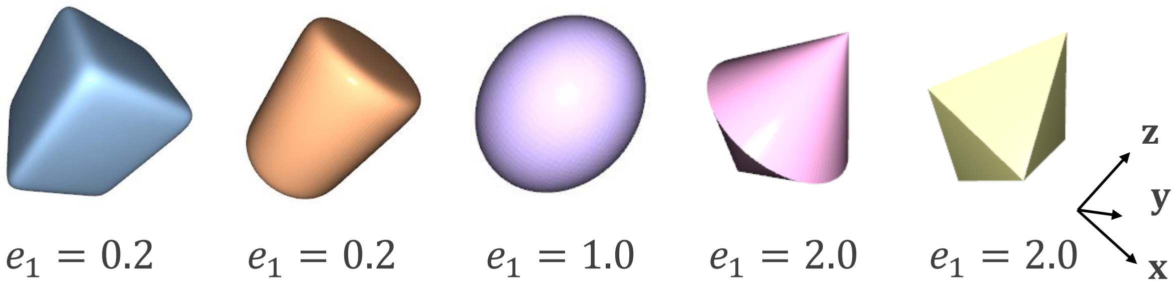
$$f(x,y,z) = \left(\left|\frac{x}{a_1}\right|^{\frac{2}{e_2}} + \left|\frac{y}{a_2}\right|^{\frac{2}{e_2}} \right)^{\frac{e_2}{e_1}} + \left|\frac{z}{a_3}\right|^{\frac{2}{e_1}} = 1$$

Deformable Superquadrics

Superquadrics

$$f(x,y,z) = \left(\left|\frac{x}{a_1}\right|^{\frac{2}{e_2}} + \left|\frac{y}{a_2}\right|^{\frac{2}{e_2}} \right)^{\frac{e_2}{e_1}} + \left|\frac{z}{a_3}\right|^{\frac{2}{e_1}} = 1$$

Box	Cylinder	Ellipsoid	Bicone	Octahedron



$e_1 = 0.2$	$e_1 = 0.2$	$e_1 = 1.0$	$e_1 = 2.0$	$e_1 = 2.0$
$e_2 = 0.2$	$e_2 = 1.0$	$e_2 = 1.0$	$e_2 = 0.2$	$e_2 = 1.0$

$\mathbf{a} = (a_1, a_2, a_3)$: size parameters

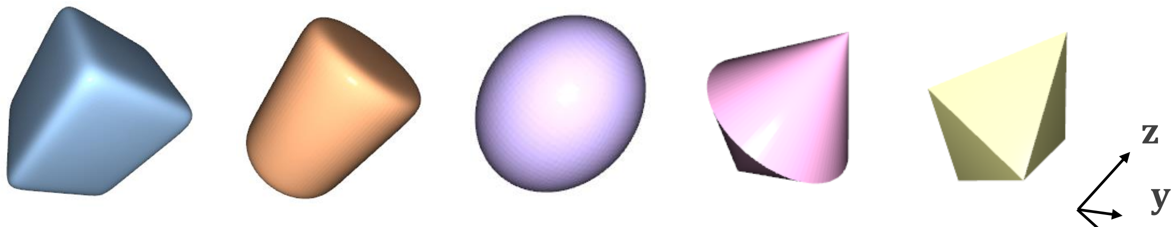
$\mathbf{e} = (e_1, e_2)$: shape parameters

Deformable Superquadrics

Superquadrics

$$f(x,y,z)=\left(\left|\frac{x}{a_1}\right|^{\frac{2}{e_2}}+\left|\frac{y}{a_2}\right|^{\frac{2}{e_2}}\right)^{\frac{e_2}{e_1}}+\left|\frac{z}{a_3}\right|^{\frac{2}{e_1}}=1$$

Box	Cylinder	Ellipsoid	Bicone	Octahedron



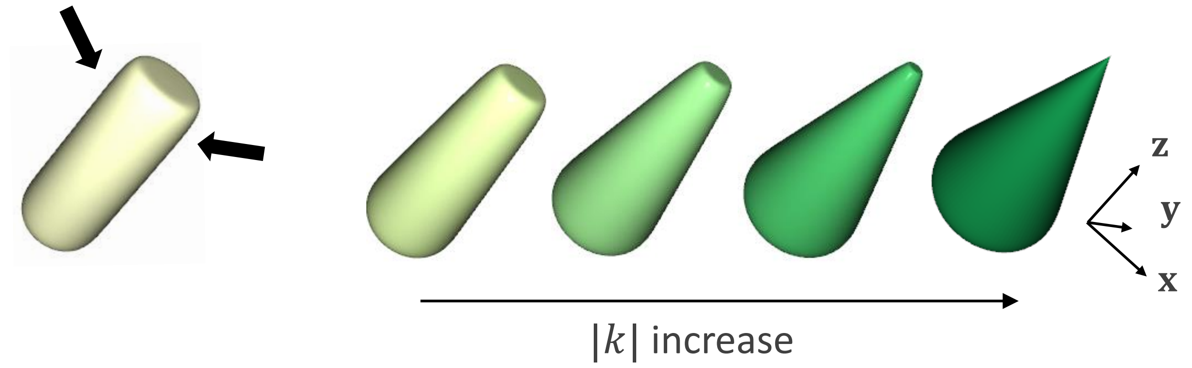
$e_1=0.2$	$e_1=0.2$	$e_1=1.0$	$e_1=2.0$	$e_1=2.0$
$e_2=0.2$	$e_2=1.0$	$e_2=1.0$	$e_2=0.2$	$e_2=1.0$

$\mathbf{a}=(a_1,a_2,a_3)$: size parameters

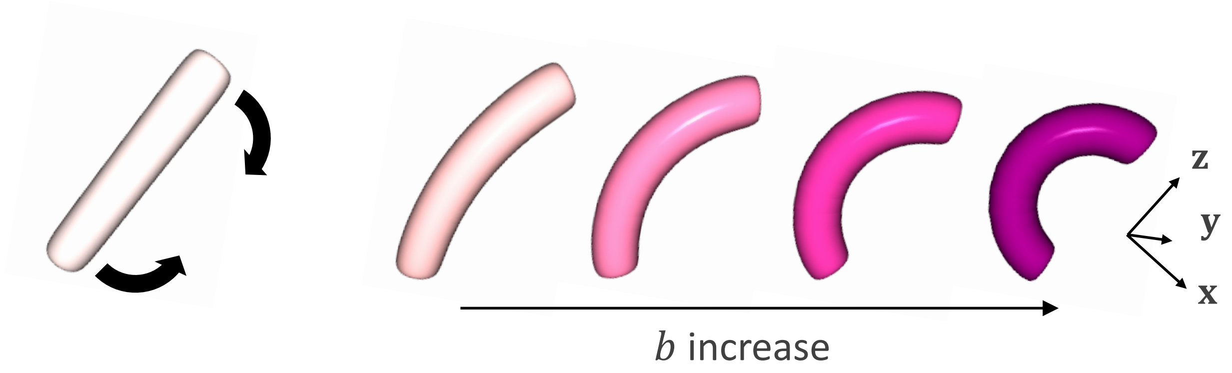
$\mathbf{e}=(e_1,e_2)$: shape parameters

Deformable Superquadrics

Tapering



Bending

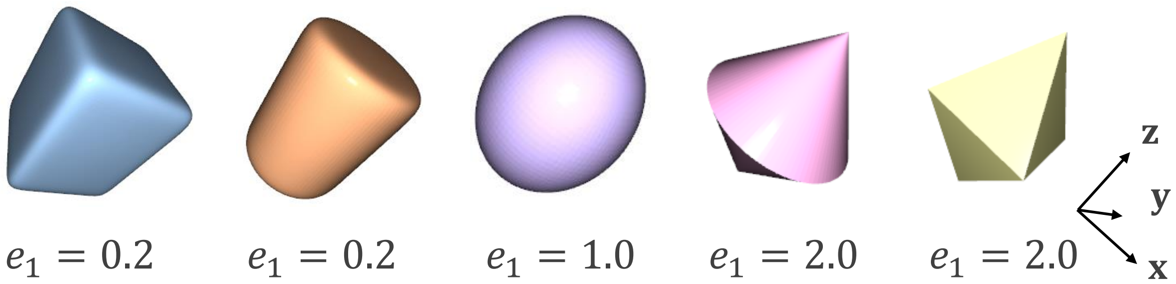


Deformable Superquadrics

Superquadrics

$$f(x,y,z) = \left(\left| \frac{x}{a_1} \right|^{\frac{2}{e_2}} + \left| \frac{y}{a_2} \right|^{\frac{2}{e_2}} \right)^{\frac{e_2}{e_1}} + \left| \frac{z}{a_3} \right|^{\frac{2}{e_1}} = 1$$

Box | Cylinder | Ellipsoid | Bicone | Octahedron



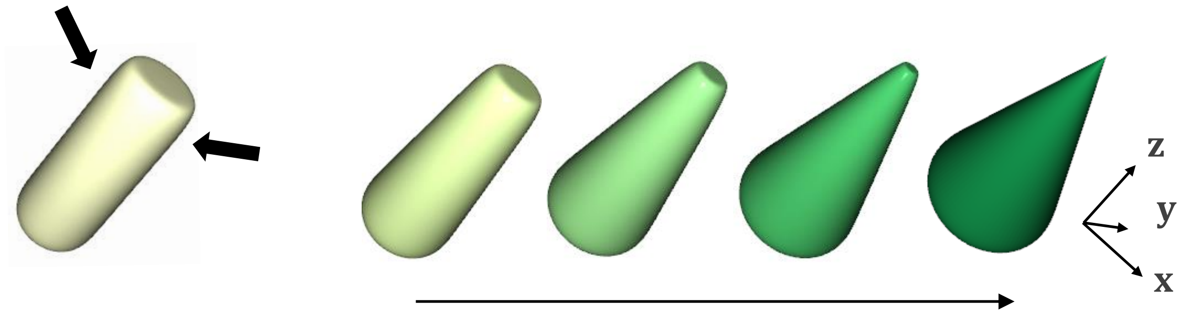
Box	Cylinder	Ellipsoid	Bicone	Octahedron
$e_1 = 0.2$	$e_1 = 0.2$	$e_1 = 1.0$	$e_1 = 2.0$	$e_1 = 2.0$
$e_2 = 0.2$	$e_2 = 1.0$	$e_2 = 1.0$	$e_2 = 0.2$	$e_2 = 1.0$

$\mathbf{a} = (a_1, a_2, a_3)$: size parameters

$\mathbf{e} = (e_1, e_2)$: shape parameters

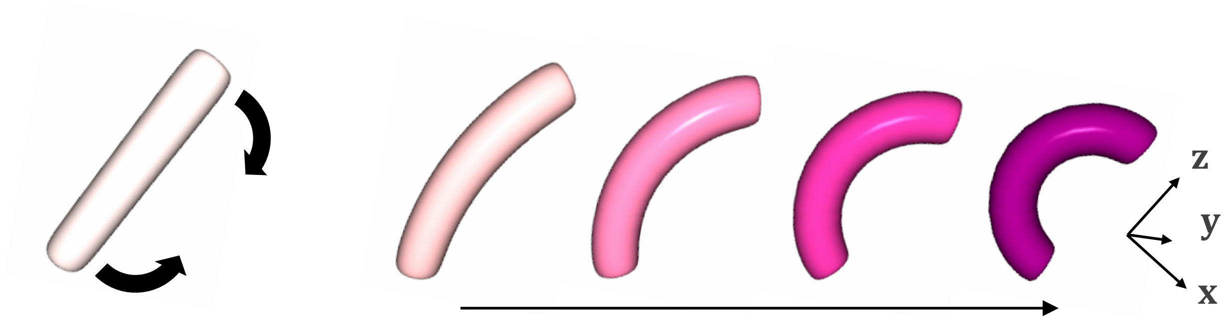
Deformable Superquadrics

Tapering with tapering parameter k



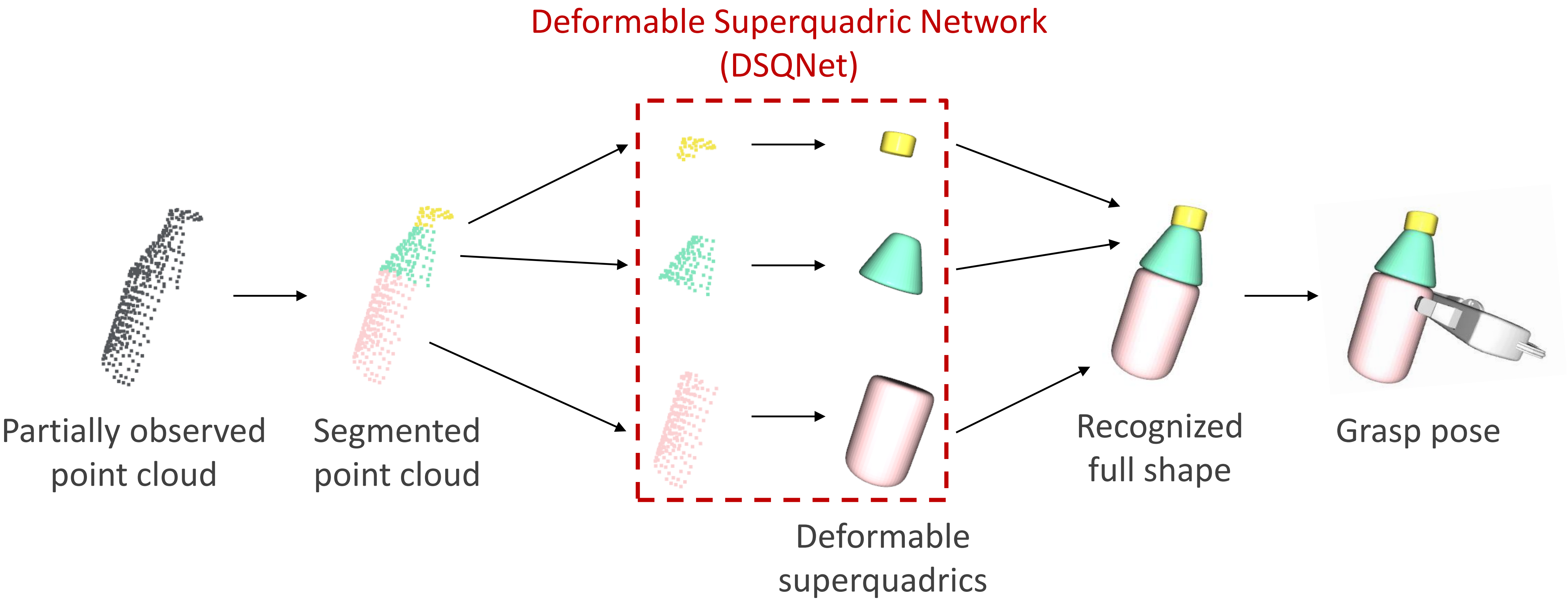
$|k|$ increase

Bending with bending parameters (b, α)



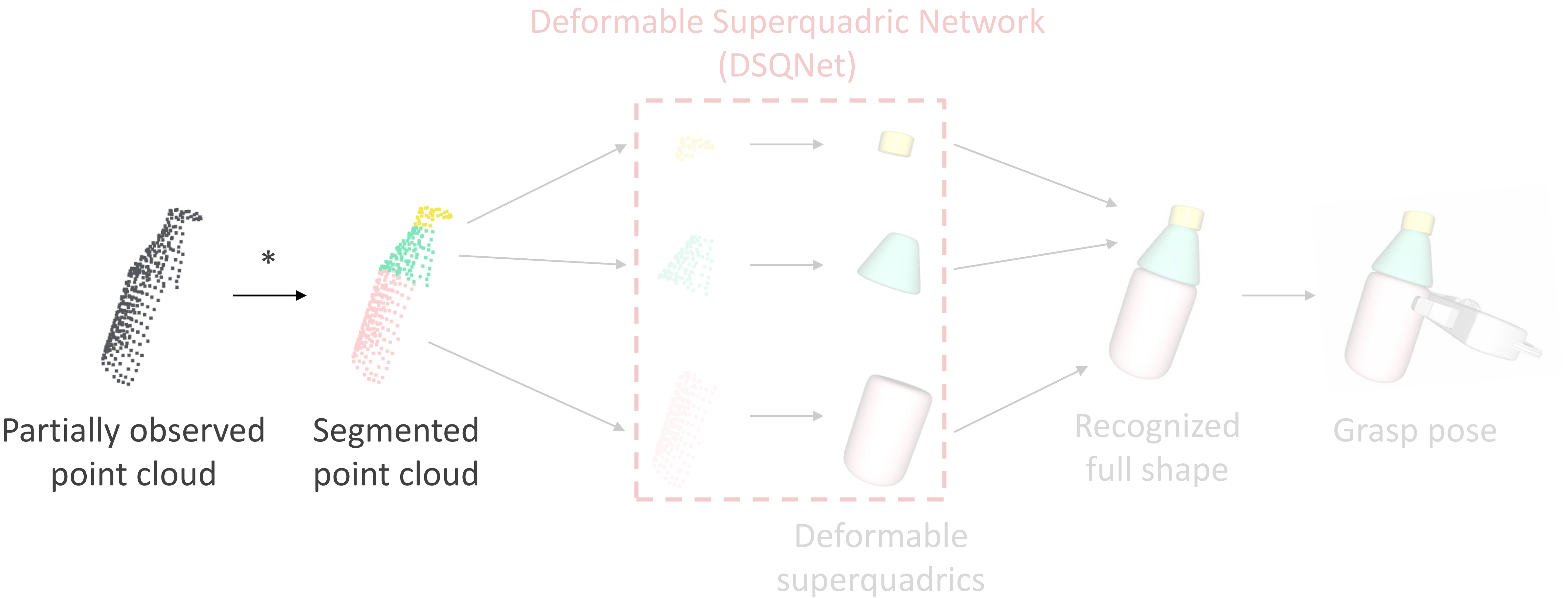
b increase

Our Method



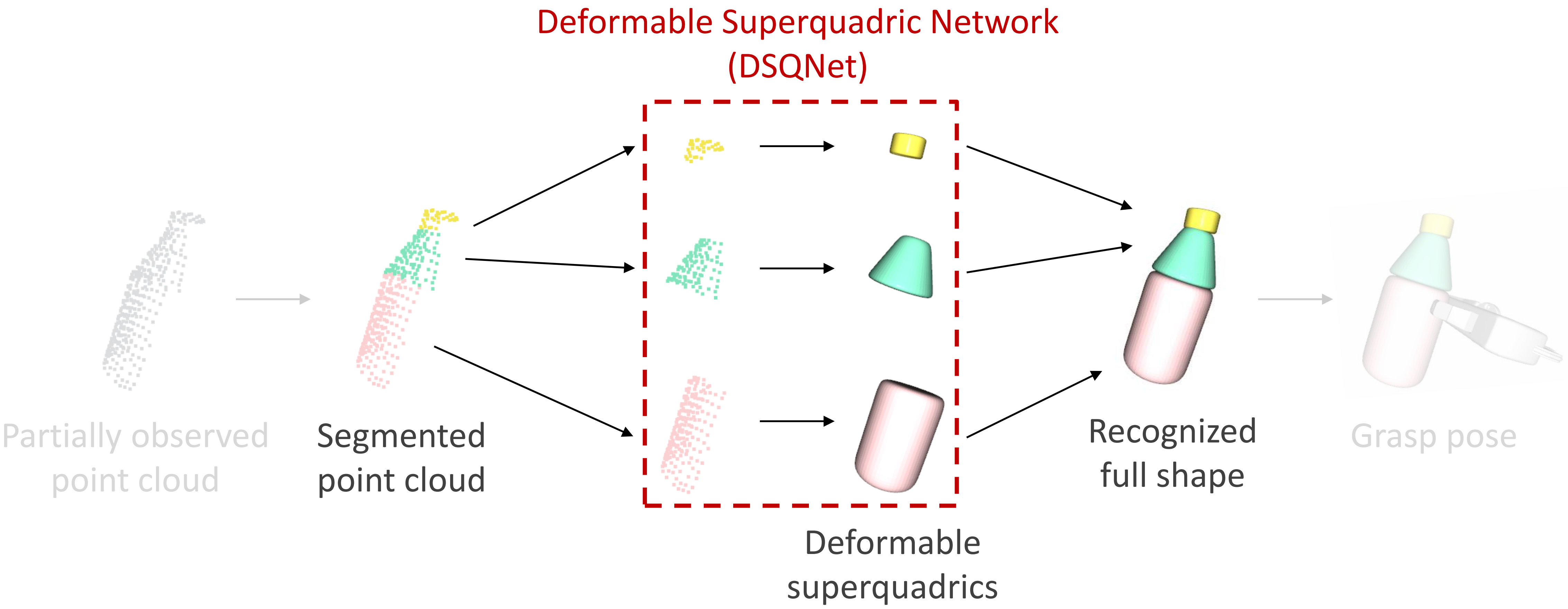
Deformable Superquadric Network
(DSQNet)
Partially observed
point cloud
Segmented
point cloud
Deformable
superquadrics
Recognized
full shape
Grasp pose

Our Method

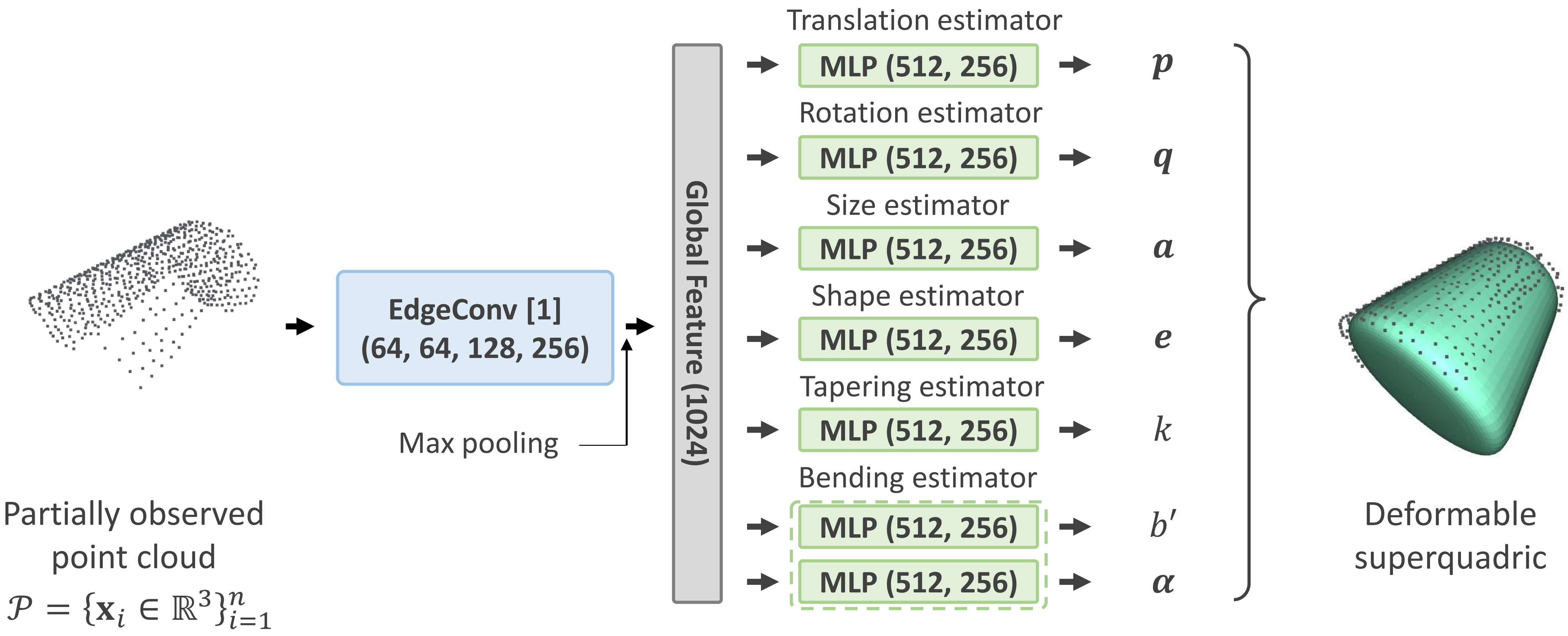


* Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," Acm Transactions On Graphics (tog), vol. 38, no. 5, pp. 1–12, 2019.

Our Method

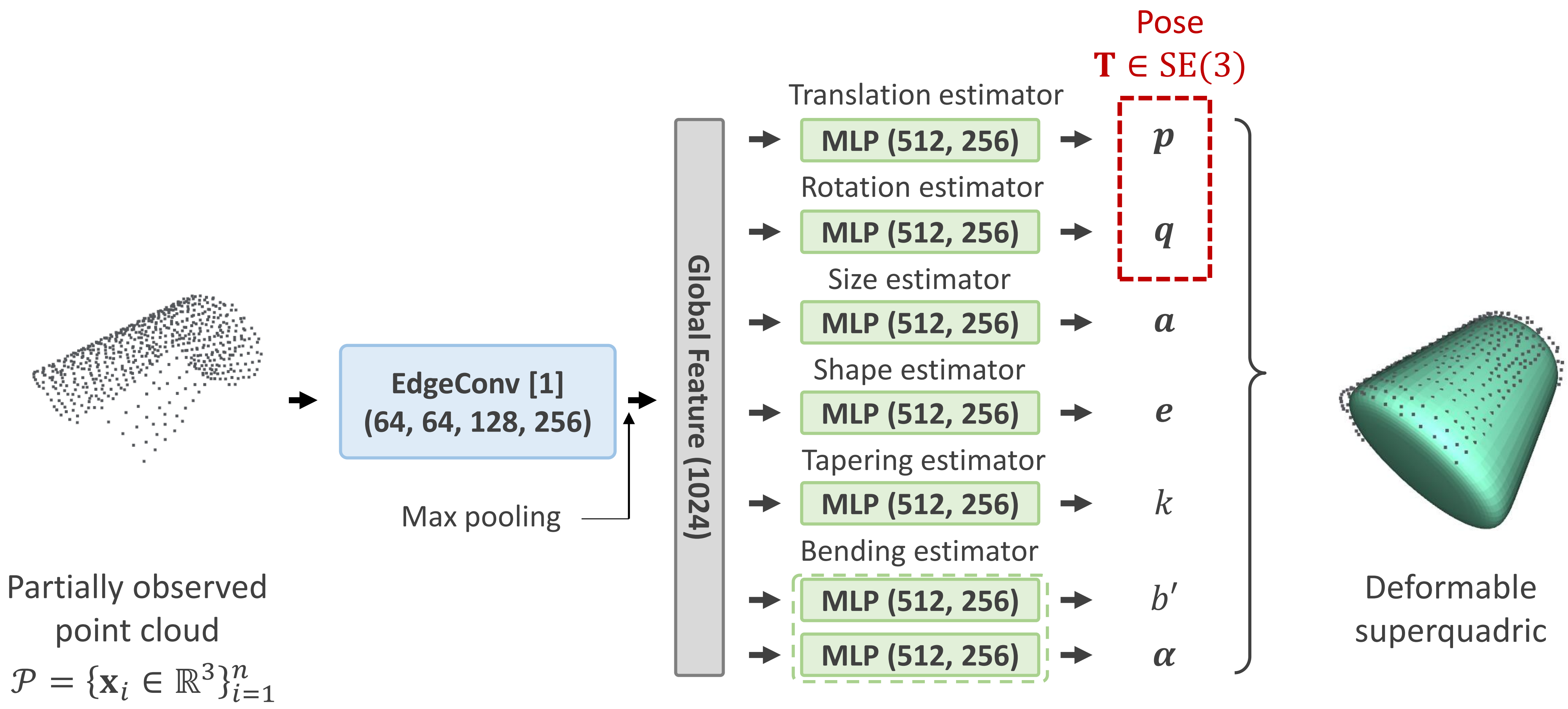

Deformable Superquadric Network
(DSQNet)
Partially observed
point cloud
Segmented
point cloud
Deformable
superquadrics
Recognized
full shape
Grasp pose

Deformable Superquadric Network



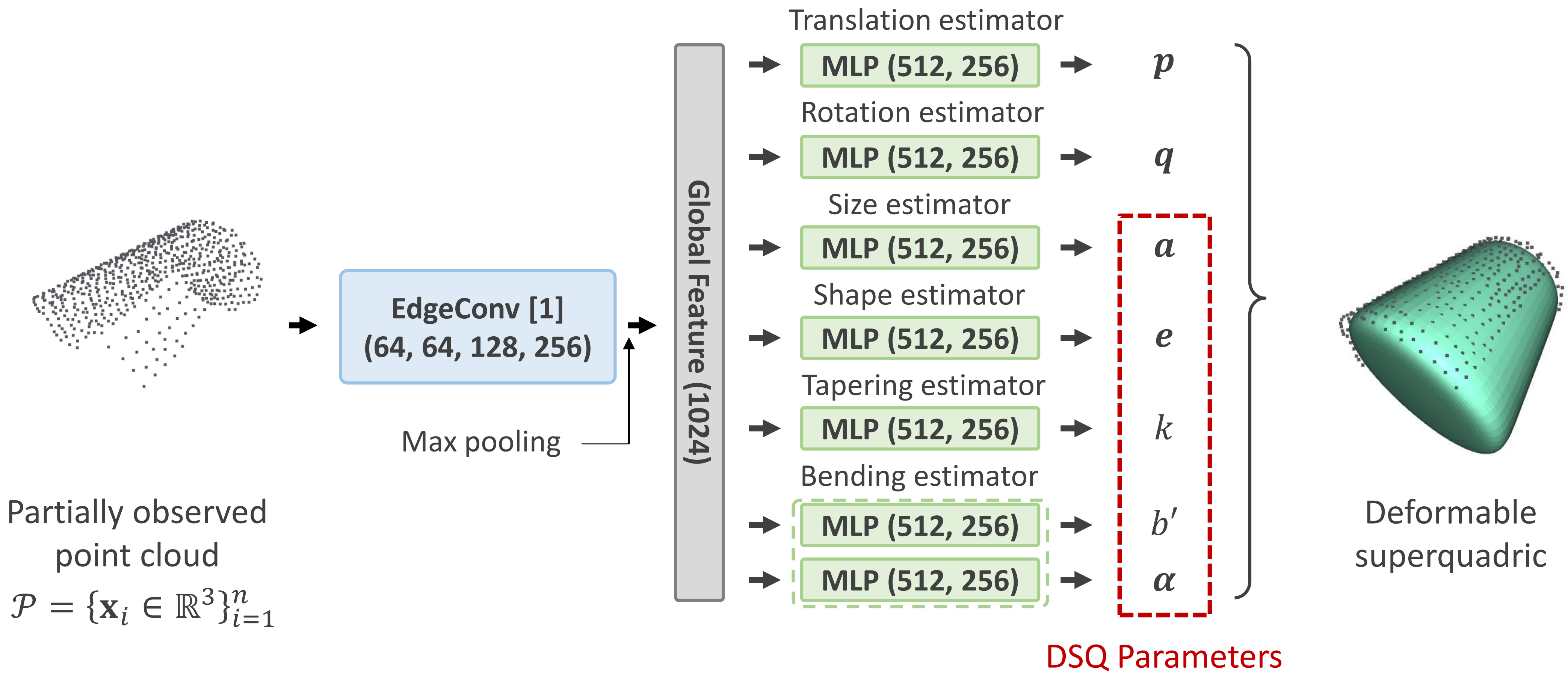
[1] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," Acm Transactions On Graphics (tog), vol. 38, no. 5, pp. 1–12, 2019.

Deformable Superquadric Network



[1] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," Acm Transactions On Graphics (tog), vol. 38, no. 5, pp. 1–12, 2019.

Deformable Superquadric Network



[1] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, “Dynamic graph cnn for learning on point clouds,” Acm Transactions On Graphics (tog), vol. 38, no. 5, pp. 1–12, 2019.

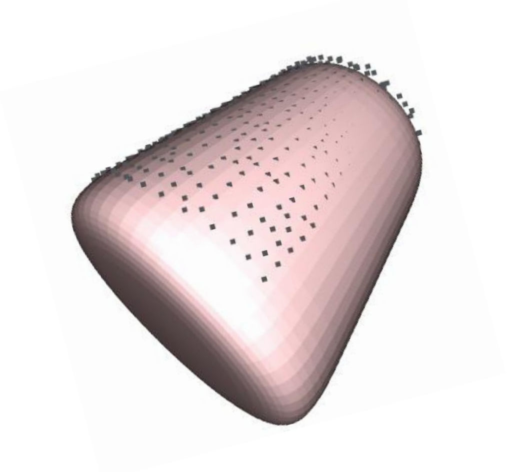
Deformable Superquadric Network

Partially observed point cloud

Deformable Superquadric Network

Partially observed point cloud

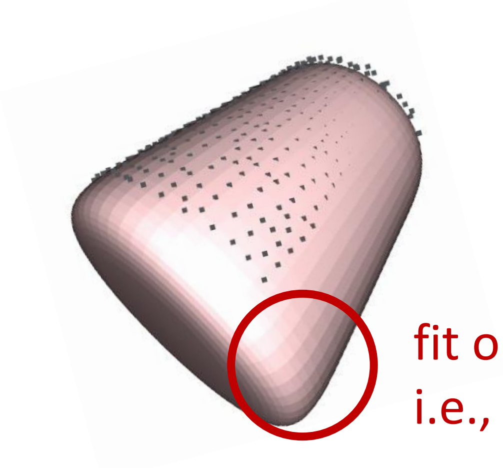
Optimization



Deformable Superquadric Network

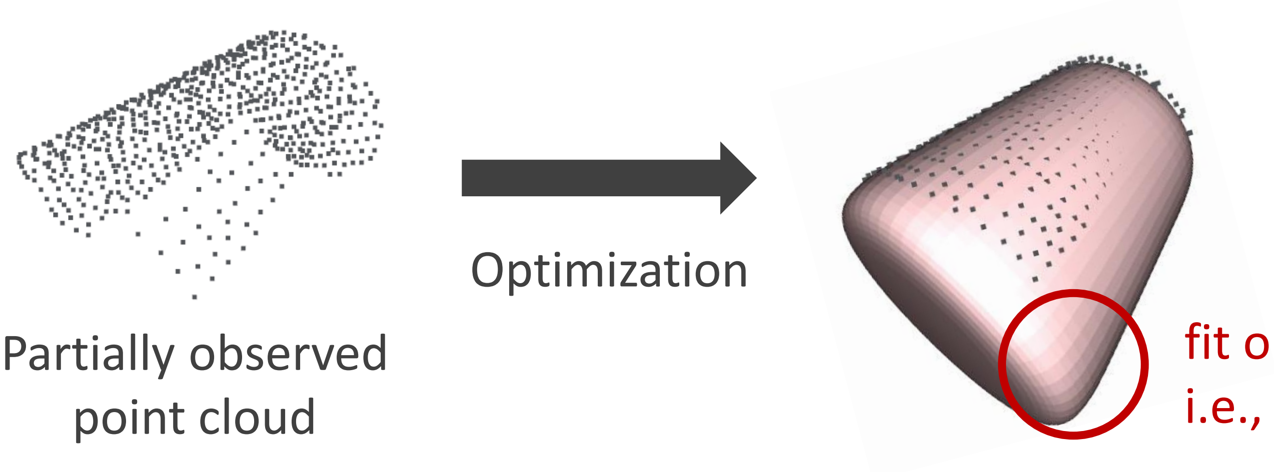
Partially observed point cloud

Optimization



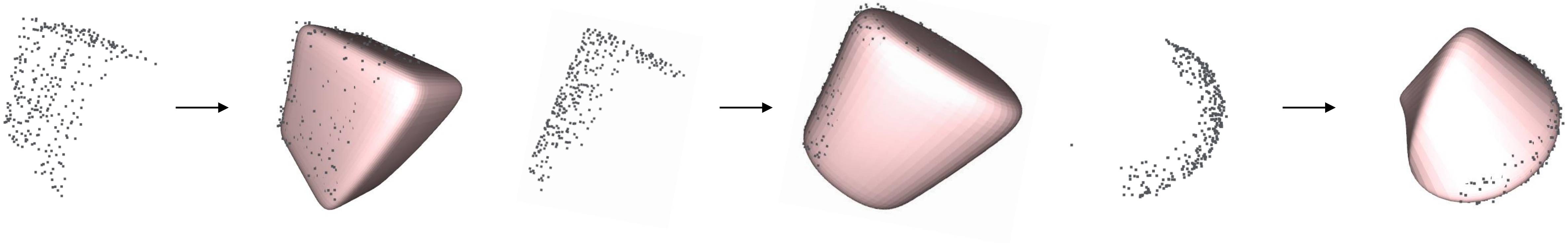
fit only the “partially observed” point clouds
i.e., do not fit the occluded parts of the objects

Deformable Superquadric Network

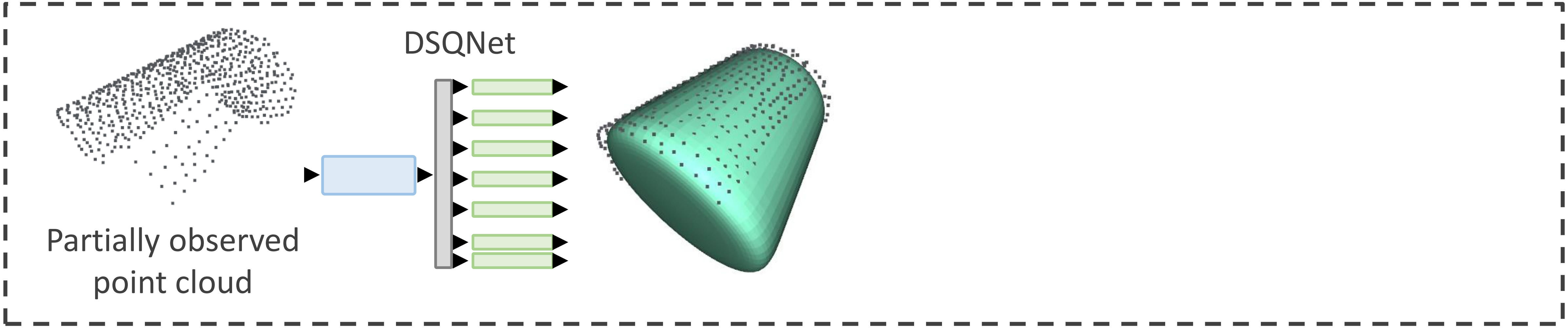


fit only the "partially observed" point clouds
i.e., do not fit the occluded parts of the objects

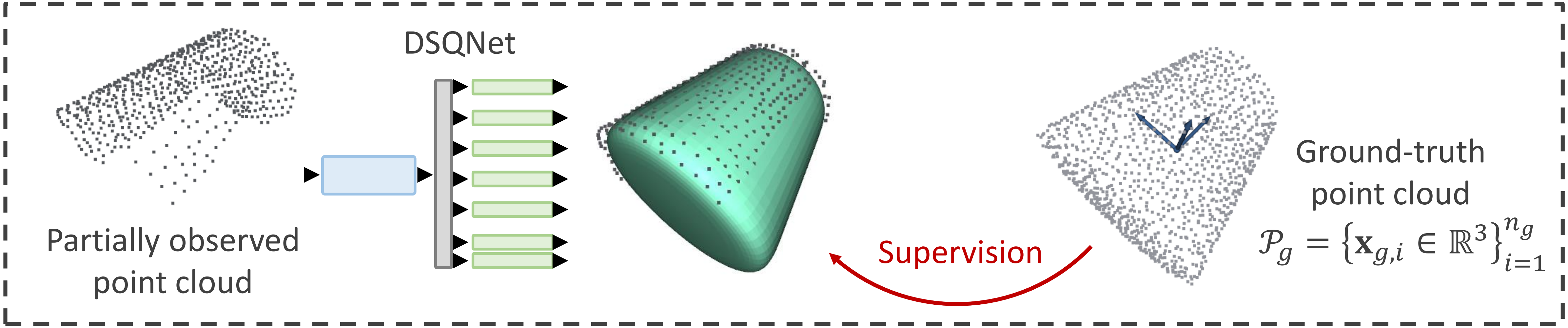
Examples of **optimization-based method**'s results



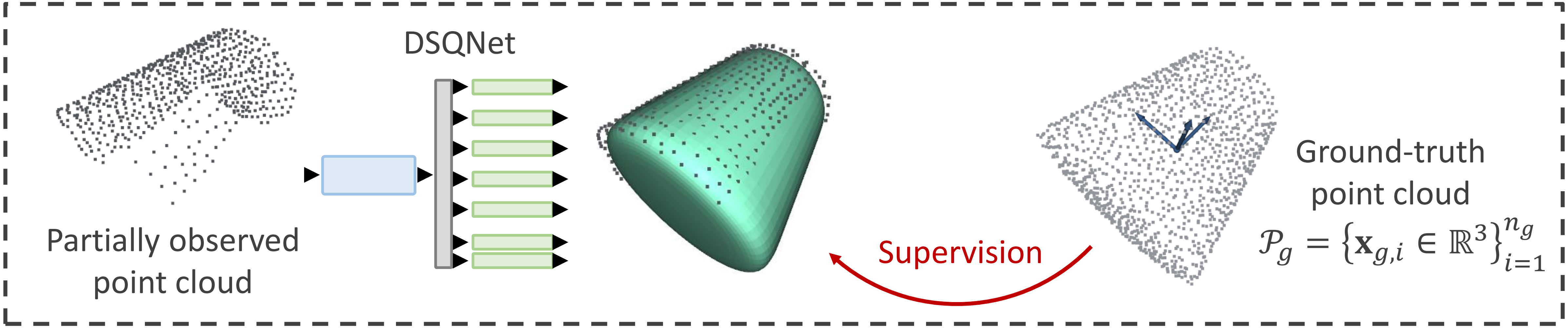
Deformable Superquadric Network



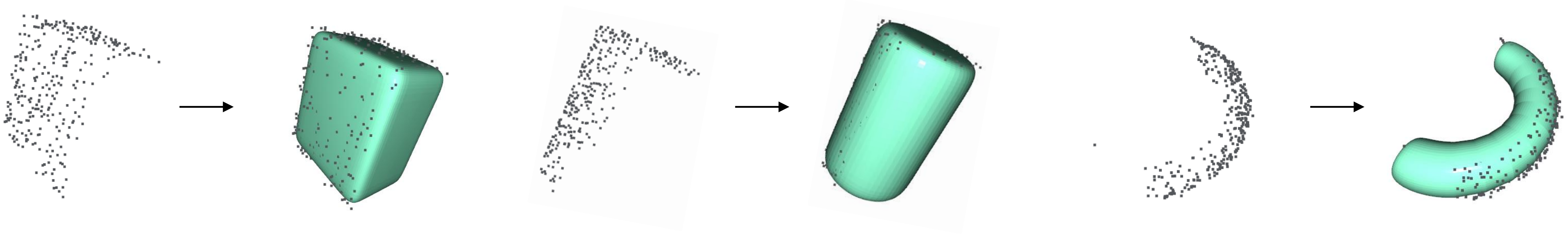
Deformable Superquadric Network



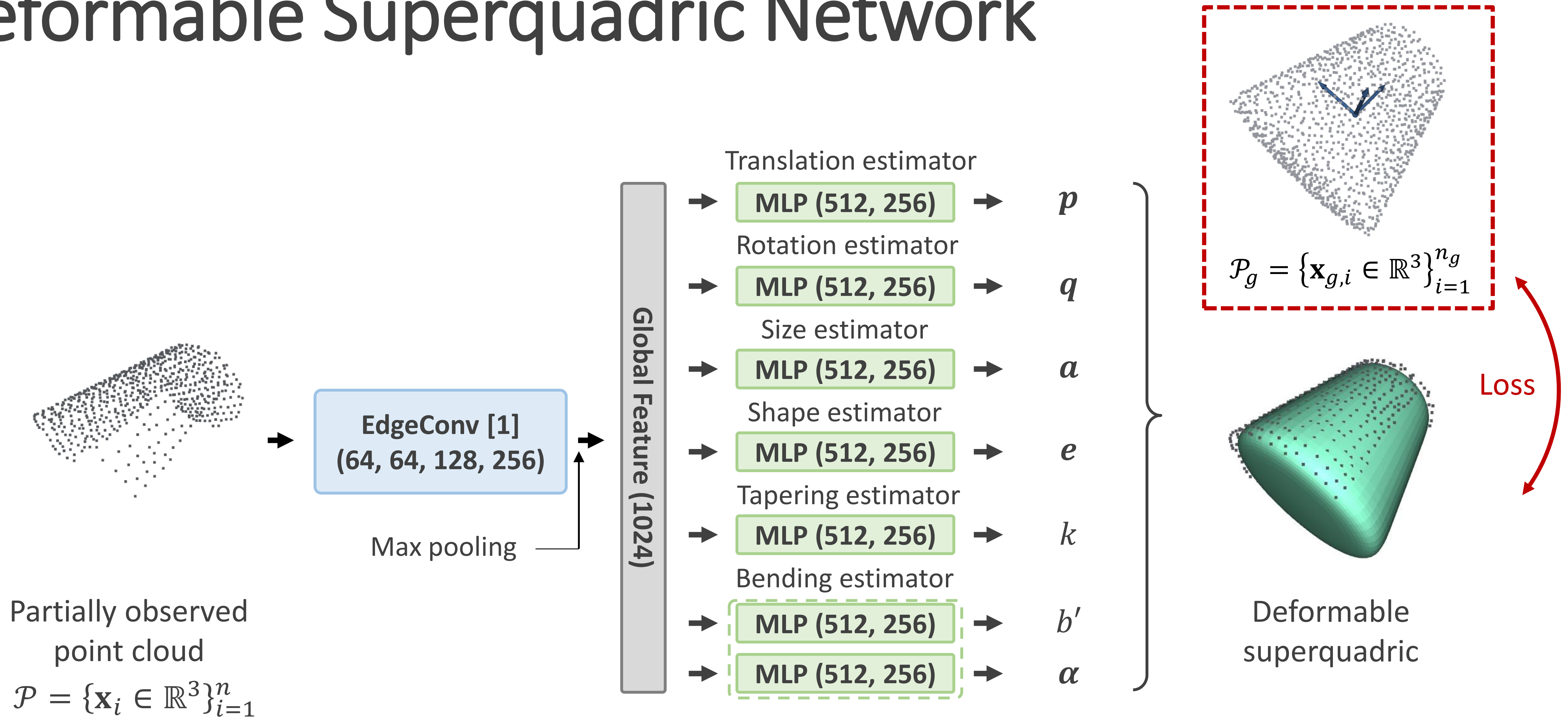
Deformable Superquadric Network



Examples of **DSQNet**'s results

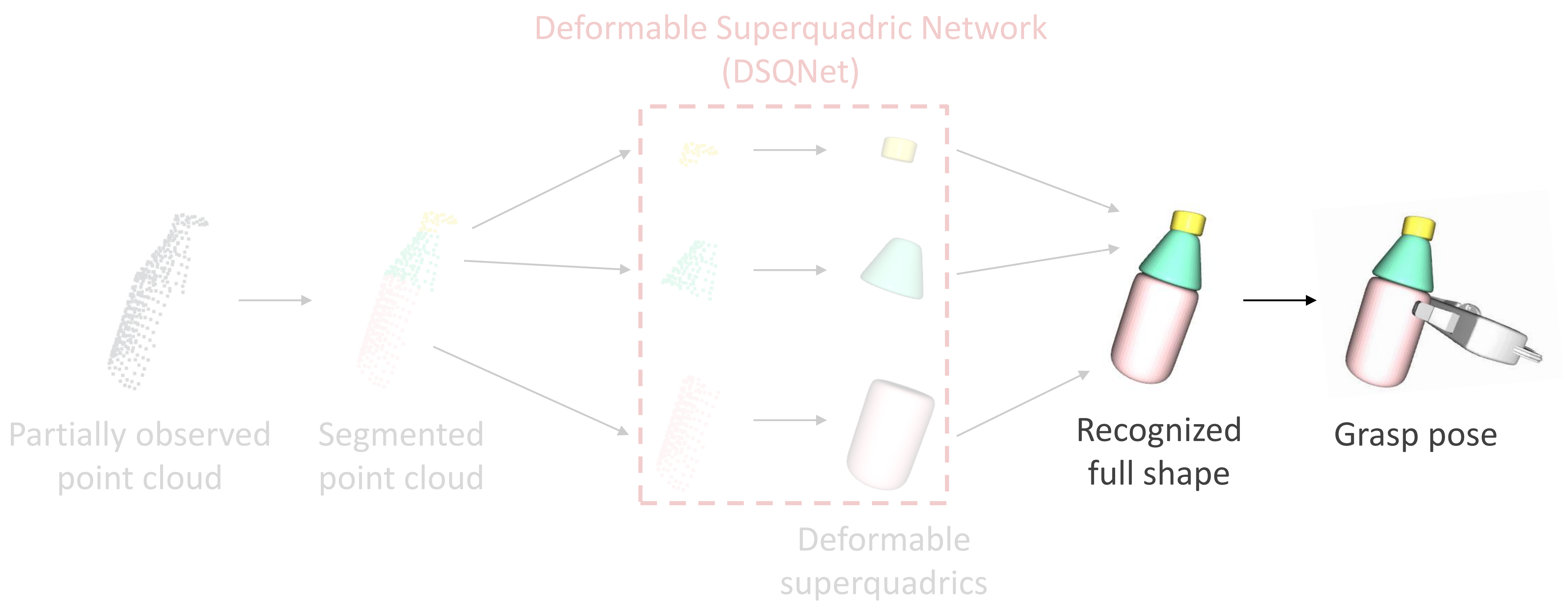


Deformable Superquadric Network

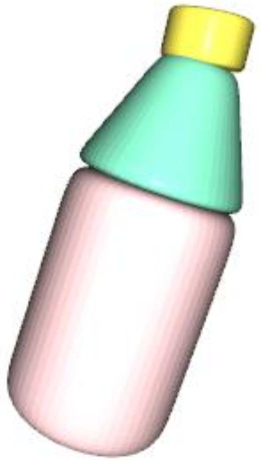


[1] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," Acm Transactions On Graphics (tog), vol. 38, no. 5, pp. 1–12, 2019.

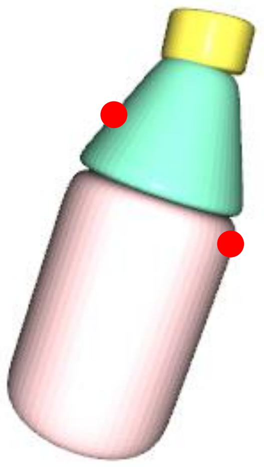
Our Method


Deformable Superquadric Network
(DSQNet)
Partially observed
point cloud
Segmented
point cloud
Deformable
superquadrics
Recognized
full shape
Grasp pose

Grasp Pose Generation

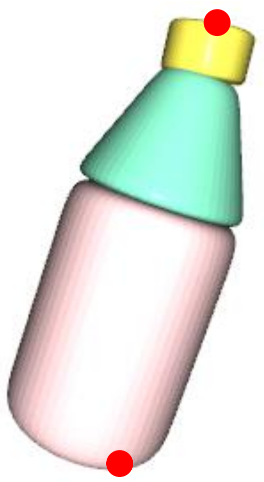


Grasp Pose Generation



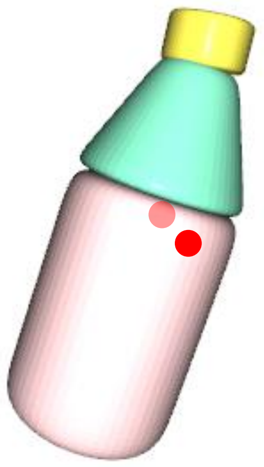
Sample antipodal points

Grasp Pose Generation



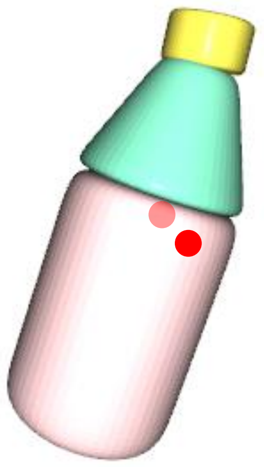
Sample antipodal points

Grasp Pose Generation

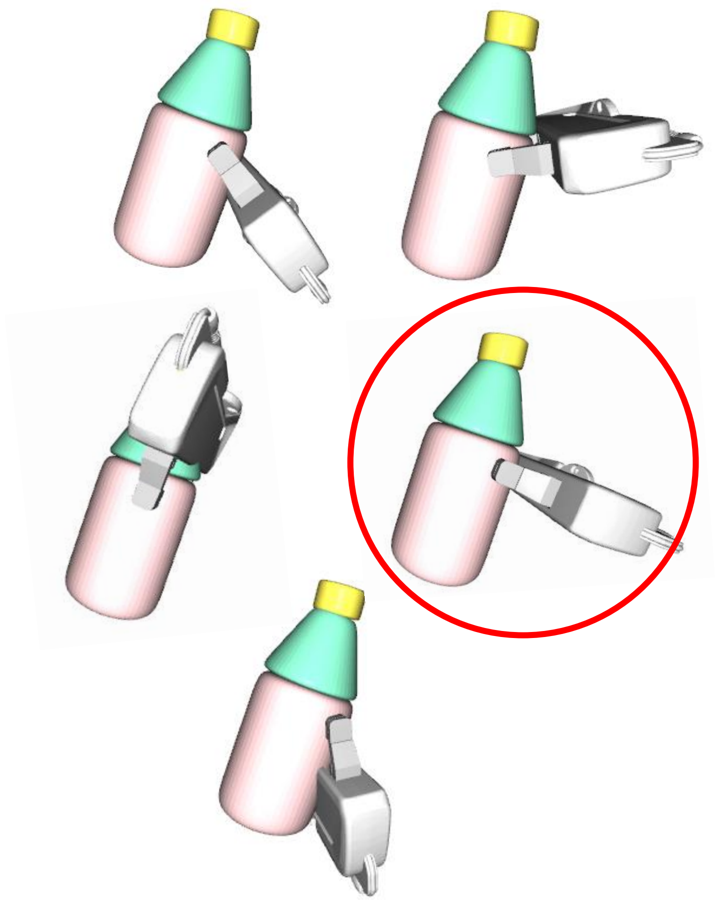


Sample antipodal points

Grasp Pose Generation

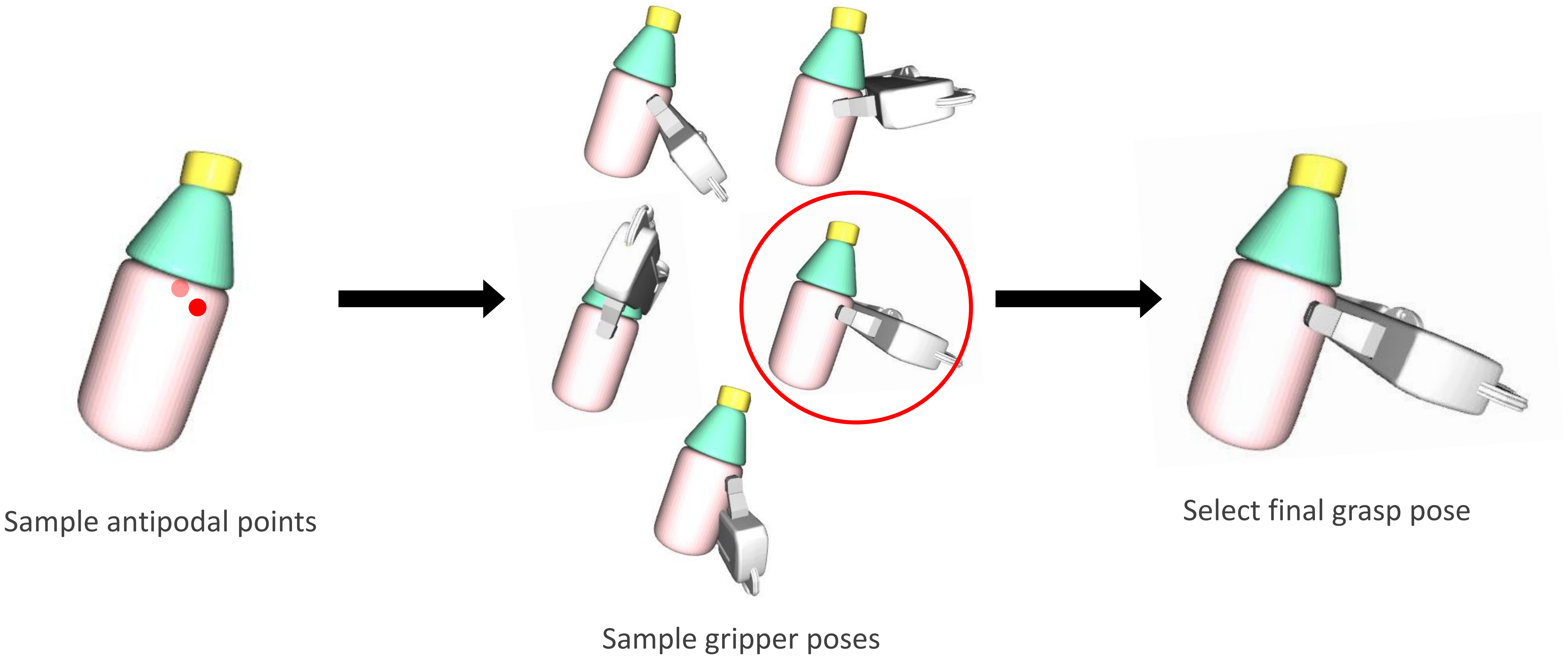


Sample antipodal points



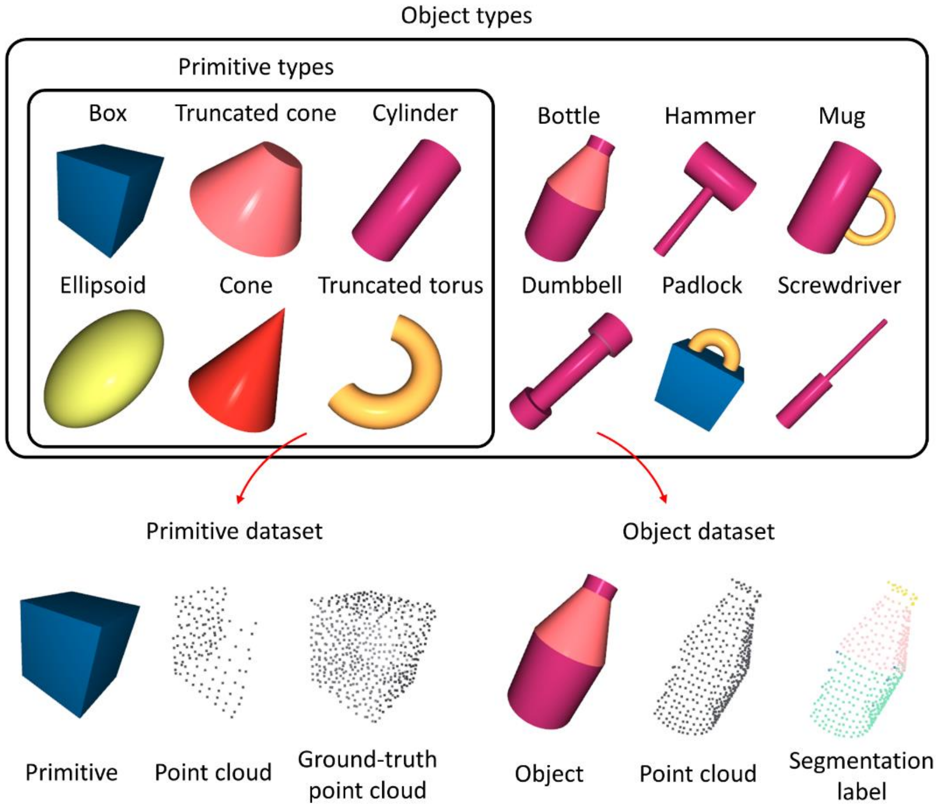
Sample gripper poses

Grasp Pose Generation



Experimental Results

Synthetic dataset
(1200 objects, about 10,000 pairs)



Real-world objects

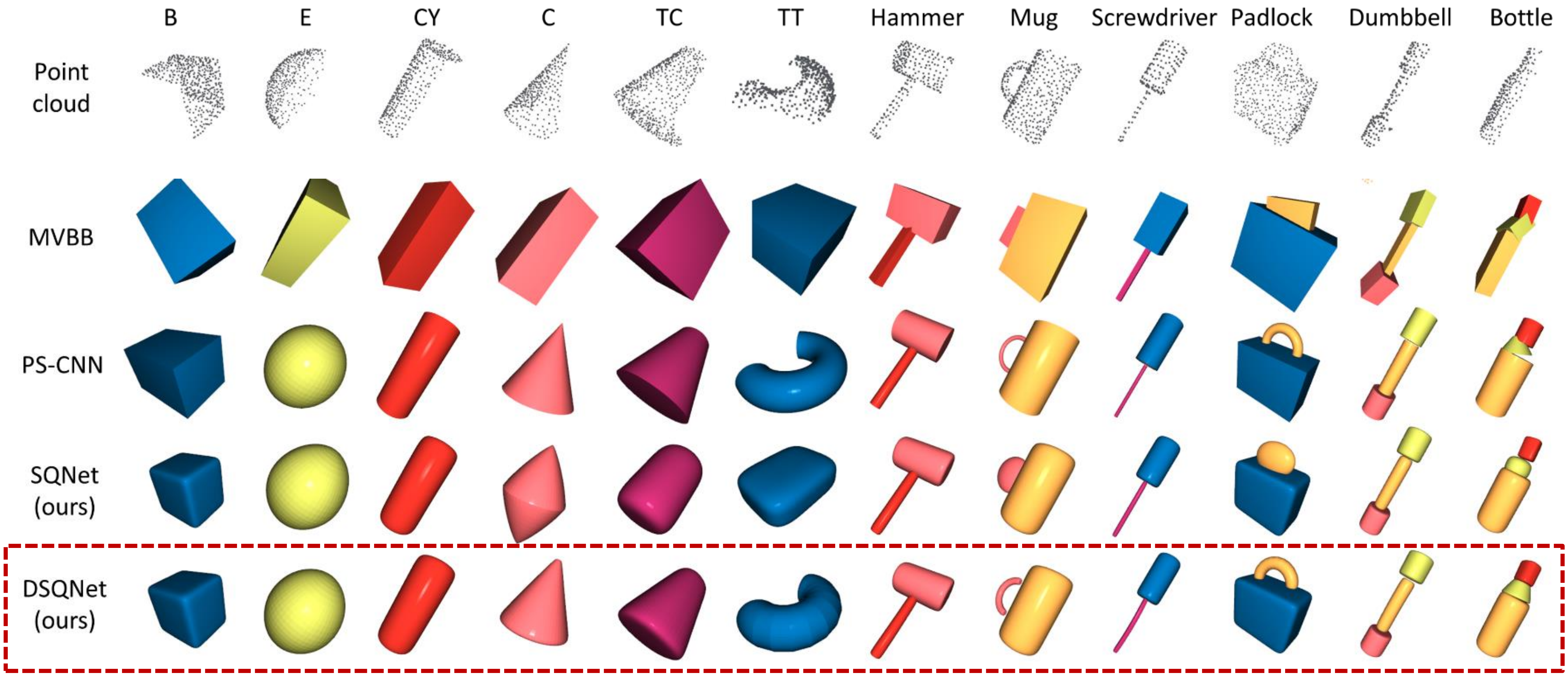
Experimental Results

TABLE III

VOLUMETRIC IOU COMPARISON BETWEEN MVBB, PS-CNN, SQNET, AND DSQNET FOR OBJECT DATASET

Objects	B	E	CY	C	TC	TT	Hammer	Cup	Screwdriver	Padlock	Dumbbell	Bottle	**Average**
MVBB	.3795	.3026	.5283	.3065	.4448	.3546	.5293	.4666	.5535	.4343	.4367	.4045	.4284
PS-CNN	.6442	.7429	.7988	.5946	.7504	.6141	.8101	.8282	.8346	.6751	**.7976**	.7610	.7376
SQNet (ours)	.8517	.8483	.8903	.5421	.7340	.3691	**.8358**	.7786	.8631	.8182	.7589	.8120	.7588
DSQNet (ours)	**.8759**	**.8666**	**.8939**	**.8039**	**.8264**	**.6759**	.8208	**.8483**	**.8655**	**.8312**	.7017	**.8189**	**.8191**

Experimental Results



Robot Grasping

Grasping single-part shapes

- Recognize the object using single deformable superquadric.

Grasping multi-part shapes

- Recognize the object using multiple deformable superquadrics.

Grasping Dinnerware Objects


Bowl


Dish

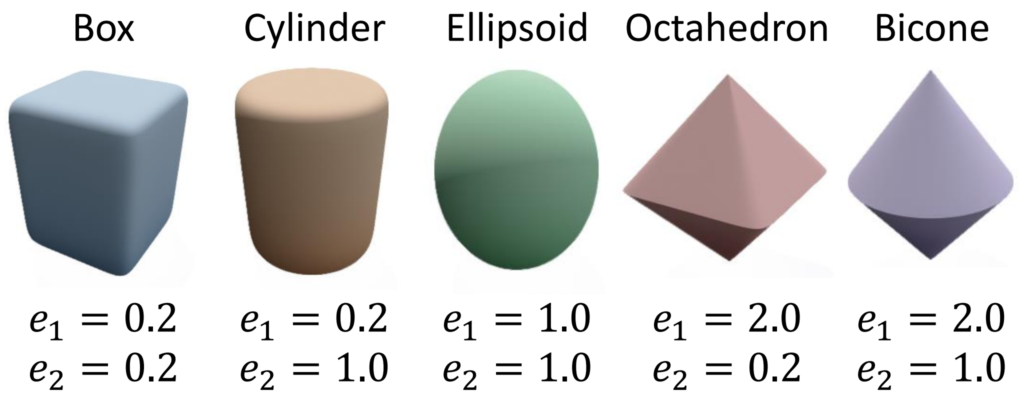

Mug


Spoon

Superparaboloids

Superquadrics

$$f(x,y,z) = \left(\left| \frac{x}{a_1} \right|^{2/e_2} + \left| \frac{y}{a_2} \right|^{2/e_2} \right)^{e_2/e_1} + \left| \frac{z}{a_3} \right|^{2/e_1} = 1$$



Superparaboloids

Superquadrics

$$f(x,y,z) = \left(\left| \frac{x}{a_1} \right|^{2/e_2} + \left| \frac{y}{a_2} \right|^{2/e_2} \right)^{e_2/e_1} + \left| \frac{z}{a_3} \right|^{2/e_1} = 1$$

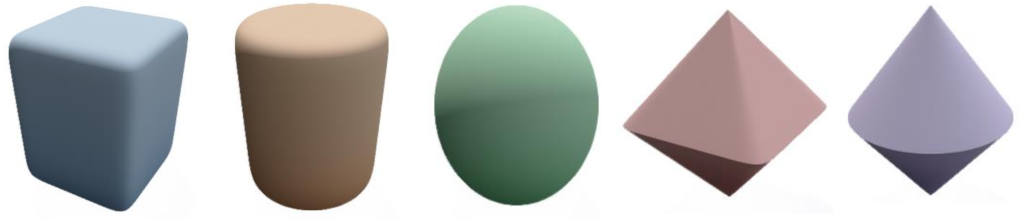
Superparaboloid

$$f(x,y,z) = \left(\left| \frac{x}{a_1} \right|^{2/e_2} + \left| \frac{y}{a_2} \right|^{2/e_2} \right)^{e_2/e_1} - \left(\frac{z}{a_3} \right) = 1$$

$\mathbf{a} = (a_1, a_2, a_3)$: size parameters

$\mathbf{e} = (e_1, e_2)$: shape parameters

Box	Cylinder	Ellipsoid	Octahedron	Bicone



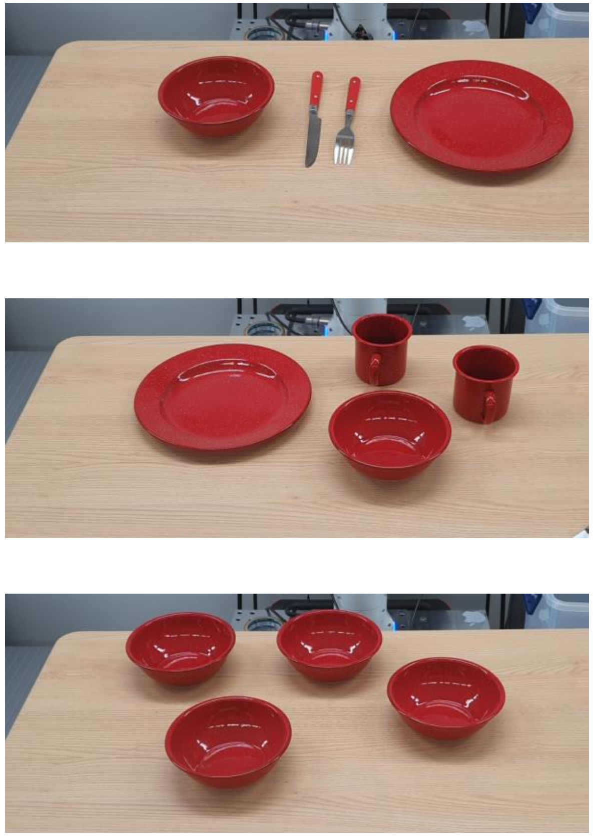
$e_1 = 0.2$	$e_1 = 0.2$	$e_1 = 1.0$	$e_1 = 2.0$	$e_1 = 2.0$
$e_2 = 0.2$	$e_2 = 1.0$	$e_2 = 1.0$	$e_2 = 0.2$	$e_2 = 1.0$

Dishes and bowls

$e_1 = 0.2$	$e_1 = 0.2$	$e_1 = 1.0$	$e_1 = 1.0$	$e_1 = 0.5$
$e_2 = 0.2$	$e_2 = 1.0$	$e_2 = 1.0$	$e_2 = 0.2$	$e_2 = 1.0$

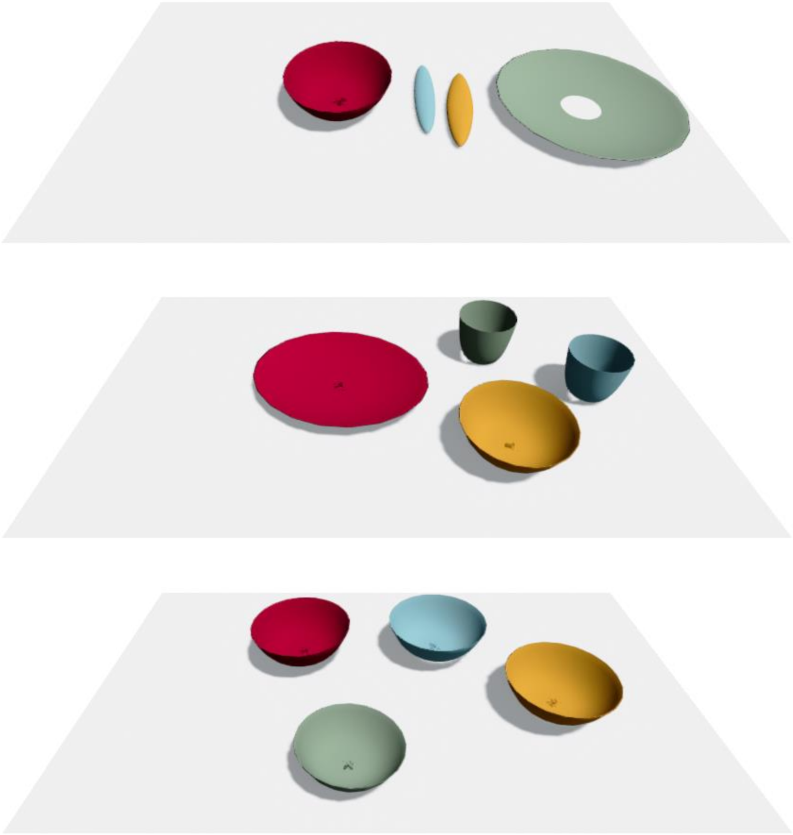
Experimental Results

Scene

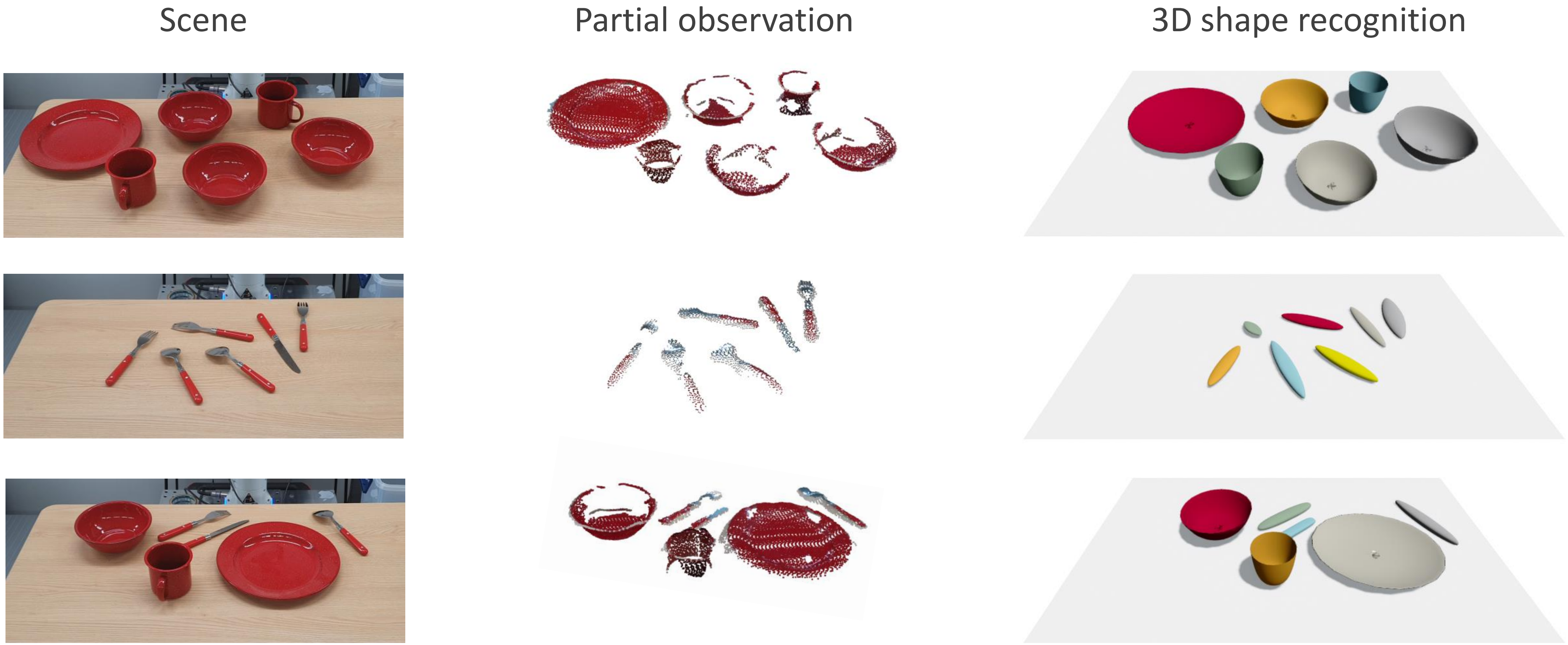


Partial observation

3D shape recognition



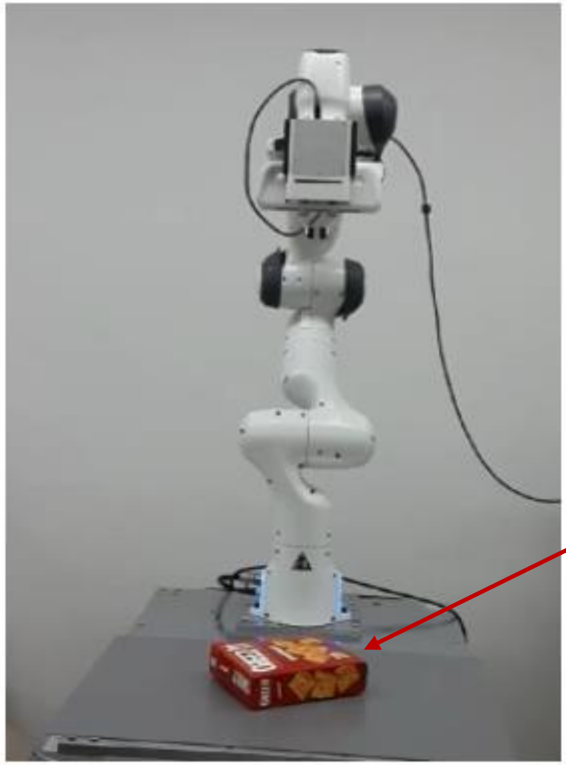
Experimental Results



Robot Grasping

- Success rate = 92% (92/100)

Needs for Non-prehensile Manipulation



Target object

- Too large to grasp

Target object

- Too cluttered environment

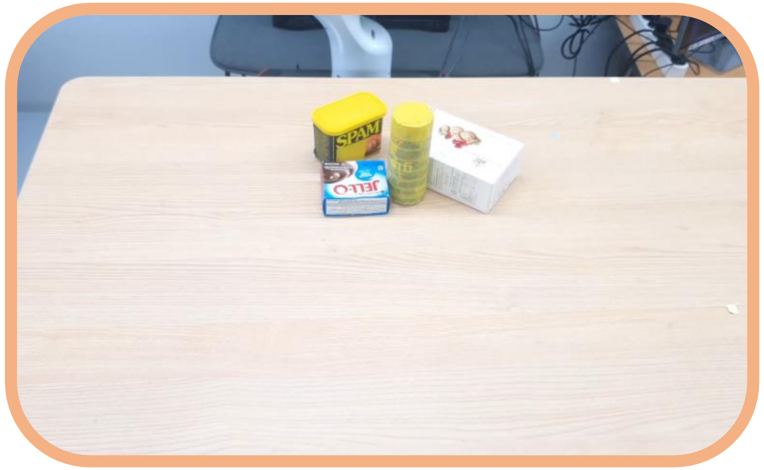
Needs for Non-prehensile Manipulation

- Move the target object

- Singulate the target object

Shape Recognition-based Approaches

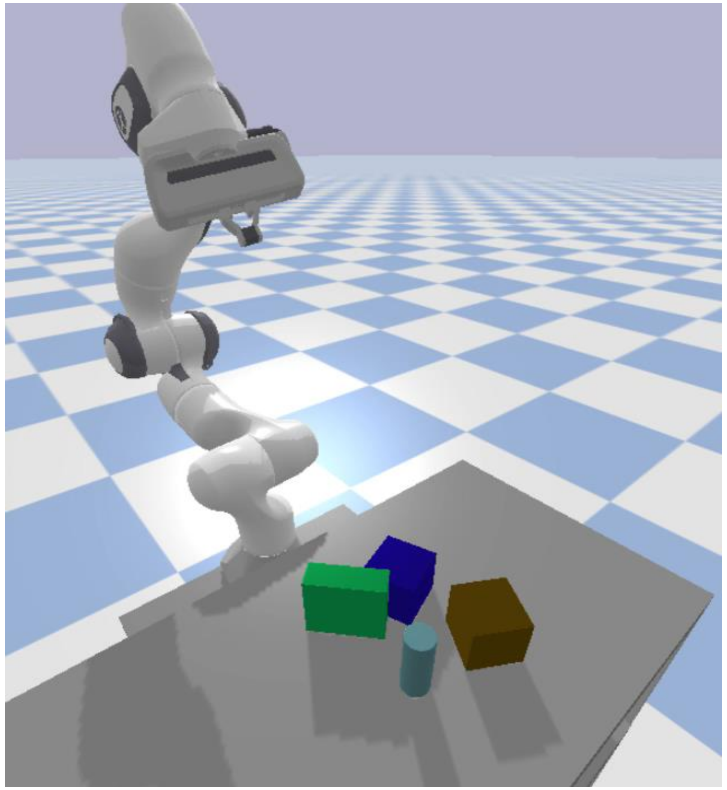
DSQNet
(S. Kim, et al., T-ASE'22)



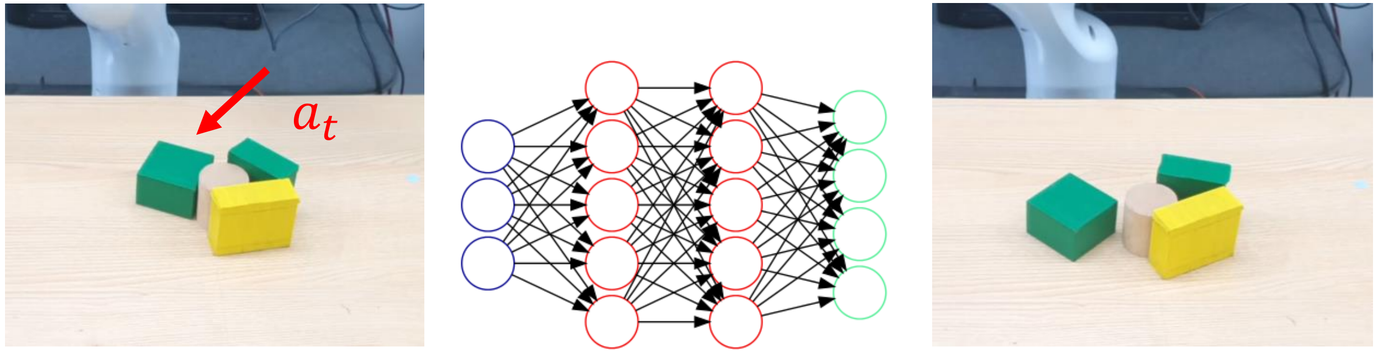
SQPDNet
(S. Kim, et al., CoRL'22)

Search-for-Grasp
(S. Kim, et al. CoRL'23)

Vision-based Pushing Manipulation



Data-driven end-to-end methods

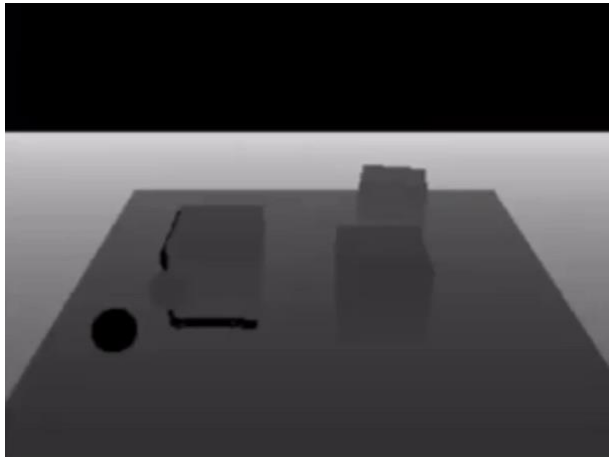


Scene s_t

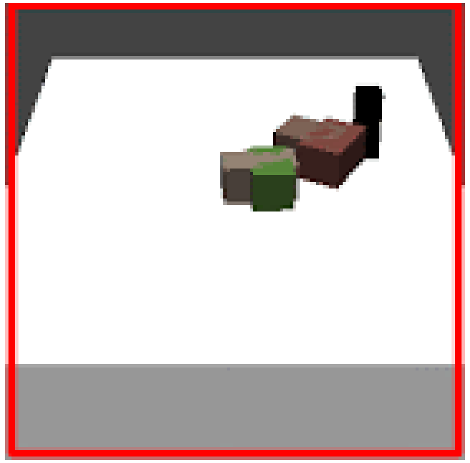
Next scene s_{t+1}

$$s_{t+1} = f(s_t, a_t)$$

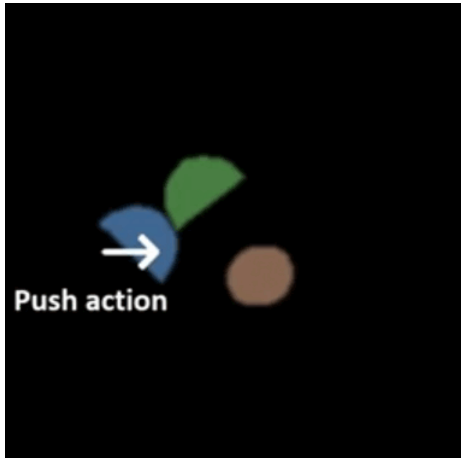
Data-Driven Methods



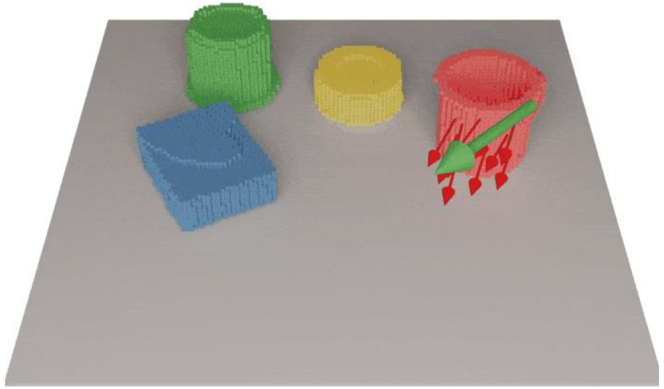
SE3-Net
(A. Byravan, et al., ICRA'17)



OC-MPC
(Y. Ye, et al., CoRL'19)

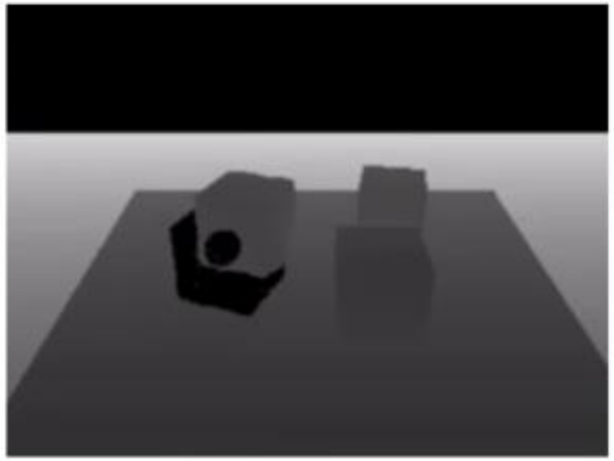


DIPN
(J. Wang, et al., ICRA'21)



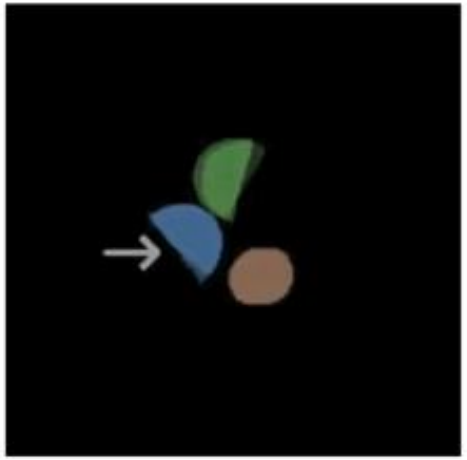
DSR-Net
(Z. Xu, et al., CoRL'20)

Data-Driven Methods

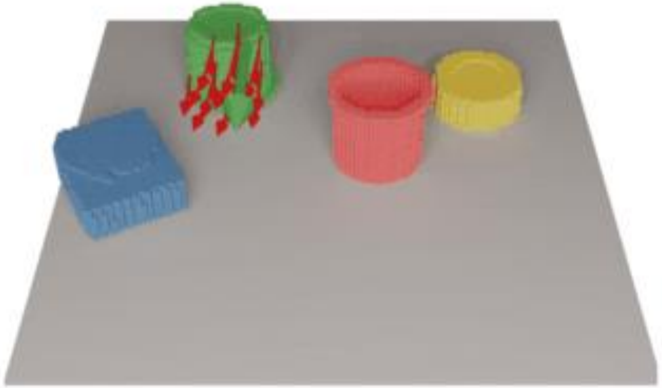


SE3-Net
(A. Byravan, et al., ICRA'17)

OC-MPC
(Y. Ye, et al., CoRL'19)



DIPN
(J. Wang, et al., ICRA'21)

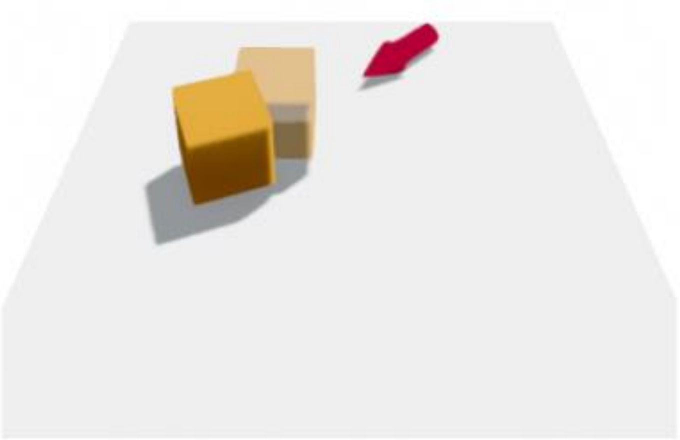


DSR-Net
(Z. Xu, et al., CoRL'20)

- Generalization performance is less-than-satisfying.
- Require large amounts of training data.

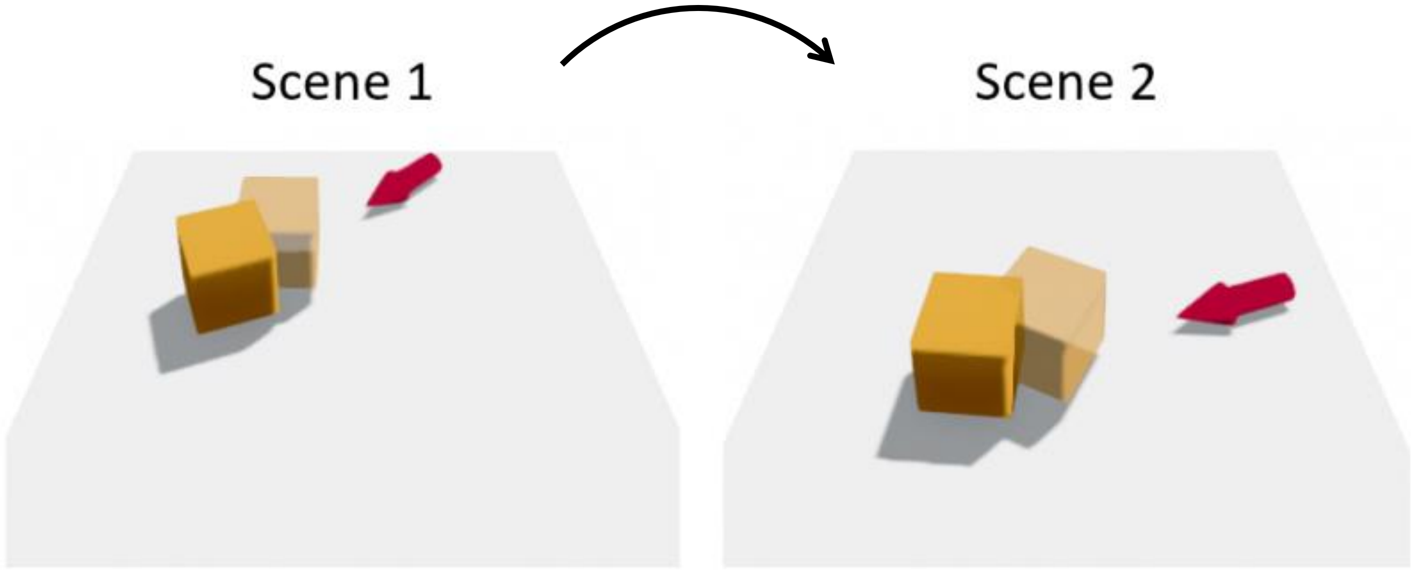
Reducing Needed Training Data: Equivariance

Scene 1



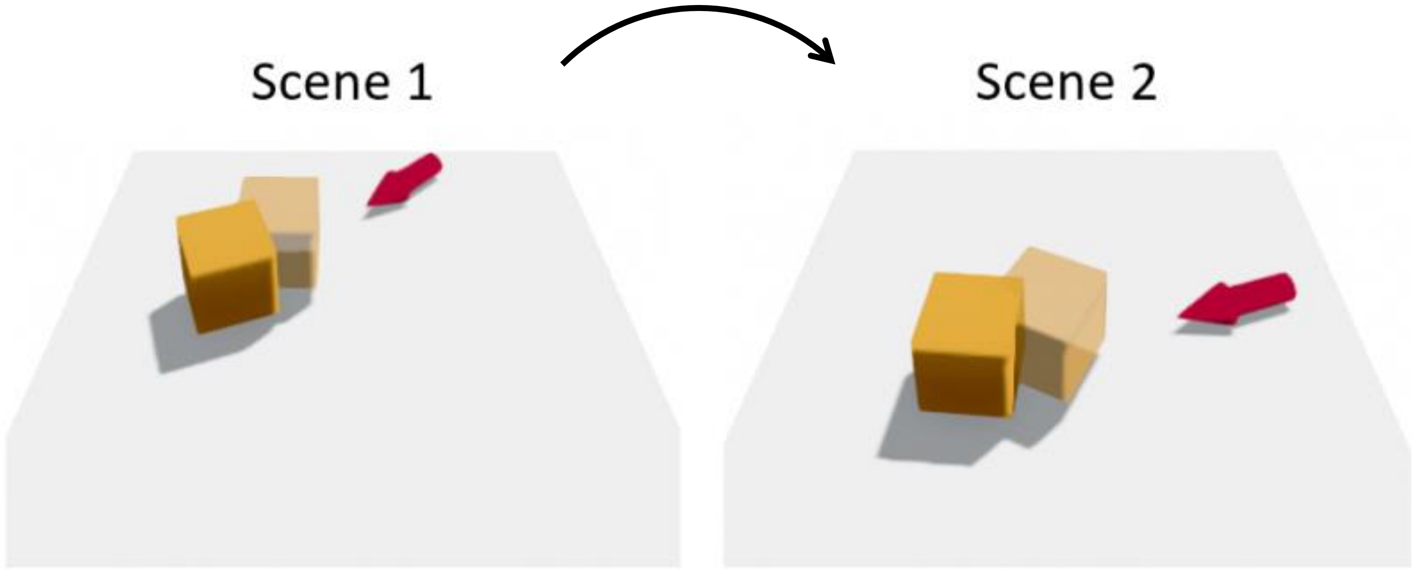
Reducing Needed Training Data: Equivariance

Objects and actions are translated and rotated.



Reducing Needed Training Data: Equivariance

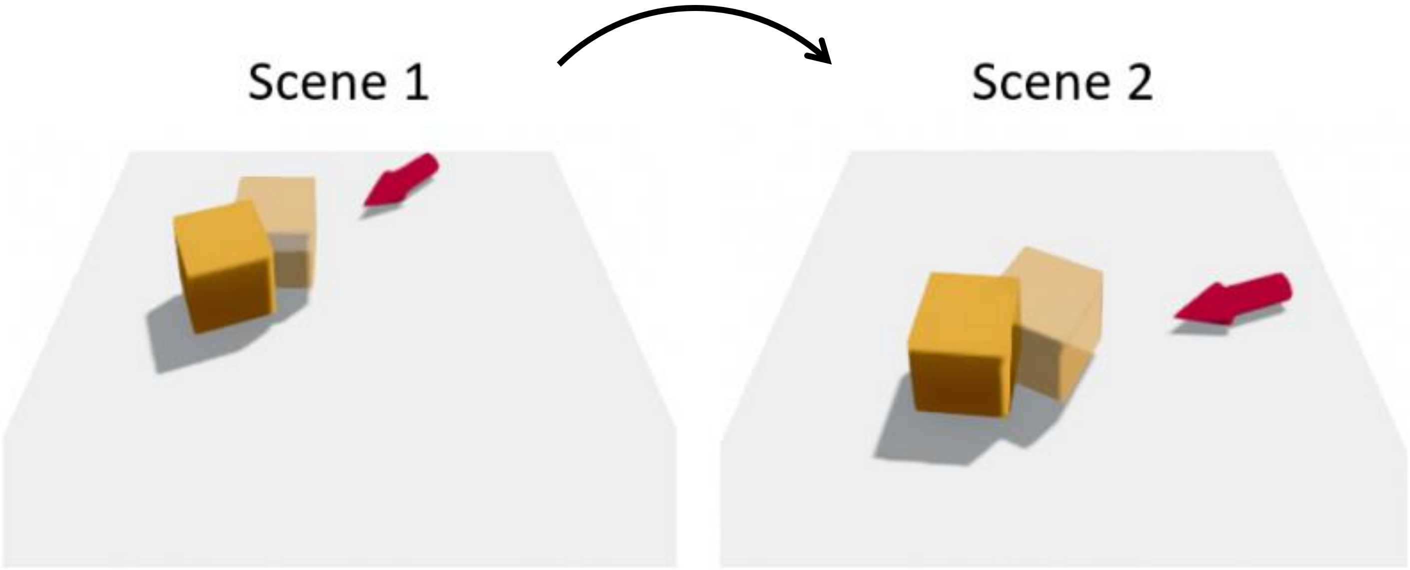
Objects and actions are translated and rotated.



A network that possesses this property is said to be equivariant with respect to translations and rotations.

Reducing Needed Training Data: Equivariance

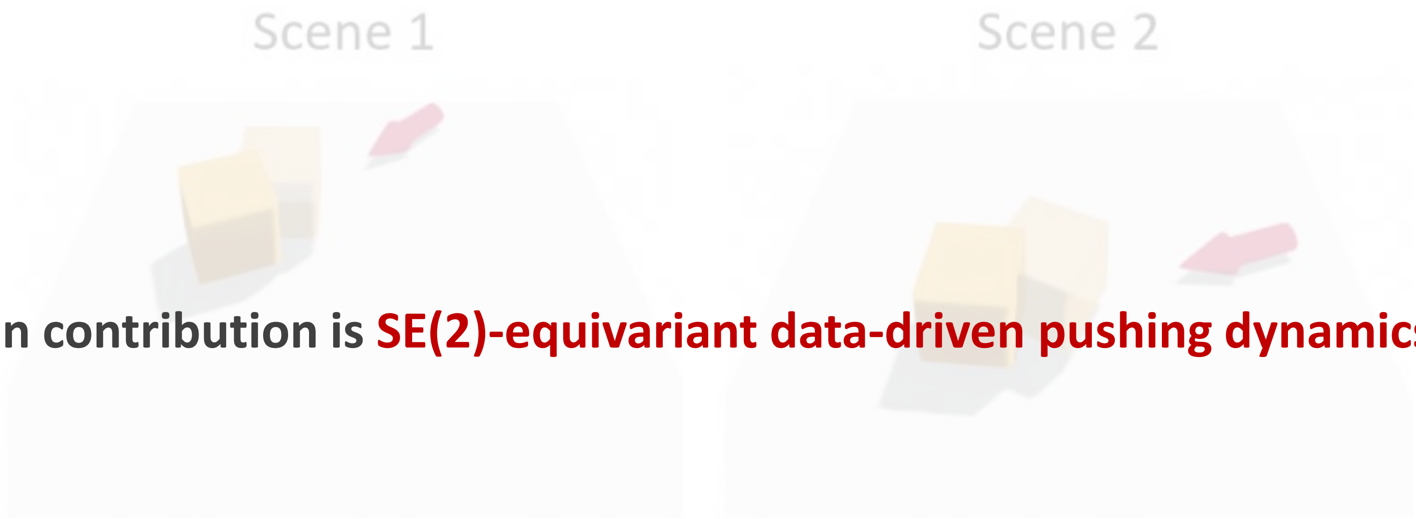
Objects and actions are translated and rotated. $\begin{bmatrix} \mathbf{Rot}(\hat{\mathbf{z}}, \theta) & \mathbf{t_{xy}} \\ 0 & 1 \end{bmatrix} \in \mathrm{SE}(2)$



A network that possesses this property is said to be equivariant with respect to translations and rotations.

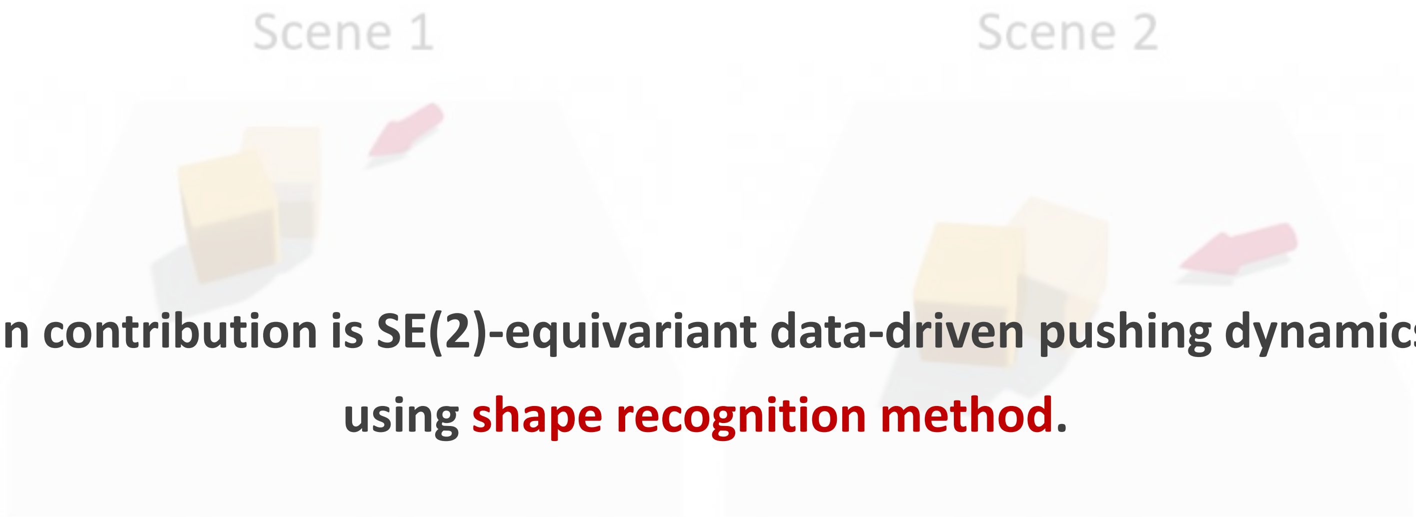
equivariant with respect to SE(2).

Reducing Needed Training Data: Equivariance



The main contribution is SE(2)-equivariant data-driven pushing dynamics model

Reducing Needed Training Data: Equivariance



The main contribution is SE(2)-equivariant data-driven pushing dynamics model using shape recognition method.

SE(2)-Equivariant Network Architecture

SE(2)-Equivariant Network Architecture

Assume that table surface is flat and orthogonal to the gravity with uniform friction coefficient.

SE(2)-Equivariant Network Architecture

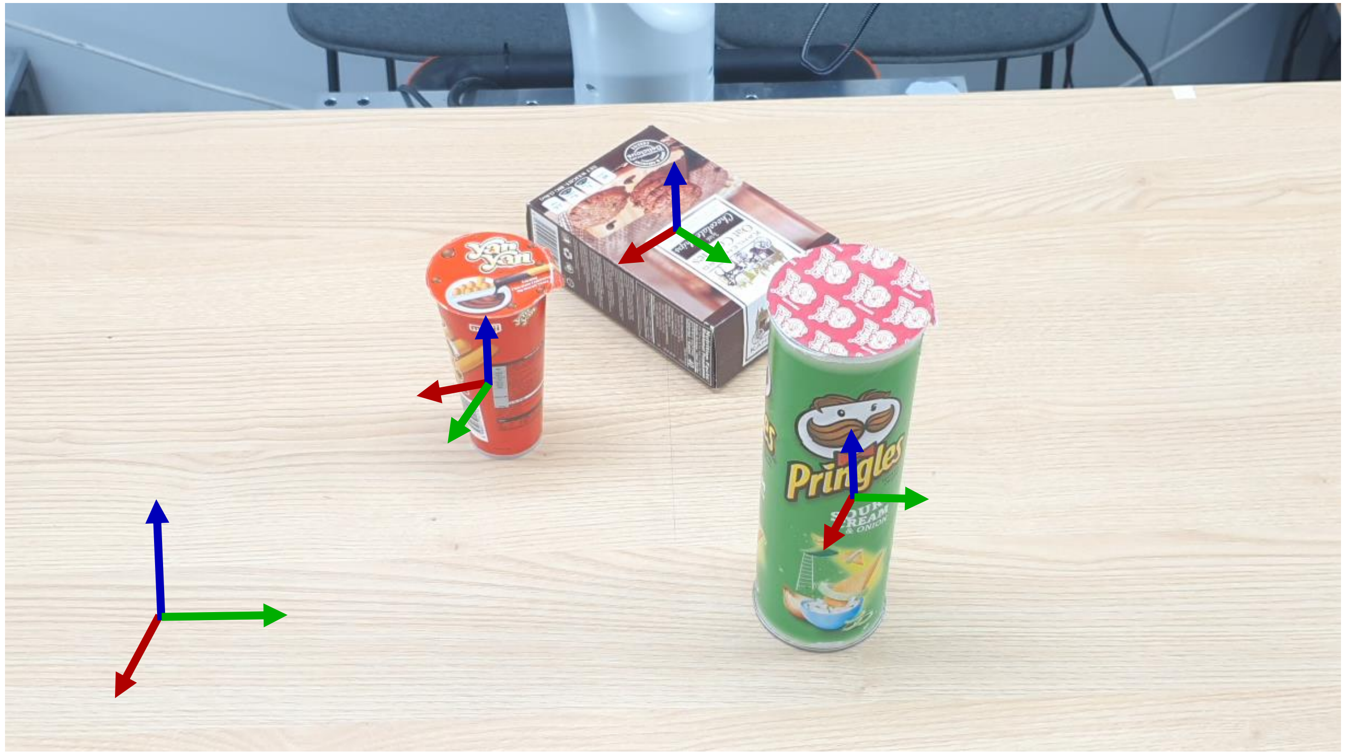
Assume that table surface is flat and orthogonal to the gravity with uniform friction coefficient.

Assume that we know the objects

$\mathbf{T}_i \in \mathrm{SE}(3)$ object pose

$\mathbf{q}_i \in \mathcal{Q}$ shape parameter

SE(2)-Equivariant Network Architecture



Assume that table surface is flat and orthogonal to the gravity with uniform friction coefficient.

Assume that we know the objects

$\boxed{\mathbf{T}_i \in \mathrm{SE}(3) \text{ object pose}}$

$\mathbf{q}_i \in \mathcal{Q}$ shape parameter

SE(2)-Equivariant Network Architecture

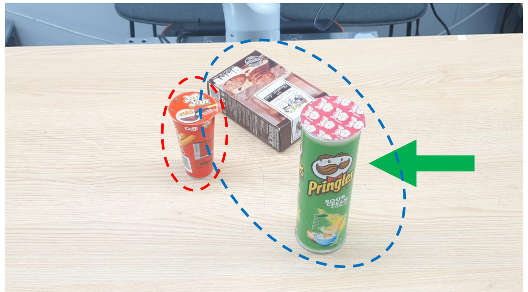
Assume that table surface is flat and orthogonal to the gravity with uniform friction coefficient.

Assume that we know the objects

$\mathbf{T}_i \in \mathrm{SE}(3)$ object pose

$\boxed{\mathbf{q}_i \in \mathcal{Q} \quad \text{shape parameter}}$

SE(2)-Equivariant Network Architecture



Assume that we know the objects

$\mathbf{T}_i \in \mathrm{SE}(3)$ object pose

$\mathbf{q}_i \in \mathcal{Q}$ shape parameter

$$\mathbf{T}_1^{t+1} = f(\underbrace{\{\mathbf{T}_1^t, \mathbf{q}_1\}}_{\text{Object 1}}, \underbrace{\{\mathbf{T}_2^t, ..., \mathbf{T}_N^t, \mathbf{q}_2, ..., \mathbf{q}_N\}}_{\text{Surrounding objects}}, \underbrace{a^t}_{\text{Action}})$$

SE(2)-Equivariant Network Architecture

Assume that we know the objects

$\mathbf{T}_i \in \mathrm{SE}(3)$ object pose

$\mathbf{q}_i \in \mathcal{Q}$ shape parameter

$$\mathbf{T}_1^{t+1} = f(\{\mathbf{T}_1^t, \mathbf{q}_1\}, \{\mathbf{T}_2^t, \ldots, \mathbf{T}_N^t, \mathbf{q}_2, \ldots, \mathbf{q}_N\}, a^t)$$

$$\mathbf{T}_2^{t+1} = f(\{\mathbf{T}_2^t, \mathbf{q}_2\}, \{\mathbf{T}_1^t, \ldots, \mathbf{T}_N^t, \mathbf{q}_1, \ldots, \mathbf{q}_N\}, a^t)$$

$$\vdots$$

$$\mathbf{T}_N^{t+1} = f(\{\mathbf{T}_N^t, \mathbf{q}_N\}, \{\mathbf{T}_1^t, \ldots, \mathbf{T}_{N-1}^t, \mathbf{q}_1, \ldots, \mathbf{q}_{N-1}\}, a^t)$$

SE(2)-Equivariant Network Architecture

Assume that we know the objects

$\mathbf{T}_i \in \mathrm{SE}(3)$ object pose

$\mathbf{q}_i \in \mathcal{Q}$ shape parameter

Apply $\mathbf{C} = \begin{bmatrix} \mathbf{Rot}(\hat{\mathbf{z}}, \theta) & \mathbf{t_{xy}} \\ 0 & 1 \end{bmatrix} \in \mathrm{SE}(2)$

$$f(\{\mathbf{T}_1^t, \mathbf{q}_1\}, \{\mathbf{T}_2^t, ..., \mathbf{T}_N^t, \mathbf{q}_2, ..., \mathbf{q}_N\}, a^t)$$

$$f(\{\mathbf{T}_2^t, \mathbf{q}_2\}, \{\mathbf{T}_1^t, ..., \mathbf{T}_N^t, \mathbf{q}_1, ..., \mathbf{q}_N\}, a^t)$$

$$\vdots$$

$$f(\{\mathbf{T}_N^t, \mathbf{q}_N\}, \{\mathbf{T}_1^t, ..., \mathbf{T}_{N-1}^t, \mathbf{q}_1, ..., \mathbf{q}_{N-1}\}, a^t)$$

SE(2)-Equivariant Network Architecture

Assume that we know the objects

$\mathbf{T}_i \in \mathrm{SE}(3)$ object pose

$\mathbf{q}_i \in \mathcal{Q}$ shape parameter

Apply $\mathbf{C} = \begin{bmatrix} \mathbf{Rot}(\hat{\mathbf{z}}, \theta) & \mathbf{t_{xy}} \\ 0 & 1 \end{bmatrix} \in \mathrm{SE}(2)$

$$f(\{\mathbf{C}\mathbf{T}_1^t, \mathbf{q}_1\}, \{\mathbf{C}\mathbf{T}_2^t, ..., \mathbf{C}\mathbf{T}_N^t, \mathbf{q}_2, ..., \mathbf{q}_N\}, \mathbf{C}a^t)$$

$$f(\{\mathbf{C}\mathbf{T}_2^t, \mathbf{q}_2\}, \{\mathbf{C}\mathbf{T}_1^t, ..., \mathbf{C}\mathbf{T}_N^t, \mathbf{q}_1, ..., \mathbf{q}_N\}, \mathbf{C}a^t)$$

$$\vdots$$

$$f(\{\mathbf{C}\mathbf{T}_N^t, \mathbf{q}_N\}, \{\mathbf{C}\mathbf{T}_1^t, ..., \mathbf{C}\mathbf{T}_{N-1}^t, \mathbf{q}_1, ..., \mathbf{q}_{N-1}\}, \mathbf{C}a^t)$$

SE(2)-Equivariant Network Architecture

Assume that we know the objects

$\mathbf{T}_i \in \mathrm{SE}(3)$ object pose

$\mathbf{q}_i \in \mathcal{Q}$ shape parameter

Apply $\mathbf{C} = \begin{bmatrix} \mathbf{Rot}(\hat{\mathbf{z}}, \theta) & \mathbf{t_{xy}} \\ 0 & 1 \end{bmatrix} \in \mathrm{SE}(2)$

$$\mathbf{C}\mathbf{T}_1^{t+1} = f(\{\mathbf{C}\mathbf{T}_1^t, \mathbf{q}_1\}, \{\mathbf{C}\mathbf{T}_2^t, ..., \mathbf{C}\mathbf{T}_N^t, \mathbf{q}_2, ..., \mathbf{q}_N\}, \mathbf{C}a^t)$$

$$\mathbf{C}\mathbf{T}_2^{t+1} = f(\{\mathbf{C}\mathbf{T}_2^t, \mathbf{q}_2\}, \{\mathbf{C}\mathbf{T}_1^t, ..., \mathbf{C}\mathbf{T}_N^t, \mathbf{q}_1, ..., \mathbf{q}_N\}, \mathbf{C}a^t)$$

$$\vdots$$

$$\mathbf{C}\mathbf{T}_N^{t+1} = f(\{\mathbf{C}\mathbf{T}_N^t, \mathbf{q}_N\}, \{\mathbf{C}\mathbf{T}_1^t, ..., \mathbf{C}\mathbf{T}_{N-1}^t, \mathbf{q}_1, ..., \mathbf{q}_{N-1}\}, \mathbf{C}a^t)$$

SE(2)-Equivariant Network Architecture

Assume that we know the objects

$\mathbf{T}_i \in \mathrm{SE}(3)$ object pose

$\mathbf{q}_i \in \mathcal{Q}$ shape parameter

Definition 1 *A pushing dynamics model f is* $\mathrm{SE}(2)$*-equivariant if*

$$\mathbf{C}\mathbf{T}_1^{t+1} = f(\{\mathbf{C}\mathbf{T}_1^t, \mathbf{q}_1\}, \{\mathbf{C}\mathbf{T}_2^t, ..., \mathbf{C}\mathbf{T}_N^t, \mathbf{q}_2, ..., \mathbf{q}_N\}, \mathbf{C}a^t)$$

$$\mathbf{C}\mathbf{T}_2^{t+1} = f(\{\mathbf{C}\mathbf{T}_2^t, \mathbf{q}_2\}, \{\mathbf{C}\mathbf{T}_1^t, ..., \mathbf{C}\mathbf{T}_N^t, \mathbf{q}_1, ..., \mathbf{q}_N\}, \mathbf{C}a^t)$$

$$\vdots$$

$$\mathbf{C}\mathbf{T}_N^{t+1} = f(\{\mathbf{C}\mathbf{T}_N^t, \mathbf{q}_N\}, \{\mathbf{C}\mathbf{T}_1^t, ..., \mathbf{C}\mathbf{T}_{N-1}^t, \mathbf{q}_1, ..., \mathbf{q}_{N-1}\}, \mathbf{C}a^t)$$

for all $\mathbf{C} \in \mathrm{SE}(2)$

SE(2)-Equivariant Network Architecture

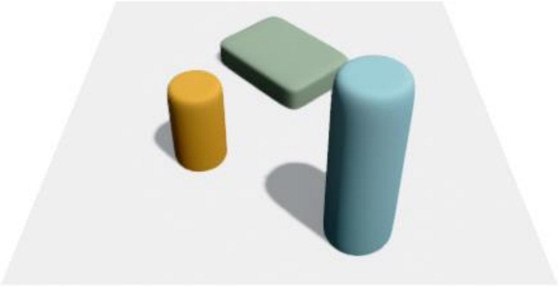
Visual observation

SE(2)-Equivariant Network Architecture

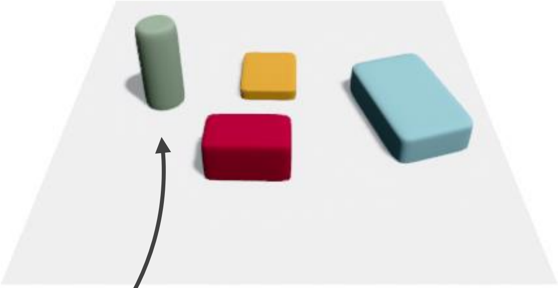
Visual observation

Recognized 3D objects

Shape Recognition



Shape Recognition



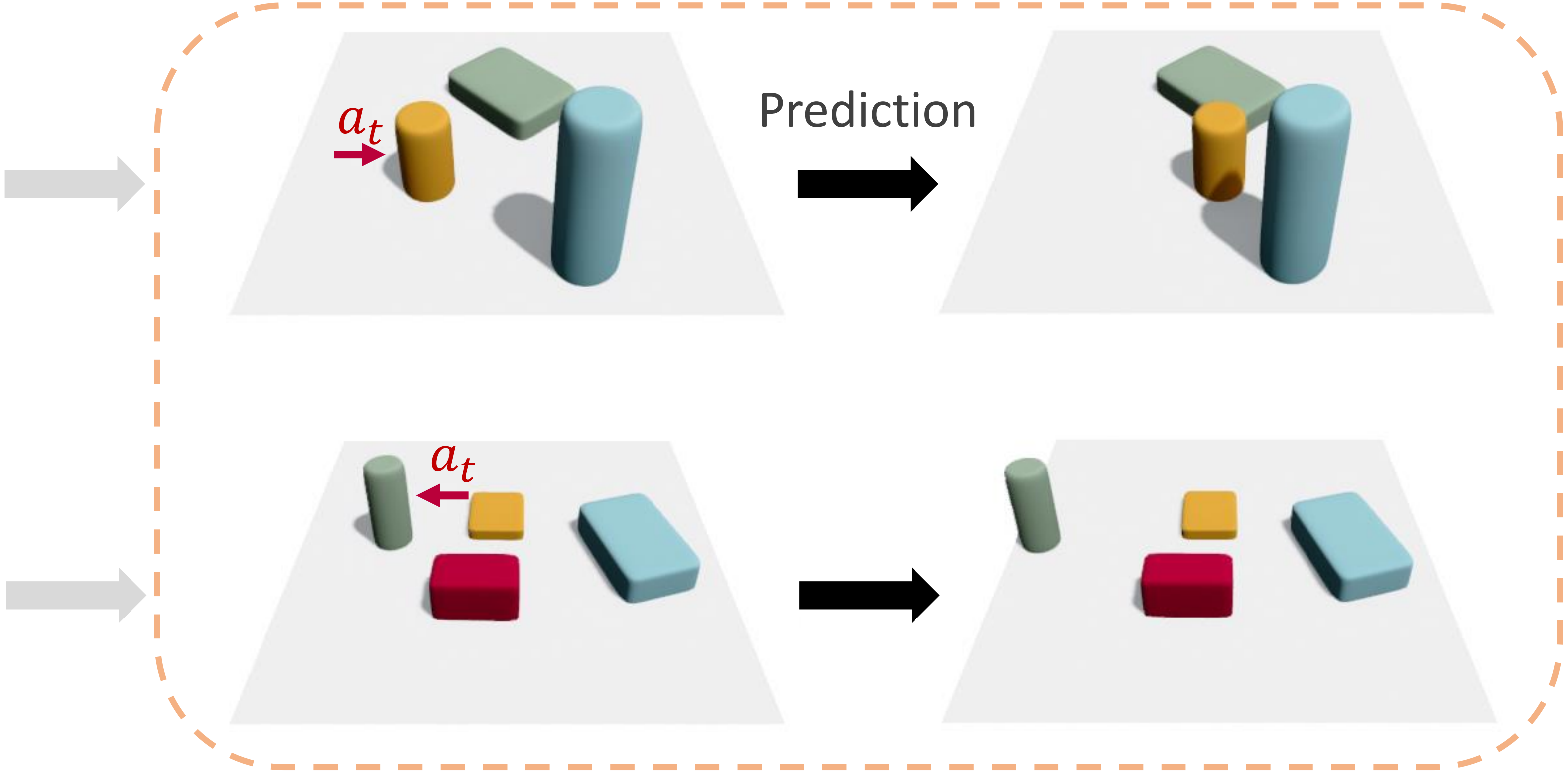
$\mathbf{T}_i \in \mathrm{SE}(3)$ Object poses

$\mathbf{q}_i$ Shape parameters

SE(2)-Equivariant Network Architecture

Visual observation

Recognized 3D objects



$\mathbf{T}_i \in \mathrm{SE}(3)$ Object poses

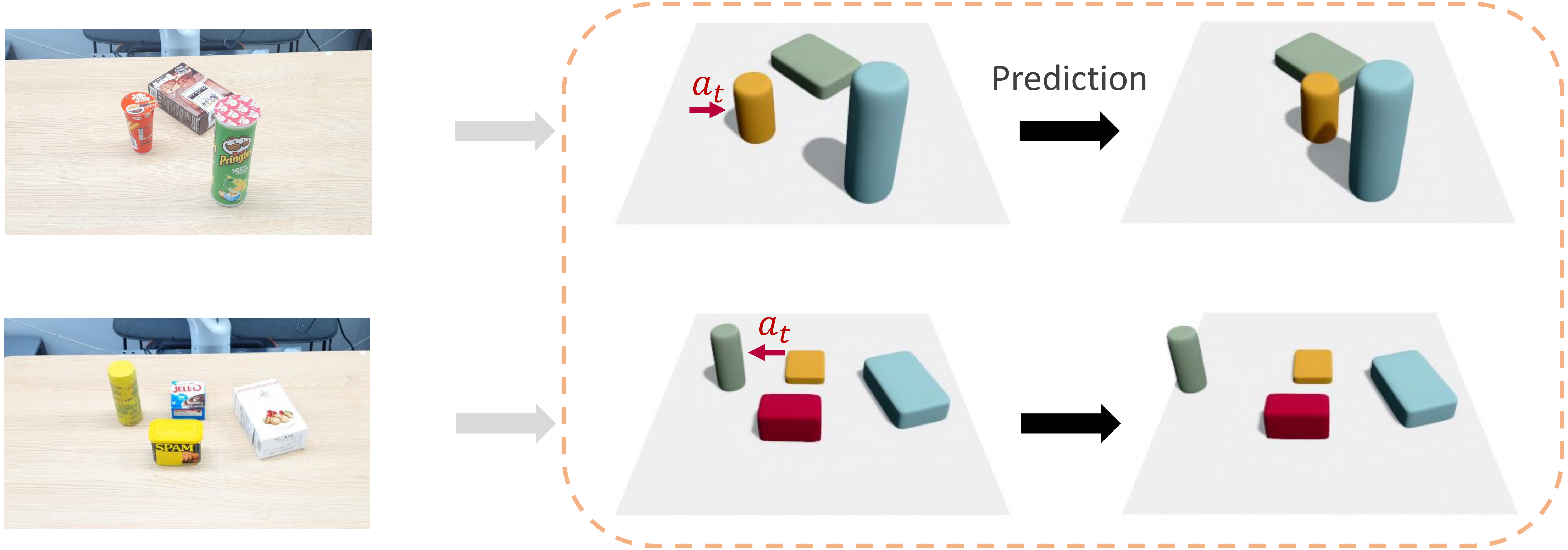
$\mathbf{q}_i$ Shape parameters

$$\{\mathbf{T}'_i\}_{i=1}^{N} = f(\{(\mathbf{T}_i, \mathbf{q}_i)\}_{i=1}^{N}, a_t)$$

SE(2)-Equivariant Network Architecture

Visual observation

Recognized 3D objects



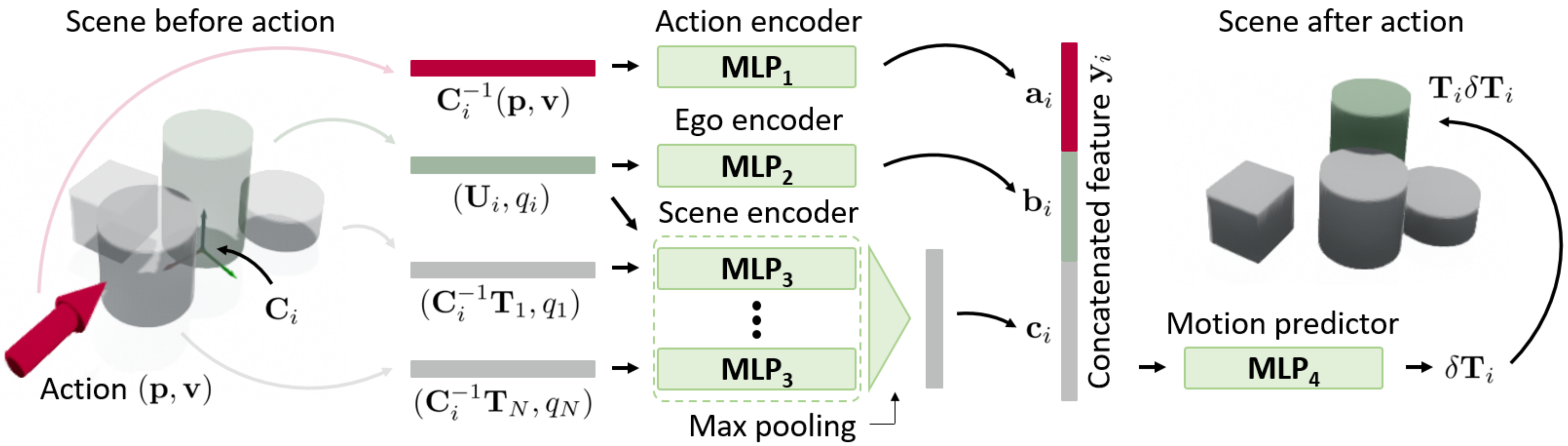
$\mathbf{T}_i \in \mathrm{SE}(3)$ Object poses

$\mathbf{q}_i$ Shape parameters

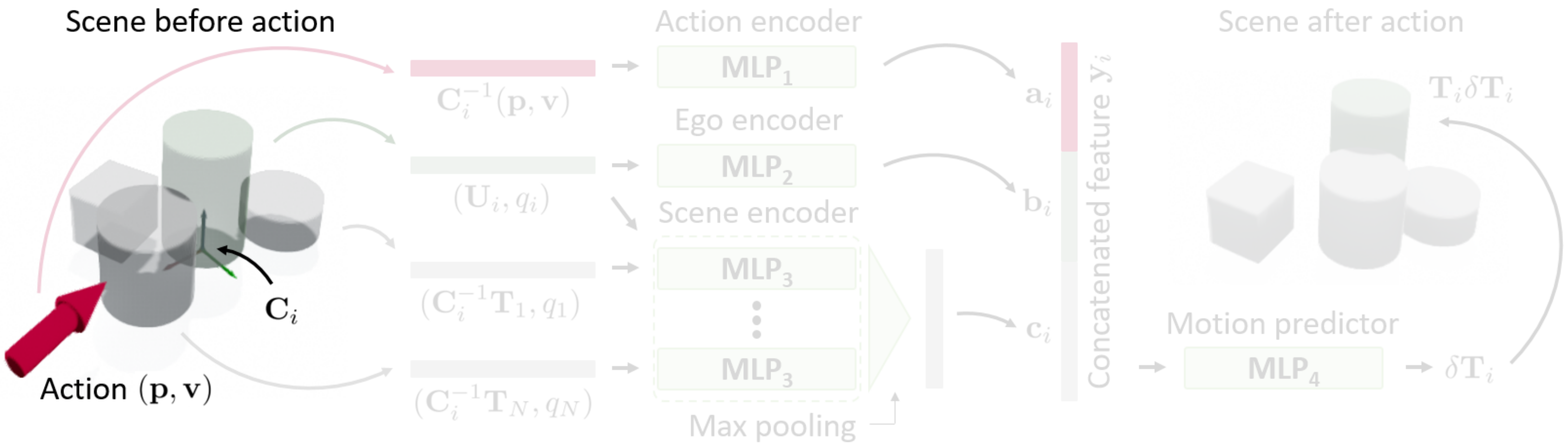
$$\{\mathbf{T}'_i\}_{i=1}^{N} = f(\{(\mathbf{T}_i, \mathbf{q}_i)\}_{i=1}^{N}, a_t)$$

Superquadric Pushing Dynamics Model (SQPDNet)

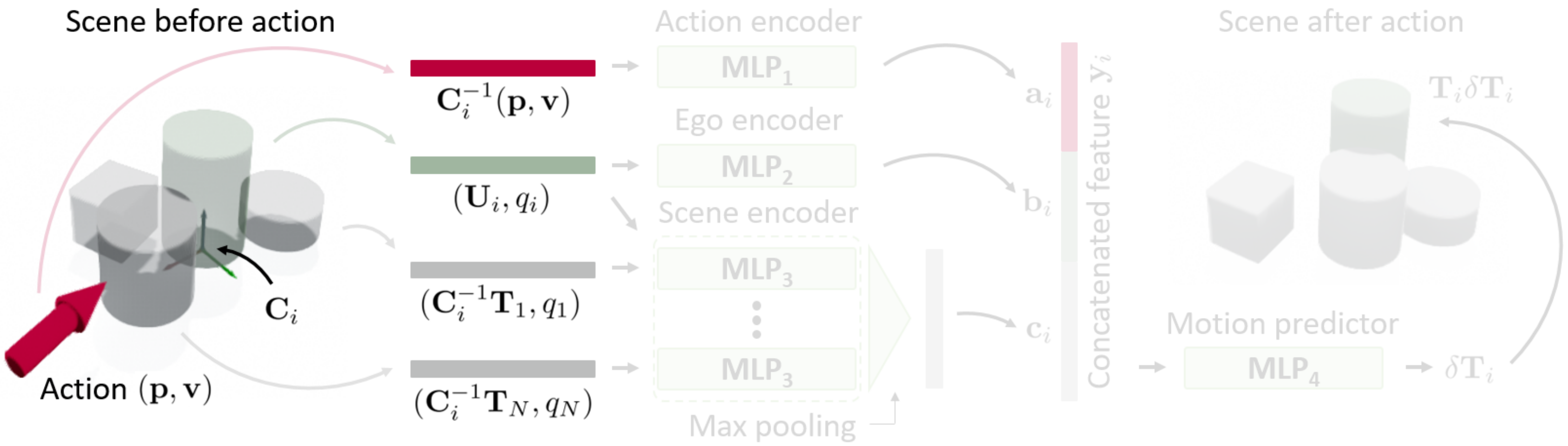
SE(2)-Equivariant Network Architecture



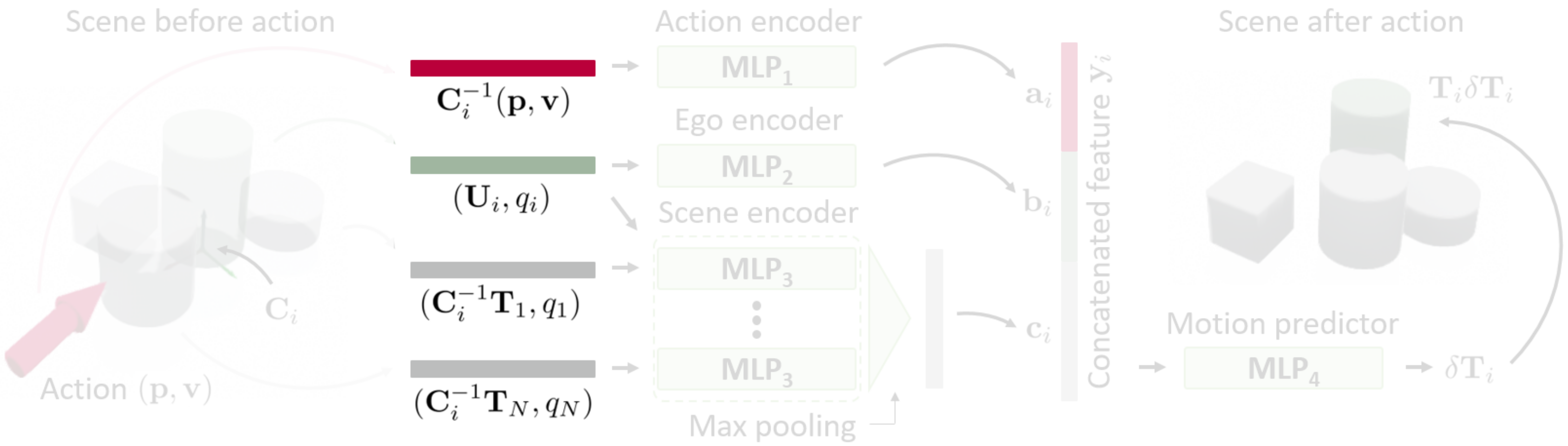
SE(2)-Equivariant Network Architecture



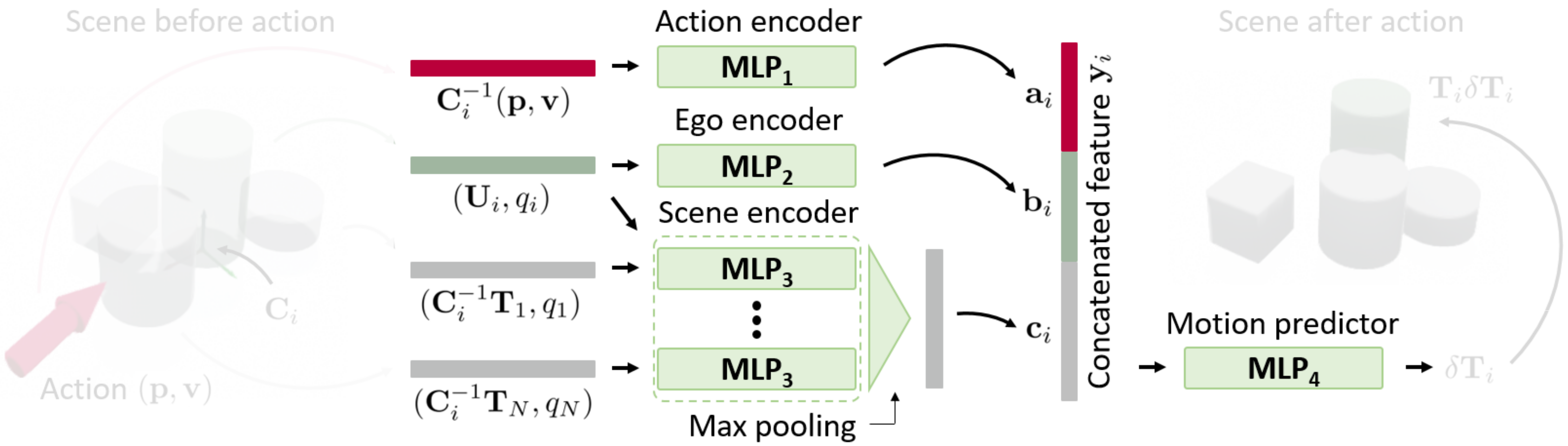
SE(2)-Equivariant Network Architecture



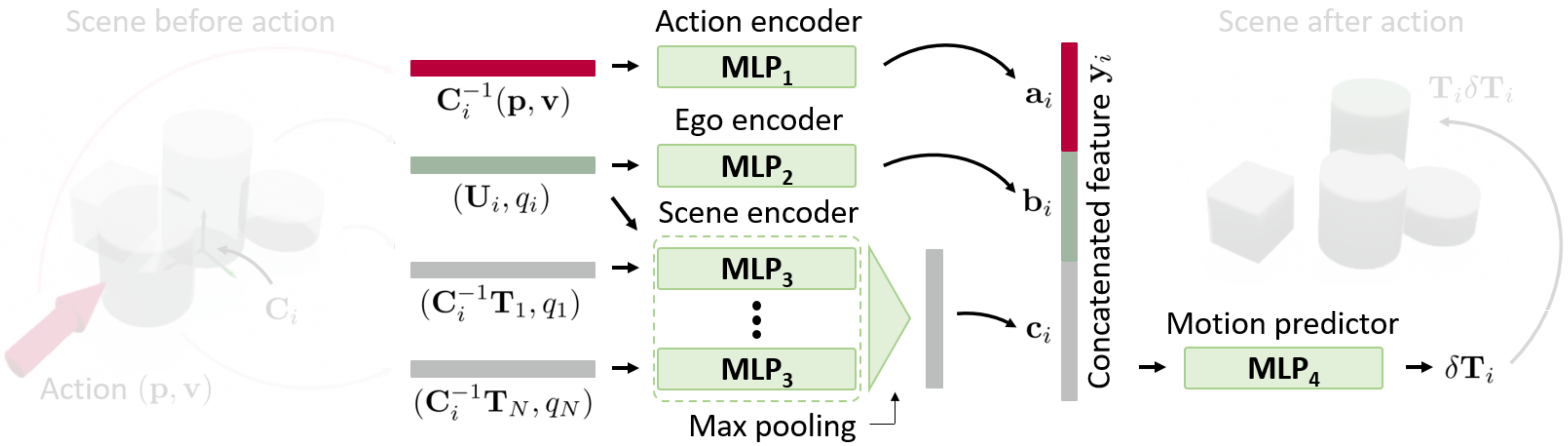
SE(2)-Equivariant Network Architecture



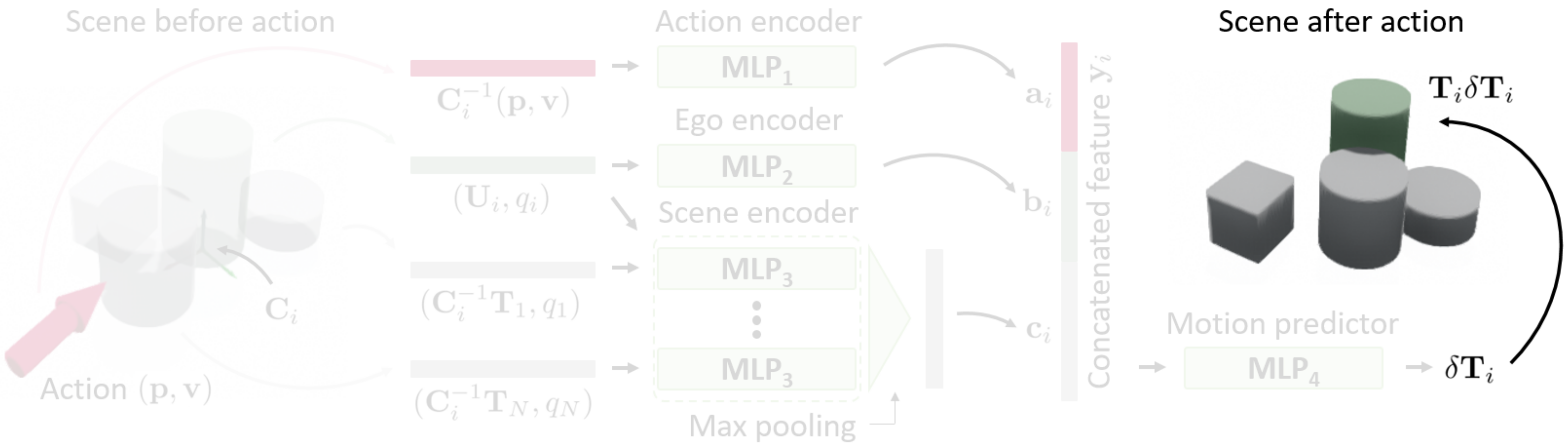
SE(2)-Equivariant Network Architecture



SE(2)-Equivariant Network Architecture

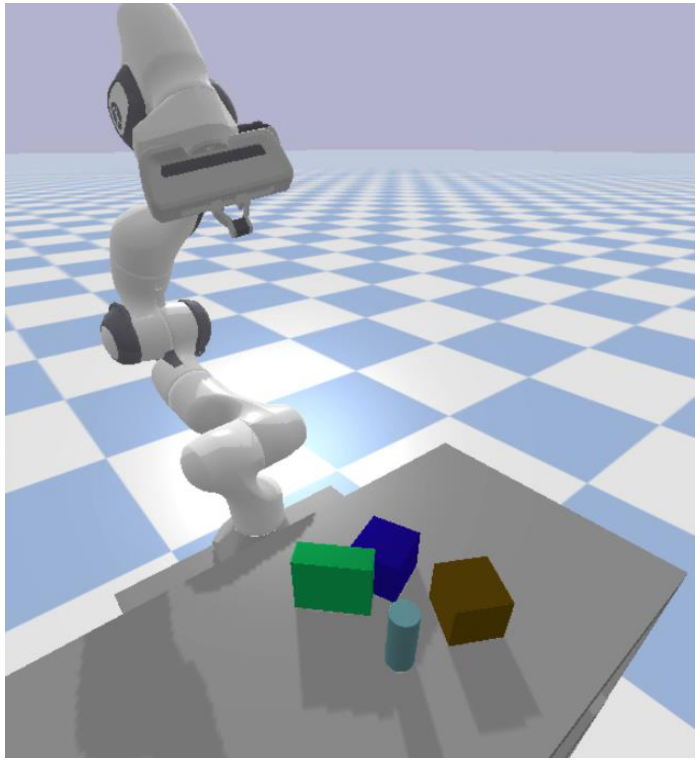


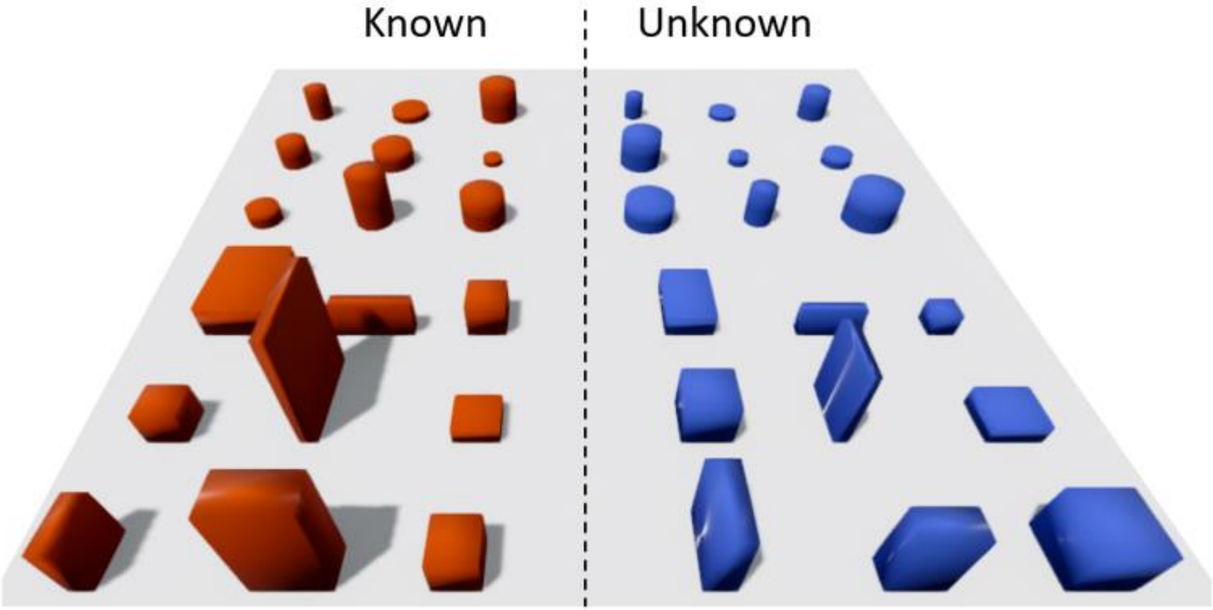
SE(2)-Equivariant Network Architecture



Experimental Results

Pushing manipulation dataset



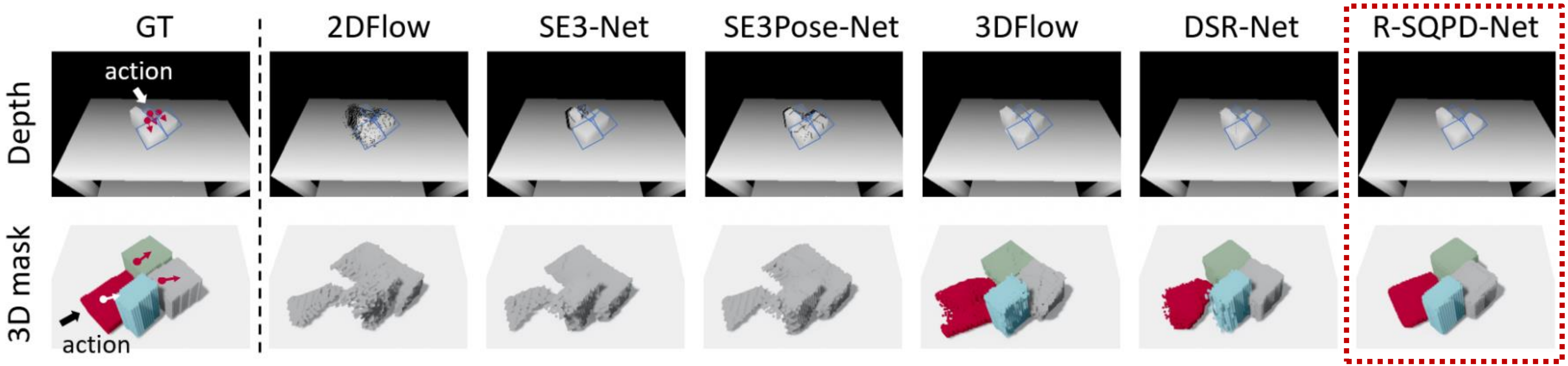


Experimental Results

METHOD	Known: Flow error (↓) visible	Known: Flow error (↓) full	Known: Mask IoU (↑) 2D	Known: Mask IoU (↑) 3D	Unknown: Flow error (↓) visible	Unknown: Flow error (↓) full	Unknown: Mask IoU (↑) 2D	Unknown: Mask IoU (↑) 3D
2DFlow [17]	2.179	-	-	-	2.180	-	-	-
SE3-Net [17]	1.631	-	-	-	1.701	-	-	-
SE3Pose-Net [18]	1.639	-	-	-	1.712	-	-	-
3DFlow [20]	1.818	1.859	0.747	0.699	1.697	1.719	0.755	0.698
DSR-Net [20]	1.325	1.331	0.720	0.705	1.531	1.524	0.665	0.632
R-SQPD-Net (ours)	**0.575**	**0.610**	**0.844**	**0.798**	**0.710**	**0.726**	**0.834**	**0.781**

Table 2: Evaluation metrics computed within test dataset (the unit of flow error is cm).

Experimental Results



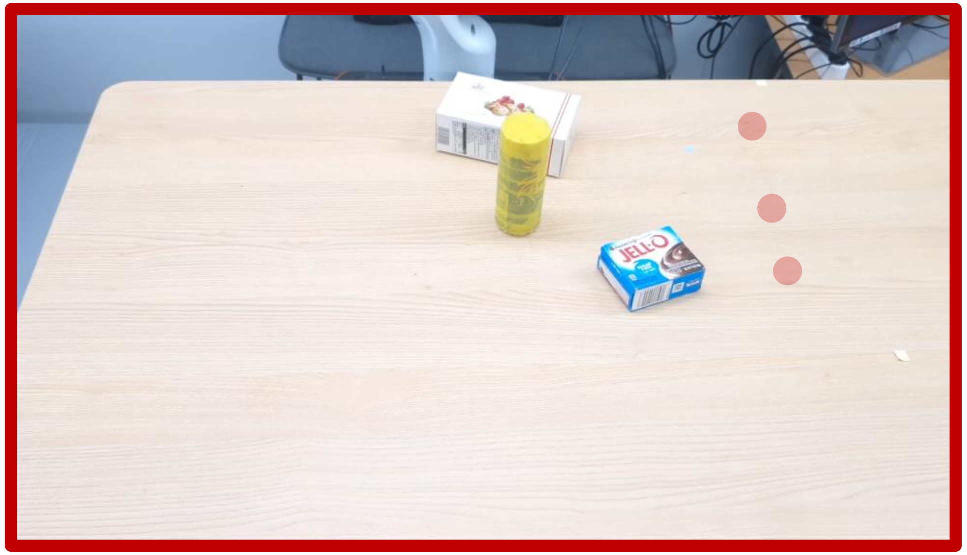
Robot Pushing Manipulation

Object moving task

Object singulation task

Robot Pushing Manipulation

Object moving task



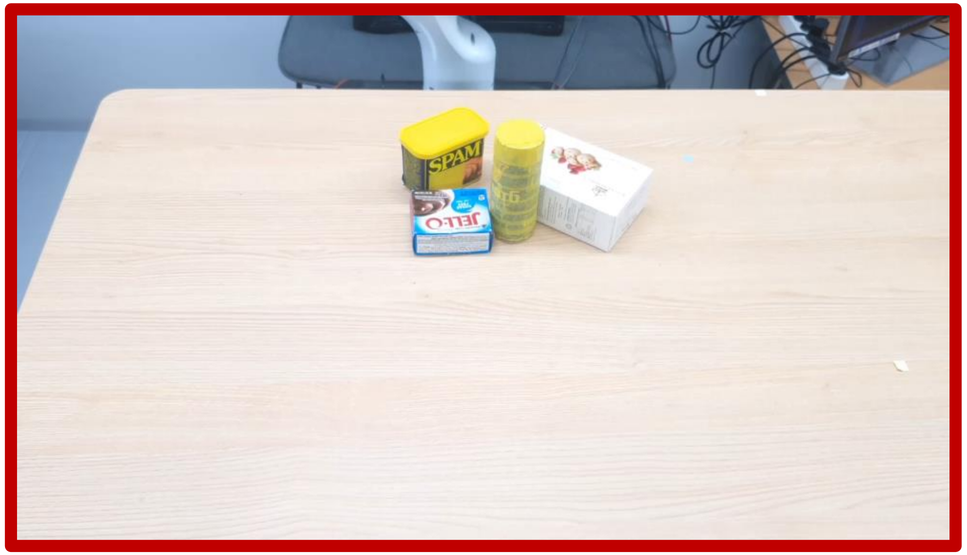
- Move the objects to their desired poses.

Object singulation task

Robot Pushing Manipulation

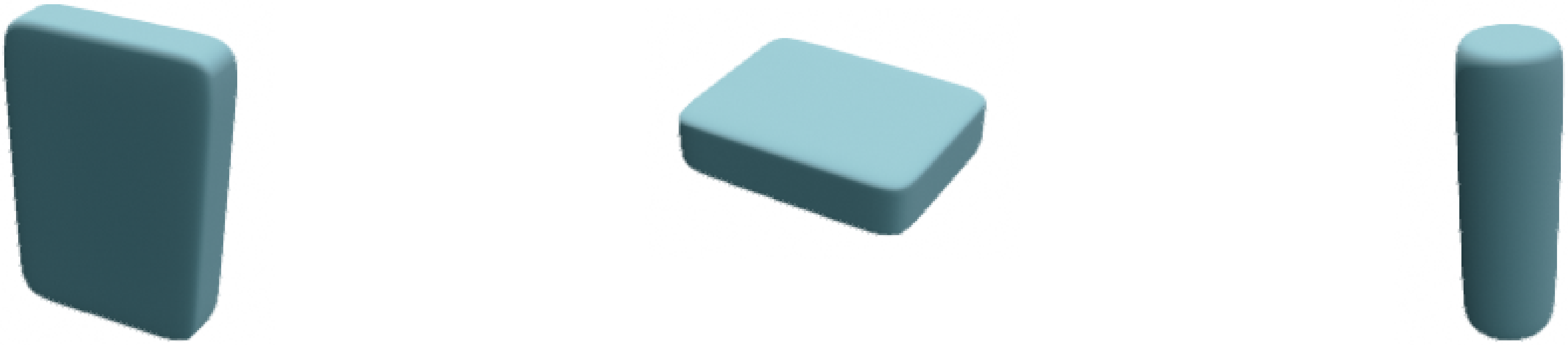
Object moving task

Object singulation task

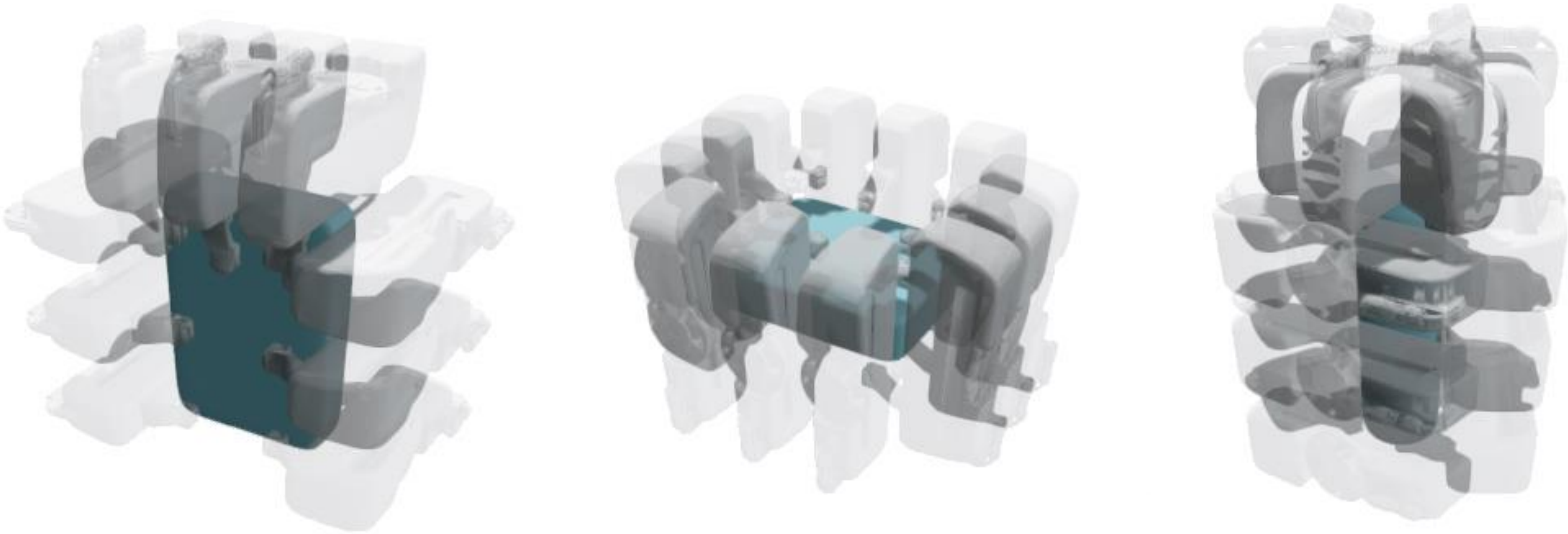


- Separate the objects by more than a certain distance τ (e.g., τ = 20cm).

Robot Pushing Manipulation

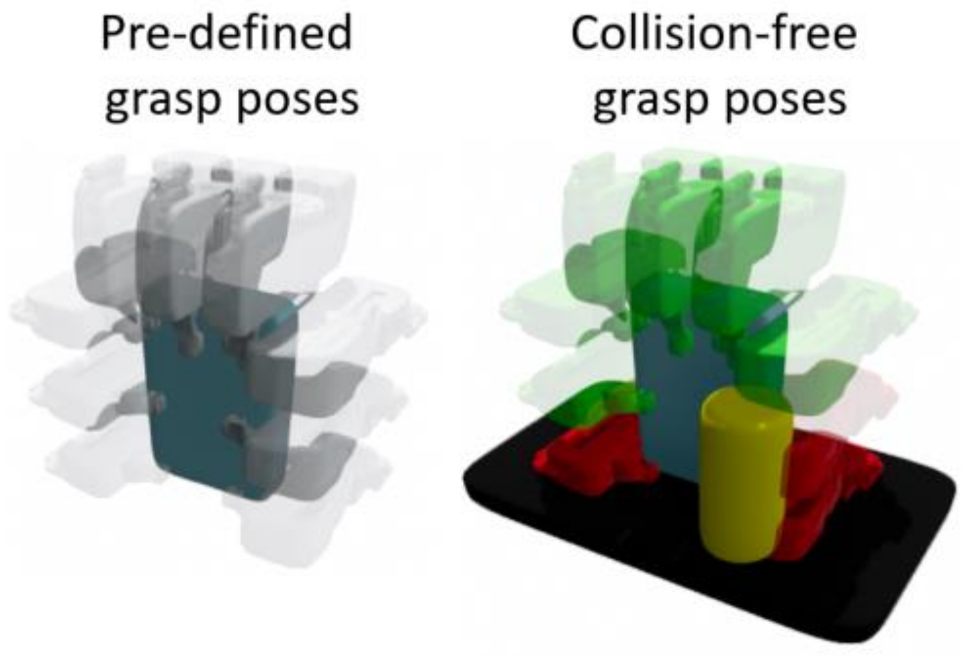


Robot Pushing Manipulation



Given the pose and shape parameters of the object, generating grasp poses is easy.

Robot Pushing Manipulation

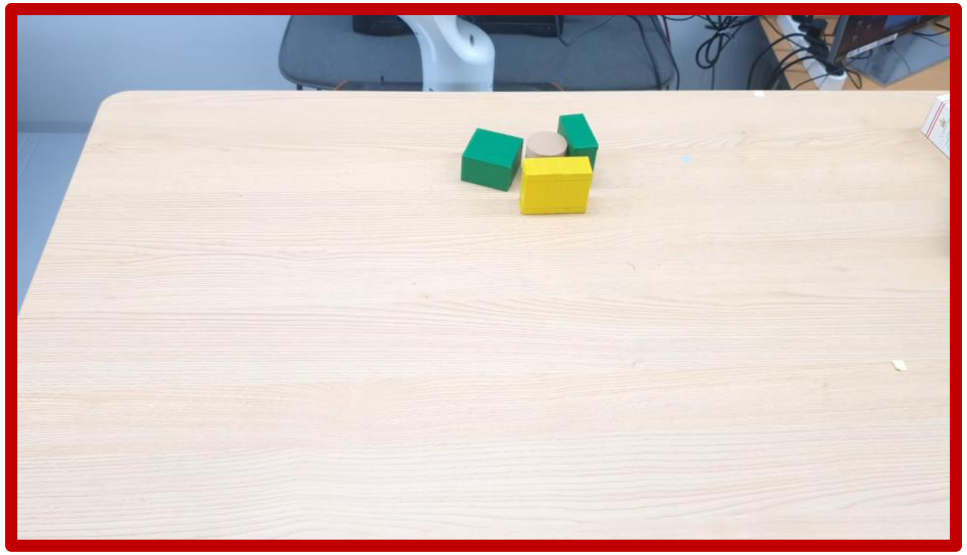


→ Grasp reward is 1 if valid grasp pose exists, 0 otherwise

Collision free from the table and other objects

Robot Pushing Manipulation

Grasping in cluttered environment



- Make the cylinder object graspable.

Grasping flat and large object

- Make Cheeze-it box graspable.

Shape Recognition-based Approaches

DSQNet
(S. Kim, et al., T-ASE'22)

SQPDNet
(S. Kim, et al., CoRL'22)

Search-for-Grasp
(S. Kim, et al. CoRL'23)

Mechanical Search on Cluttered Shelves

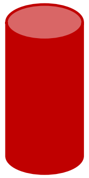
Mechanical Search on Cluttered Shelves

RGB-D image

Mechanical Search on Cluttered Shelves

RGB-D image

Target object

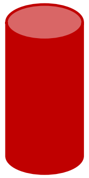


Find and grasp the desired target object on a cluttered shelf!

Mechanical Search on Cluttered Shelves

RGB-D image

Target object



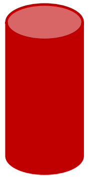
- Occluded by other objects
- Initially not visible to a camera

Find and grasp the desired target object on a cluttered shelf!

Mechanical Search on Cluttered Shelves

RGB-D image

Target object



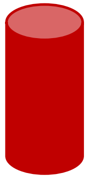
- Occluded by other objects
- Initially not visible to a camera

Find and grasp the desired target object on a cluttered shelf!

Mechanical Search on Cluttered Shelves

RGB-D image

Target object



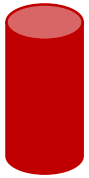
- Occluded by other objects
- Initially not visible to a camera

Find and grasp the desired target object on a cluttered shelf!

Mechanical Search on Cluttered Shelves

RGB-D image

Target object



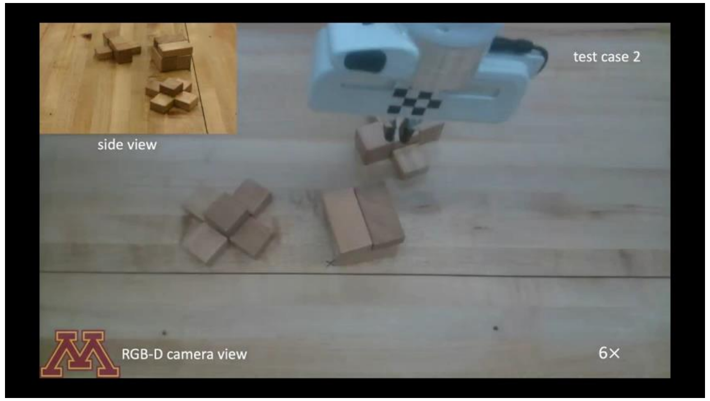
Found!!

- Occluded by other objects
- Initially not visible to a camera

Find and grasp the desired target object on a cluttered shelf!

Mechanical Search Methods

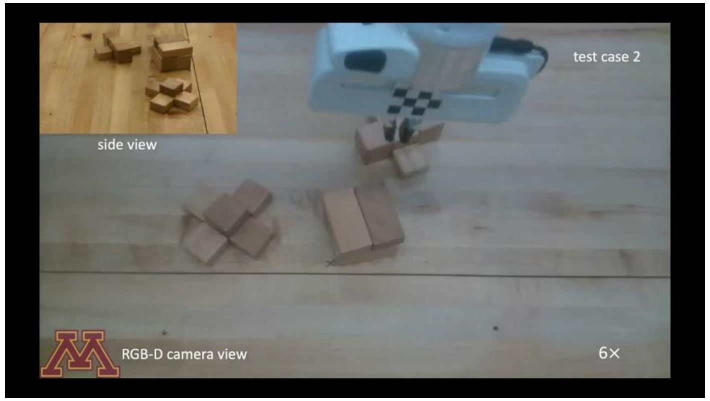
X-RAY (M. Danielczuk, et al., IROS'20)



Grasping Invisible (Y. Yang, et al., RA-L'20)

Mechanical Search Methods

X-RAY (M. Danielczuk, et al., IROS'20)

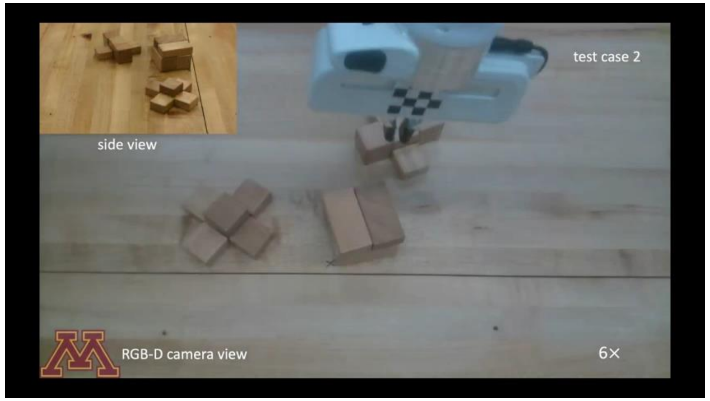


Grasping Invisible (Y. Yang, et al., RA-L'20)

Cannot be directly applied to the shelf environment!

Mechanical Search Methods

X-RAY (M. Danielczuk, et al., IROS'20)



Grasping Invisible (Y. Yang, et al., RA-L'20)

Cannot be directly applied to the shelf environment!

- **Limited action space** of the manipulator
- **Limited** amount of **visual information**

Mechanical Search Methods

X-RAY (M. Danielczuk, et al., IROS'20)

Grasping Invisible (Y. Yang, et al., RA-L'20)

Cannot be directly applied to the shelf environment!

- Limited action space of the manipulator
- Limited amount of visual information is limited

LAX-RAY (H. Huang, et al., IROS'21)

Bluction-DAR (H. Huang, et al., ICRA'22)

Mechanical Search Methods

X-RAY (M. Danielczuk, et al., IROS'20)

Grasping Invisible (Y. Yang, et al., RA-L'20)

Cannot be directly applied to the shelf environment!

- Limited action space of the manipulator
- Limited amount of visual information is limited

LAX-RAY (H. Huang, et al., IROS'21)

Bluction-DAR (H. Huang, et al., ICRA'22)

- They use a **custom long suction gripper** specialized for mechanical search.

A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

Existence Function $f\colon \mathrm{SE}(3) \rightarrow \{0, 1\}$

Graspability Function $g\colon \mathrm{SE}(3) \rightarrow \{0, 1\}$

A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

Existence Function $f\colon \mathrm{SE}(3) \rightarrow \{0, 1\}$

$f(x)$ indicates whether the target object **can be present** at the pose x or not.

Graspability Function $g\colon \mathrm{SE}(3) \rightarrow \{0, 1\}$

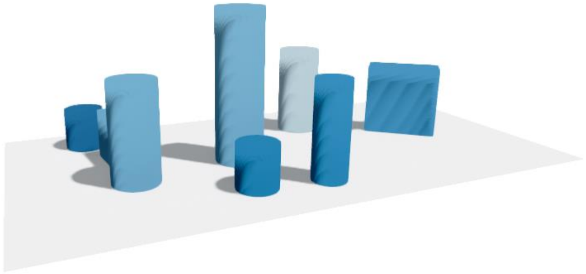
$g(x)$ indicates whether the target object at the pose x is **graspable** or not.

A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

Existence Function $f : \mathrm{SE}(3) \to \{0, 1\}$

Given Observation

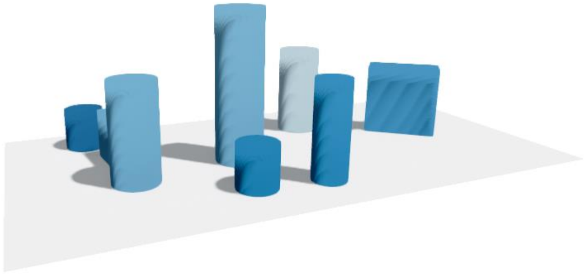


A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

Existence Function $f: \mathrm{SE}(3) \rightarrow \{0, 1\}$

Given Observation



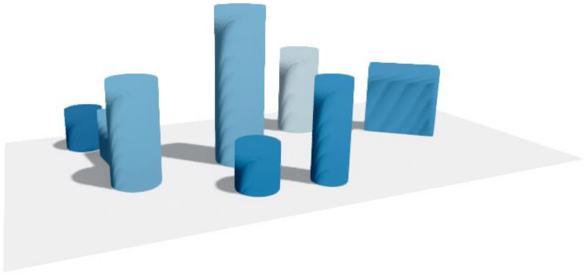
Candidate poses

A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

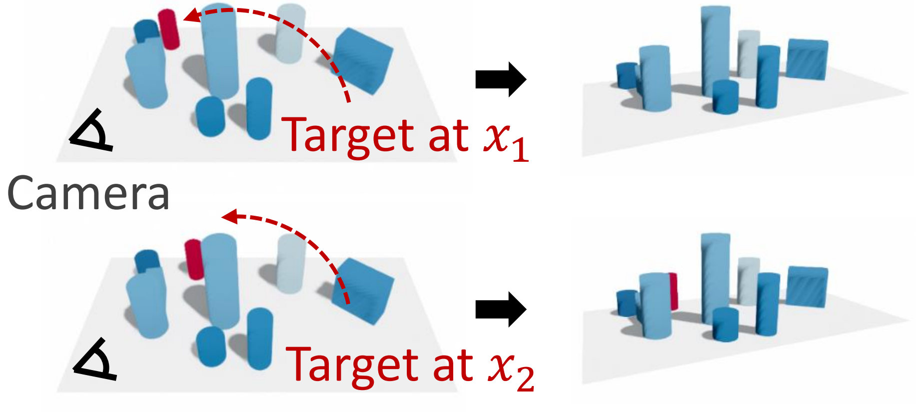
Existence Function $f: \mathrm{SE}(3) \rightarrow \{0, 1\}$

Given Observation



Candidate poses

Observation

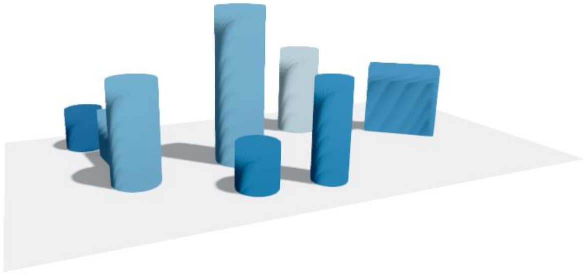


A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

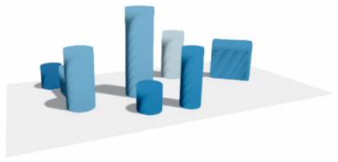
Existence Function $f: \mathrm{SE}(3) \rightarrow \{0, 1\}$

Given Observation



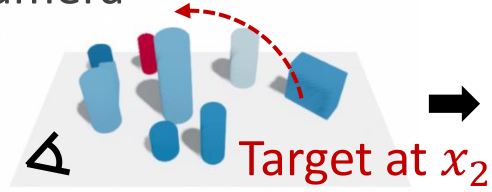
Candidate poses

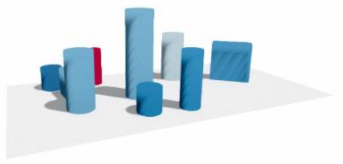
Observation



$f(x_1) = 1$

Camera



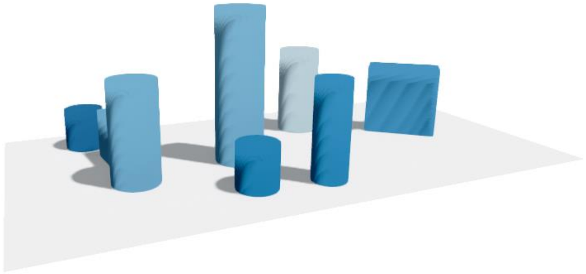


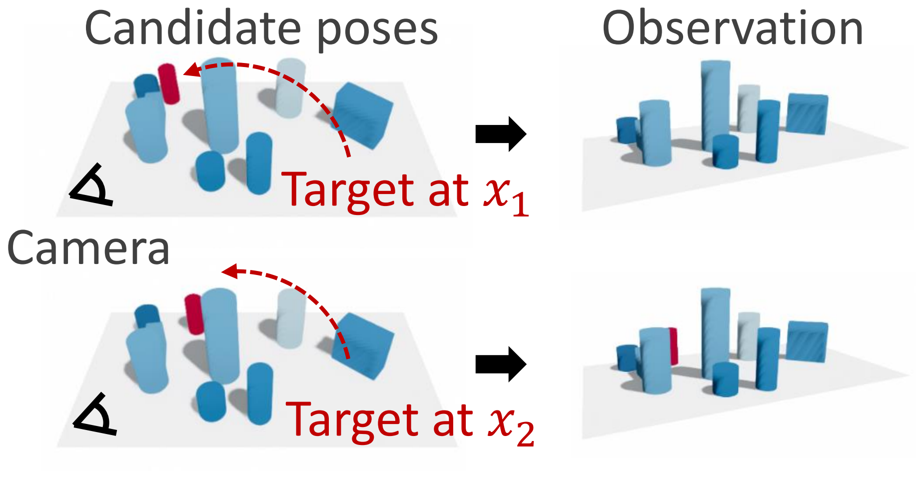
A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

Existence Function $f\colon \mathrm{SE}(3) \to \{0, 1\}$

Given Observation





$f(x_1) = 1$

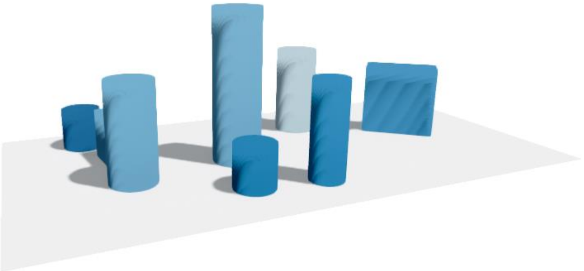
$f(x_2) = 0$

A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

Existence Function $f\colon \mathrm{SE}(3) \to \{0,1\}$

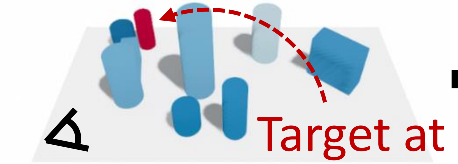
Given Observation



$$\sum_{x\in\mathcal{X}} f(x) \uparrow$$

uncertainty of actual target pose ↑

Candidate poses

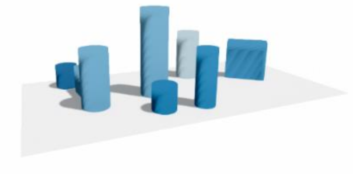


Target at x_1

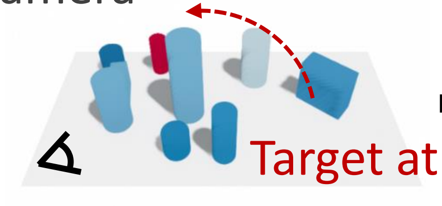
Camera

➡

Observation



$f(x_1) = 1$



Target at x_2

➡

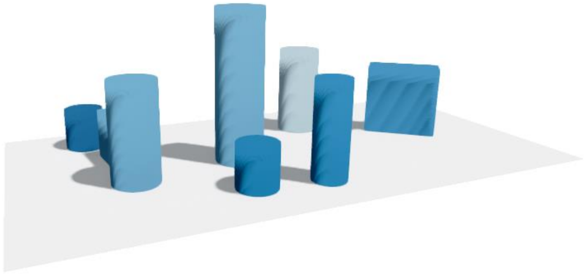
$f(x_2) = 0$

A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

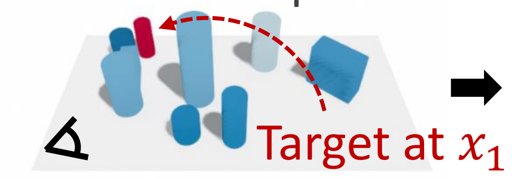
Existence Function $f: \mathrm{SE}(3) \rightarrow \{0, 1\}$

Given Observation



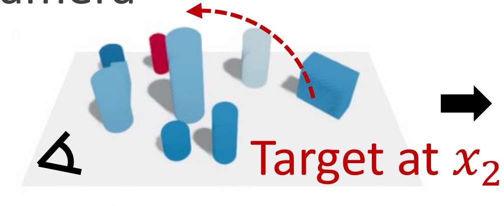
Candidate poses

Observation



$f(x_1) = 1$

Camera



$f(x_2) = 0$

$$\sum_{x \in \mathcal{X}} f(x) \uparrow$$

uncertainty of actual target pose ↑

To find the fully-occluded target object, we should minimize $\sum_{x \in \mathcal{X}} f(x)$

A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

Existence Function $f\colon \mathrm{SE}(3) \to \{0, 1\}$

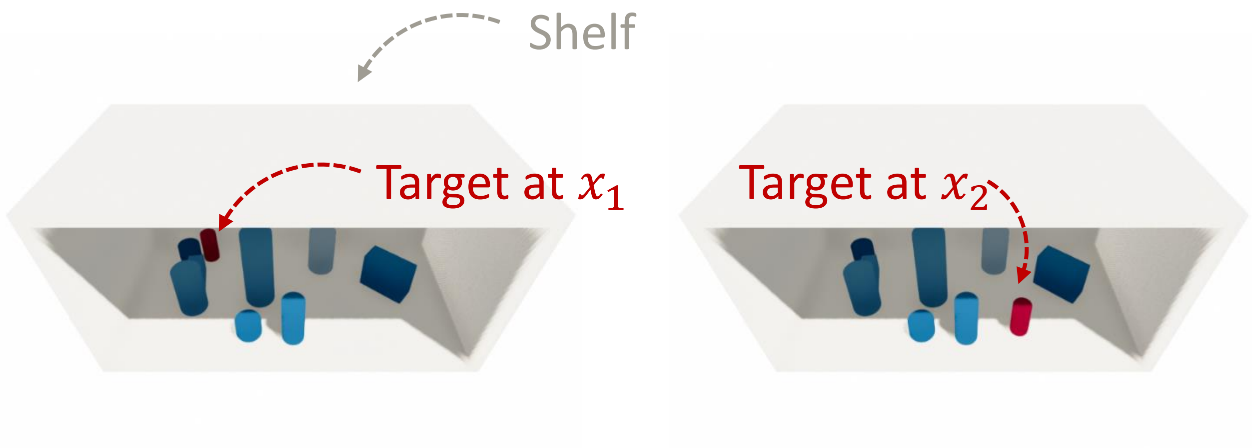
Graspability Function $g\colon \mathrm{SE}(3) \to \{0, 1\}$

A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

Existence Function $f: \mathrm{SE}(3) \rightarrow \{0, 1\}$

Graspability Function $g: \mathrm{SE}(3) \rightarrow \{0, 1\}$

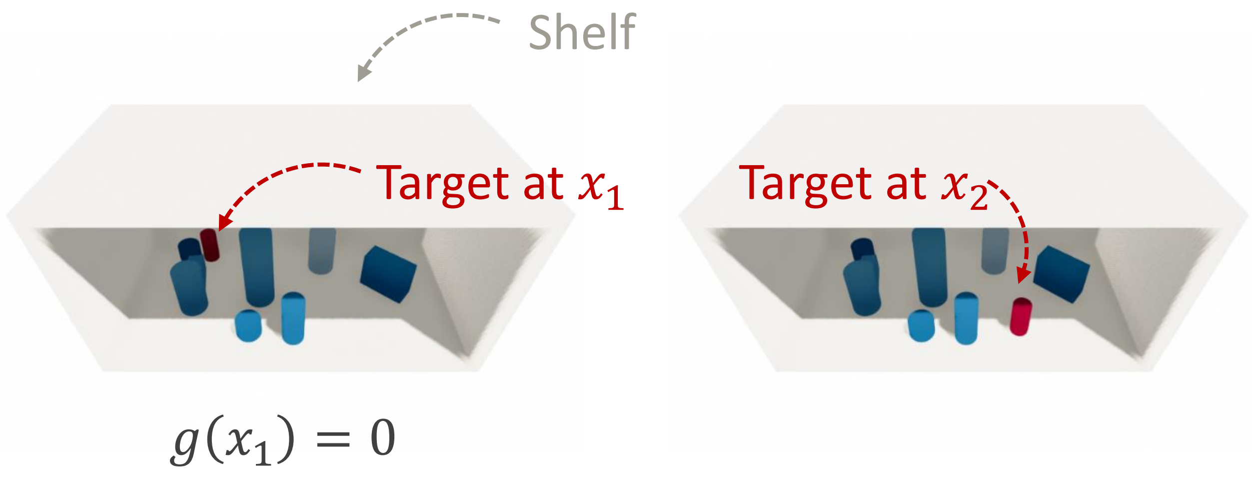


A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

Existence Function $f: \mathrm{SE}(3) \rightarrow \{0, 1\}$

Graspability Function $g: \mathrm{SE}(3) \rightarrow \{0, 1\}$

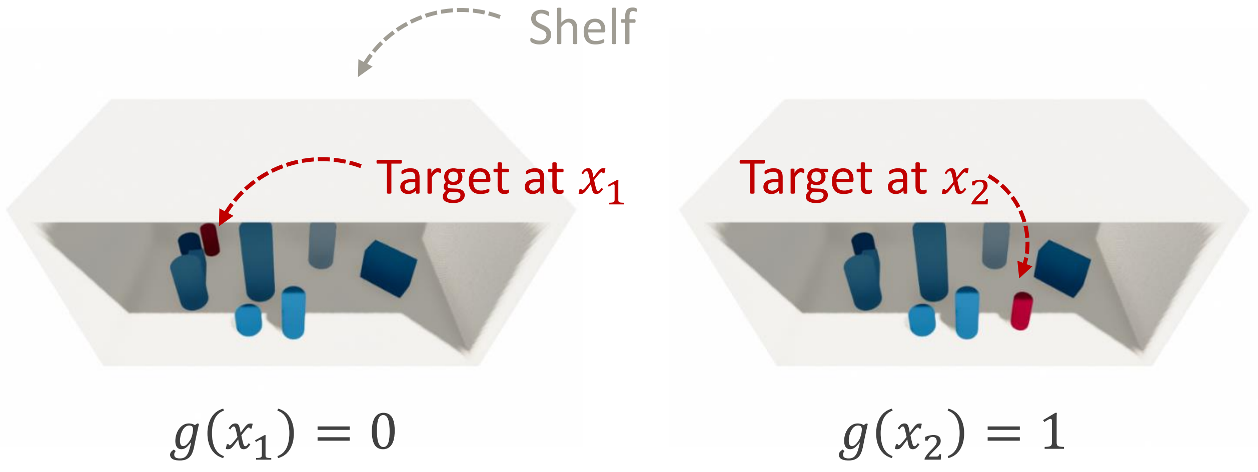


A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

Existence Function $f: \mathrm{SE}(3) \rightarrow \{0, 1\}$

Graspability Function $g: \mathrm{SE}(3) \rightarrow \{0, 1\}$



A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

Existence Function $f\colon \mathrm{SE}(3) \to \{0, 1\}$

Graspability Function $g\colon \mathrm{SE}(3) \to \{0, 1\}$

A General Framework for Mechanical Search

Find and grasp the desired target object on a cluttered shelf!

$\boxed{\text{Existence Function } f\colon \mathrm{SE}(3) \to \{0,1\}}$

$\boxed{\text{Graspability Function } g\colon \mathrm{SE}(3) \to \{0,1\}}$

Optimal control formulation

$$\min_{\{a_i\}_{i=1}^{T}} \sum_{x \in \mathcal{X}} \boxed{f_T(x)} + \boxed{\alpha f_T(x)\big(1 - g_T(x)\big)}$$

Leveraging Shape Recognition

Find and grasp the desired target object on a cluttered shelf!

Existence Function $f\colon \mathrm{SE}(3) \to \{0, 1\}$

Graspability Function $g\colon \mathrm{SE}(3) \to \{0, 1\}$

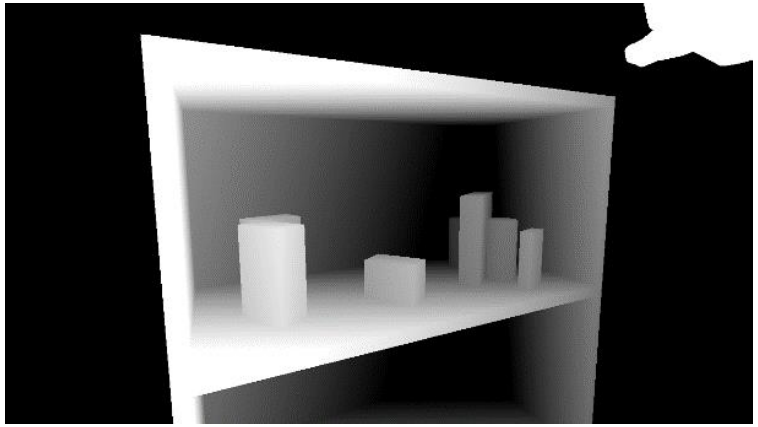
Leveraging Shape Recognition

Find and grasp the desired target object on a cluttered shelf!

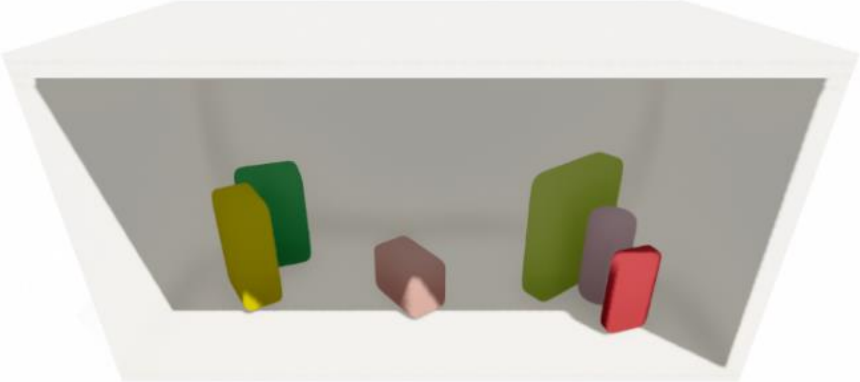
Existence Function $\hat{f}: \mathrm{SE}(3) \rightarrow \{0, 1\}$

Graspability Function $\hat{g}: \mathrm{SE}(3) \rightarrow \{0, 1\}$

Observed Depth image



Recognized 3D objects



Leveraging Shape Recognition

Find and grasp the desired target object on a cluttered shelf!

Existence Function $f\colon \mathrm{SE}(3) \to \{0,1\}$

Graspability Function $g\colon \mathrm{SE}(3) \to \{0,1\}$

Optimal control formulation

$$\min_{\{a_i\}_{i=1}^{T}} \sum_{x\in\mathcal{X}} f_T(x) + \alpha f_T(x)\big(1 - g_T(x)\big)$$

Leveraging Shape Recognition

Find and grasp the desired target object on a cluttered shelf!

Existence Function $f: \mathrm{SE}(3) \rightarrow \{0, 1\}$

Graspability Function $g: \mathrm{SE}(3) \rightarrow \{0, 1\}$

Optimal control formulation

$$\min_{\{a_i\}_{i=1}^{T}} \sum_{x \in \mathcal{X}} f_T(x) + \alpha f_T(x)\big(1 - g_T(x)\big)$$

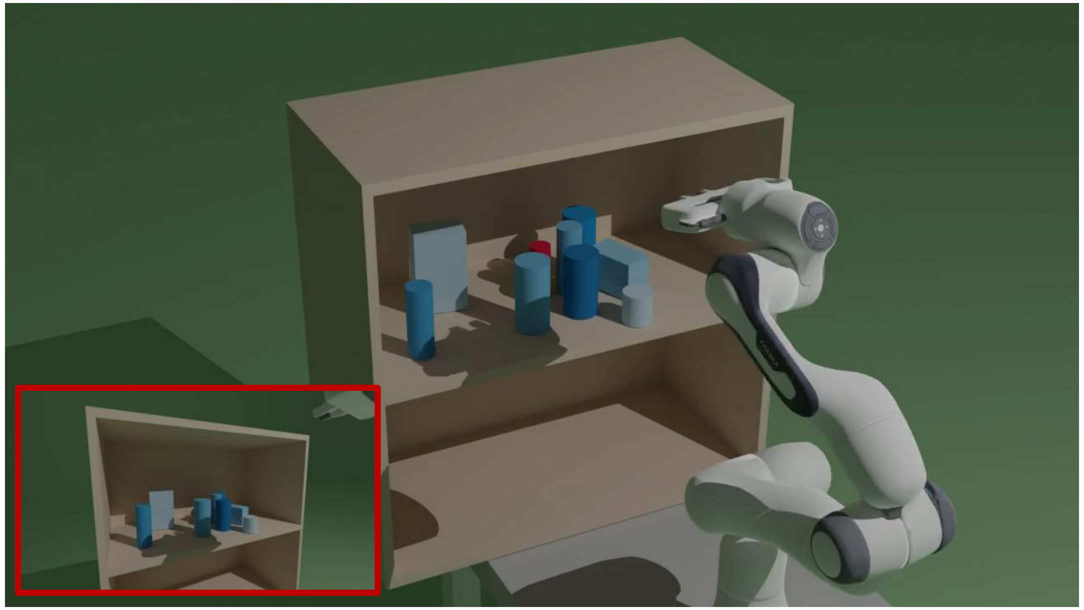
Approximate by leveraging 3D shape recognition

Tractable optimal control formulation

$$\min_{\{a_i\}_{i=1}^{T}} \sum_{x \in \mathcal{X}} \hat{f}_T(x) + \alpha \hat{f}_T(x)\big(1 - \hat{g}_T(x)\big)$$

Experimental Results

Simulation environment



Real-world environment

Experimental Results

METHOD		The number of objects 2 Find	2 Grasp	4 Find	4 Grasp	6 Find	6 Grasp	8 Find	8 Grasp
O-Search-and-Grasp	Succ.	0.98	0.96	1.0	0.88	1.0	0.84	0.98	0.66
	Steps	1.163	1.132	1.32	2.136	1.86	3.286	1.694	3.485
O-Search-for-Grasp	Succ.	1.0	0.98	1.0	0.82	1.0	0.8	1.0	0.66
	Steps	1.24	1.408	1.36	1.854	1.66	2.5	1.74	3.212
R-Search-and-Grasp	Succ.	1.0	0.96	0.96	0.84	0.98	0.66	0.98	0.56
	Steps	1.46	1.551	1.562	2.065	1.653	3.3	2.102	3.73
R-Search-for-Grasp	Succ.	1.0	0.98	1.0	0.88	1.0	0.72	0.98	0.6
	Steps	1.34	1.531	1.74	2.543	1.8	2.4	1.653	3.846

Table 1: Simulation manipulation results

Experimental Results

Cluttered shelf with 3~4 occluding objects

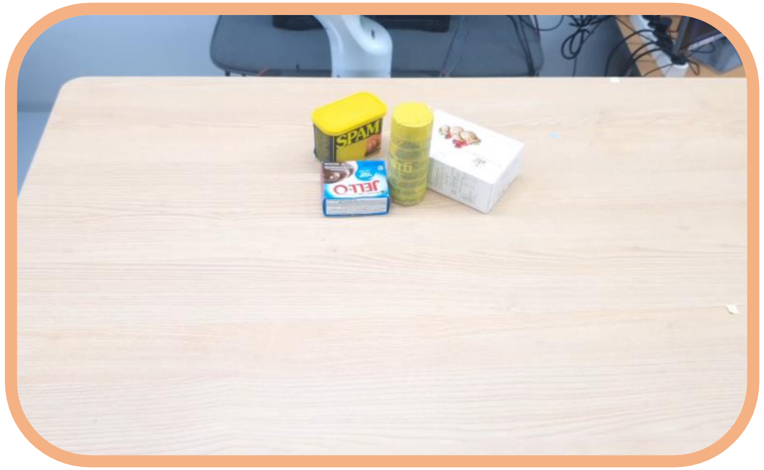
- Find and grasp the target red cylinder.

Cluttered shelf with 5~6 occluding objects

- Find and grasp the target red cylinder.

Conclusion

DSQNet
(S. Kim, et al., T-ASE'22)



SQPDNet
(S. Kim, et al., CoRL'22)

Search-for-Grasp
(S. Kim, et al. CoRL'23)

Conclusion

- We propose a shape recognition-based approach for learning vision-based object manipulation.

Conclusion

- We propose a shape recognition-based approach for learning vision-based object manipulation.

- **DSQNet**
  - We have proposed a novel shape recognition-based grasping using deformable superquadrics and deep neural networks.
  - Our method shows the best success rates among shape recognition-based grasping methods.

Conclusion

- We propose a shape recognition-based approach for learning vision-based object manipulation.

- **DSQNet**
  - We have proposed a novel shape recognition-based grasping using deformable superquadrics and deep neural networks.
  - Our method shows the best success rates among shape recognition-based grasping methods.

- **SQPDNet**
  - We have proposed a SE(2)-equivariant pushing dynamics model using recognized object shapes.
  - Our method significantly outperforms the existing visual pushing dynamics models.

Conclusion

- We propose a shape recognition-based approach for learning vision-based object manipulation.

- **DSQNet**
  - We have proposed a novel shape recognition-based grasping using deformable superquadrics and deep neural networks.
  - Our method shows the best success rates among shape recognition-based grasping methods.

- **SQPDNet**
  - We have proposed a SE(2)-equivariant pushing dynamics model using recognized object shapes.
  - Our method significantly outperforms the existing visual pushing dynamics models.

- **Search-for-Grasp**
  - We have proposed a novel mechanical search framework leveraging shape recognition.
  - Using standard two-finger gripper, our method can successfully find and grasp the target object by rearranging occluding objects.

References

- DSQNet
  - **Seungyeon Kim***, Taegyun Ahn*, Yonghyeon Lee, Jihwan Kim, Michael Yu Wang, and Frank C. Park. *DSQNet: A Deformable Model-Based Supervised Learning Algorithm for Grasping Unknown Occluded Objects*. IEEE Transactions on Automation Science and Engineering (2022).
- SQPDNet
  - **Seungyeon Kim**, Byeongdo Lim, Yonghyeon Lee, and Frank C. Park. *SE (2)-Equivariant Pushing Dynamics Models for Tabletop Object Manipulations*. Conference on Robot Learning (2022).
- Search-for-Grasp
  - **Seungyeon Kim***, Young Hun Kim*, Yonghyeon Lee, and Frank C. Park. *Leveraging 3D Reconstruction for Mechanical Search on Cluttered Shelves.* Conference on Robot Learning (2023).

Thank you for listening!

Contact: ksy@robotics.snu.ac.kr
Homepage: https://seungyeon-k.github.io