

Ph. D. Thesis Defense

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

Seungyeon Kim

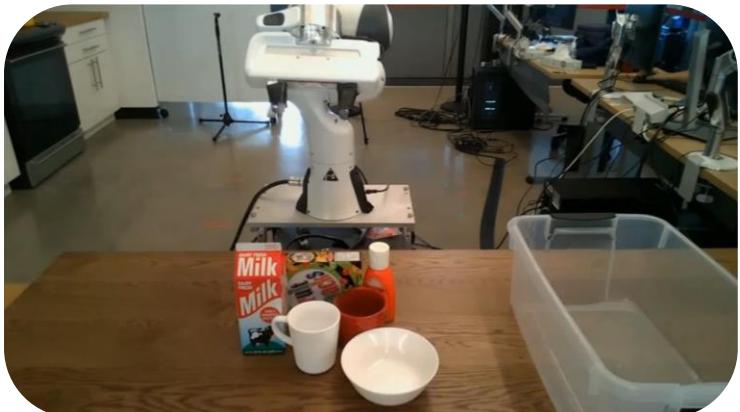
ksy@robotics.snu.ac.kr

<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-grasnet: Efficient 6-dof grasp generation in cluttered scenes." 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021.

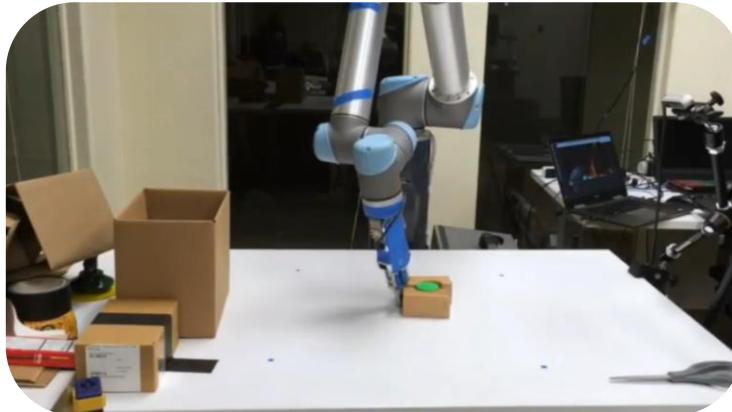
Zeng, Andy, et al. "Learning synergies between pushing and grasping with self-supervised deep reinforcement learning." 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018.

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-grasnet: Efficient 6-dof grasp generation in cluttered scenes." 2021 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2021.

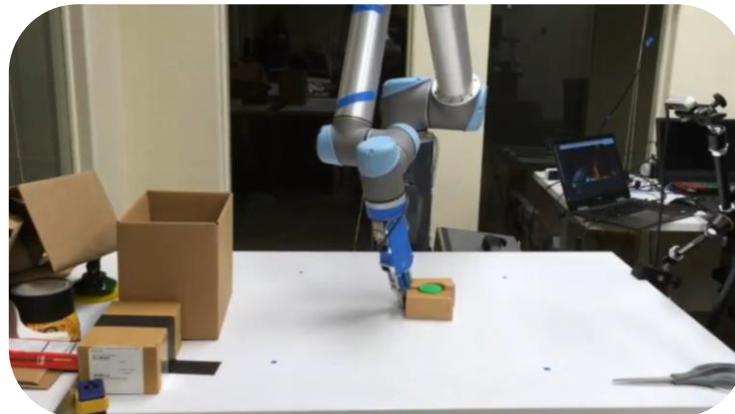
Zeng, Andy, et al. "Learning synergies between pushing and grasping with self-supervised deep reinforcement learning." 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE, 2018.

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-grasnet: 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.

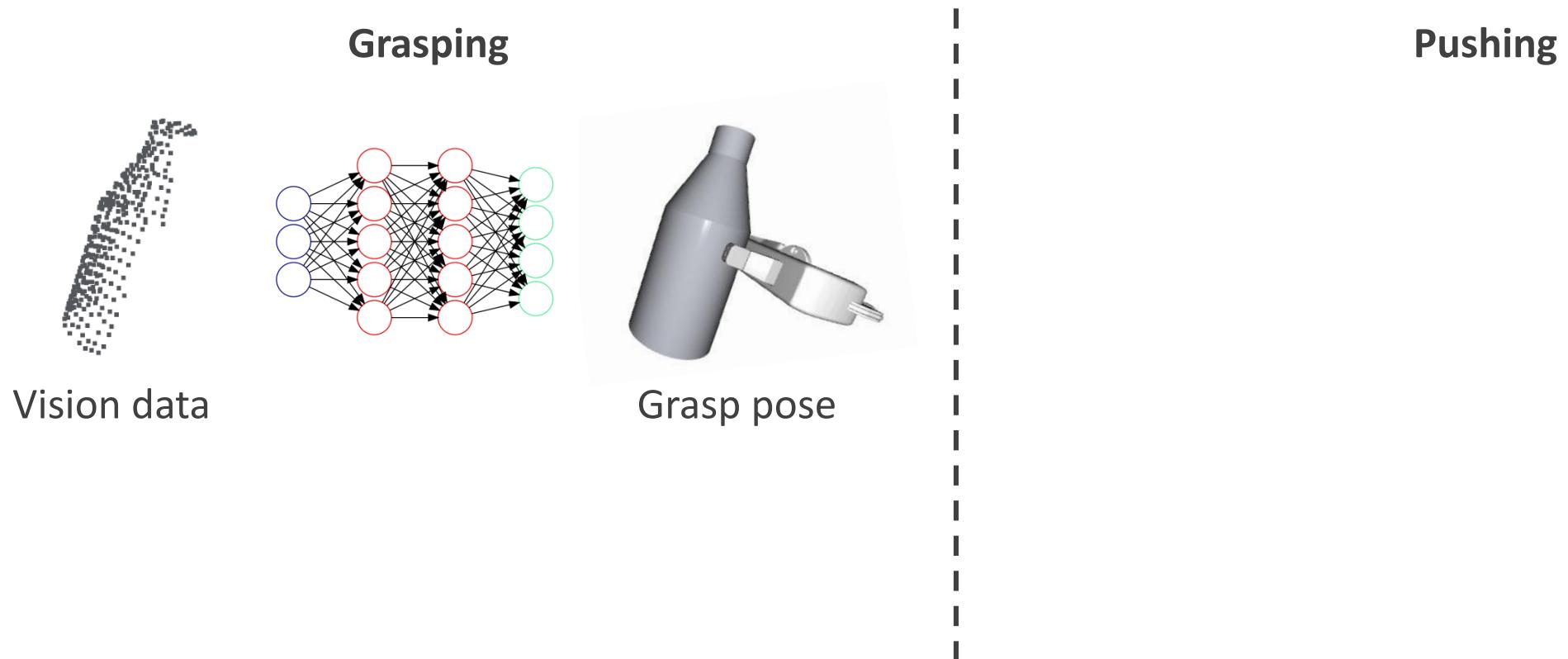
Zeng, Andy, et al. "Tossingbot: Learning to throw arbitrary objects with residual physics." IEEE Transactions on Robotics 36.4 (2020): 1307-1319.

End-to-end Approaches

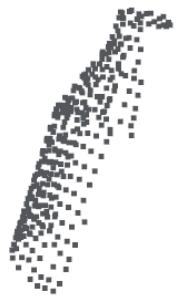
Grasping

Pushing

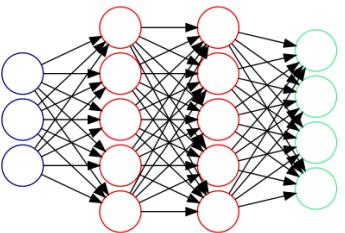
End-to-end Approaches



End-to-end Approaches



Grasping



Vision data



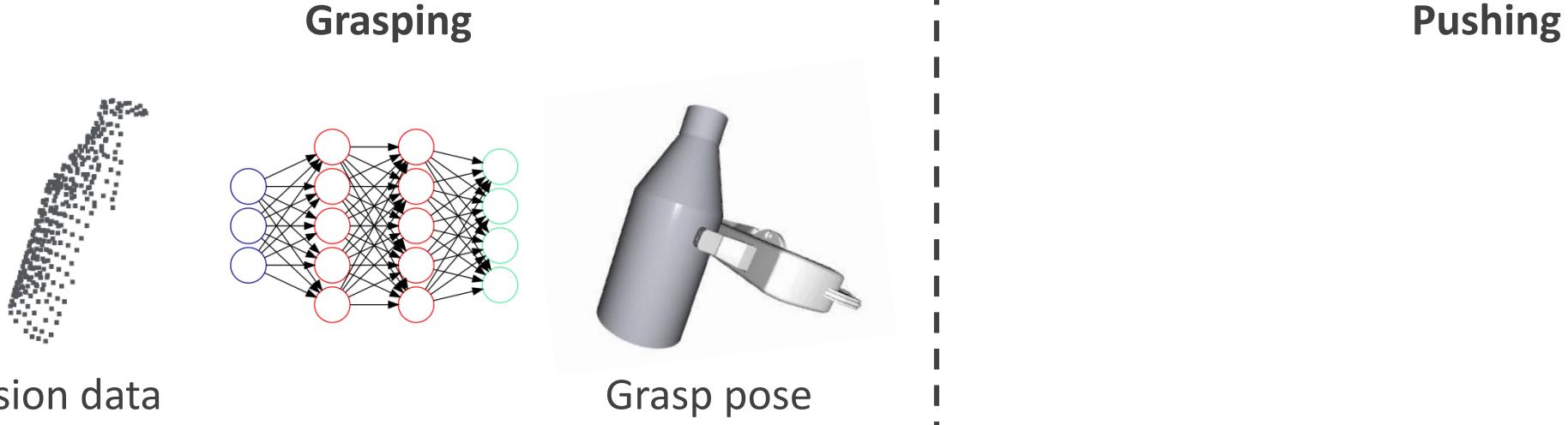
Grasp pose

($\sim 10,000,000$ pairs)

- Require large amounts of training data.

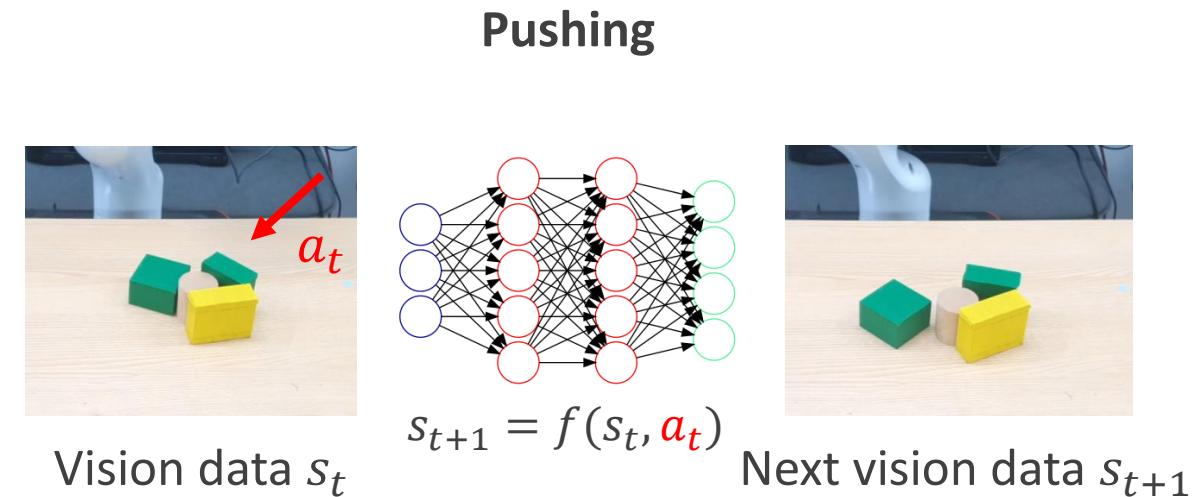
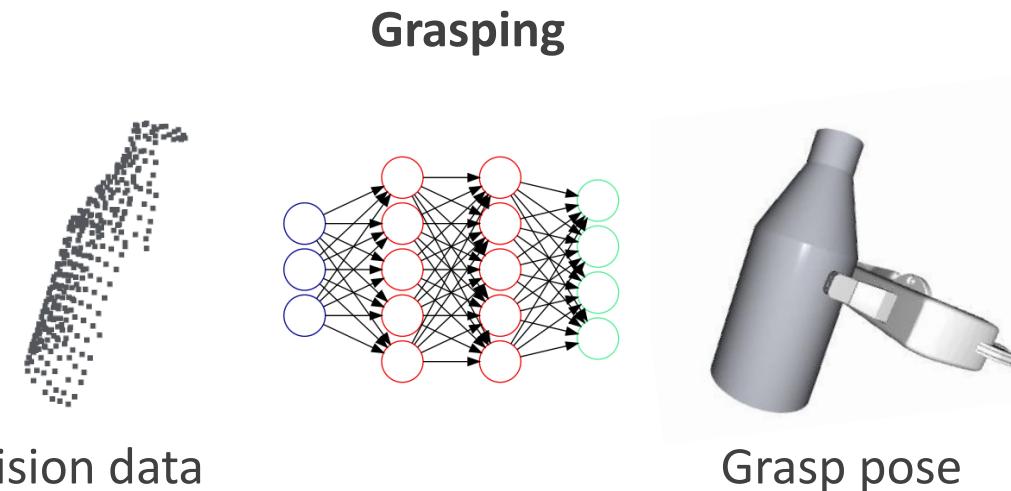
Pushing

End-to-end Approaches



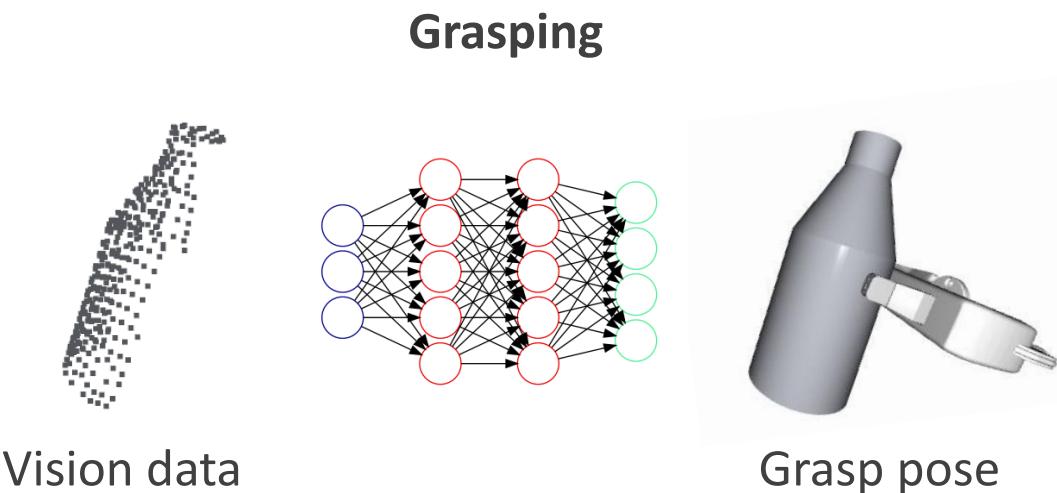
- Require large amounts of training data.
- The trained network will only work reliably for the gripper used to collect the training data.

End-to-end Approaches

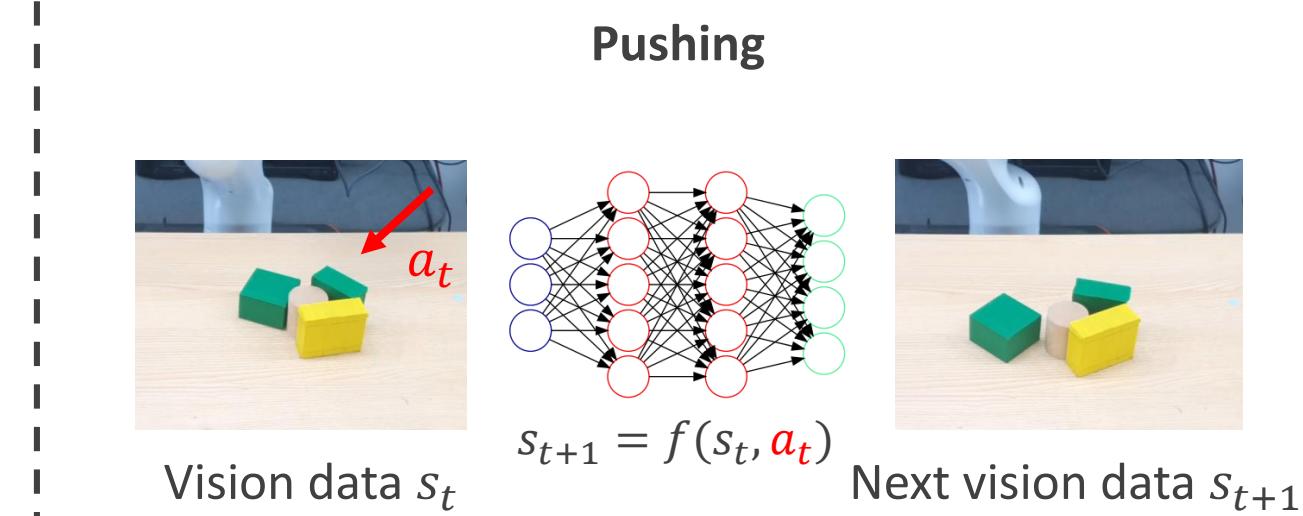


- Require large amounts of training data.
- The trained network will only work reliably for the gripper used to collect the training data.

End-to-end Approaches

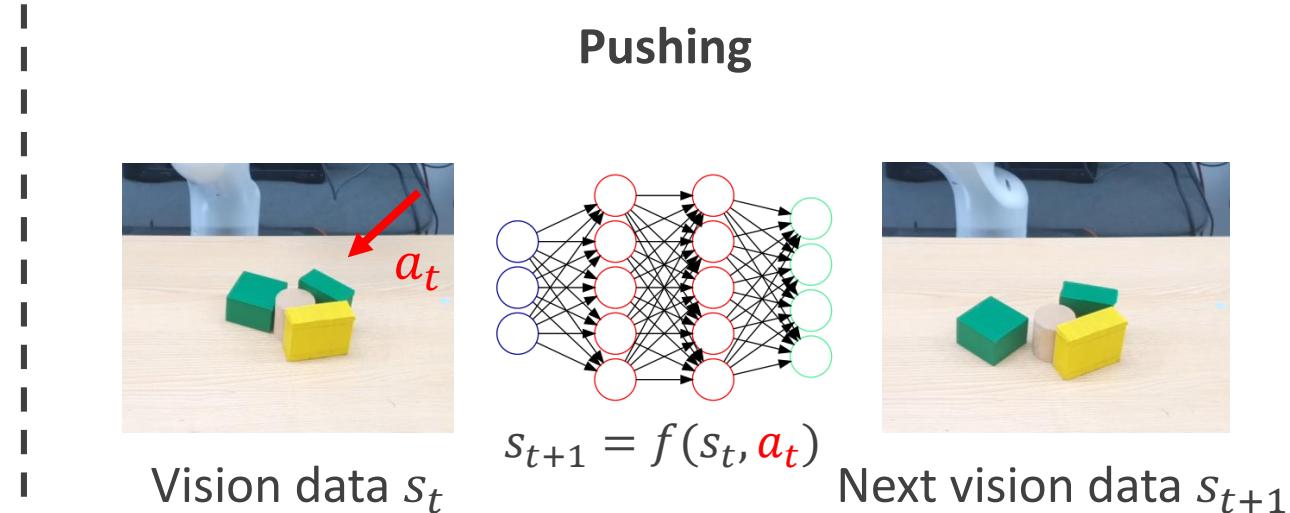
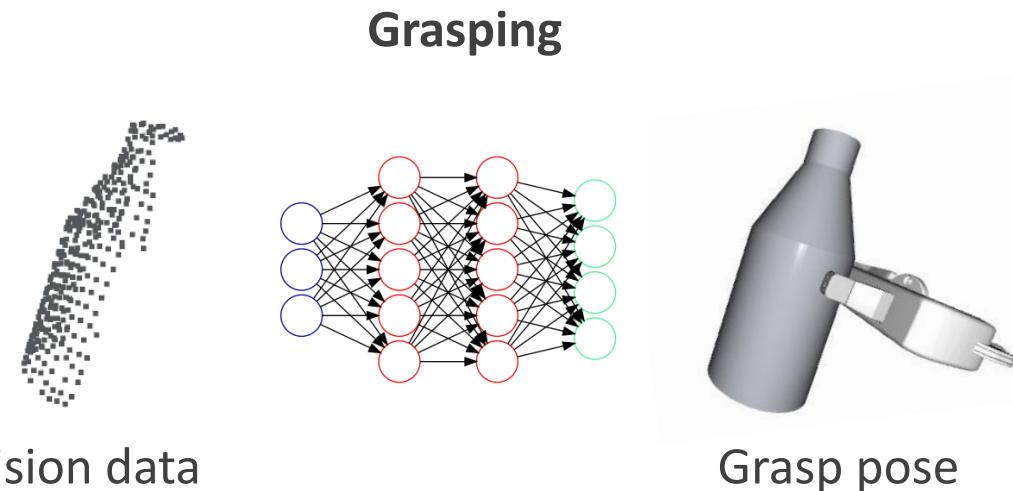


- Require large amounts of training data.
 - The trained network will only work reliably for the gripper used to collect the training data.



- Require large amounts of training data.

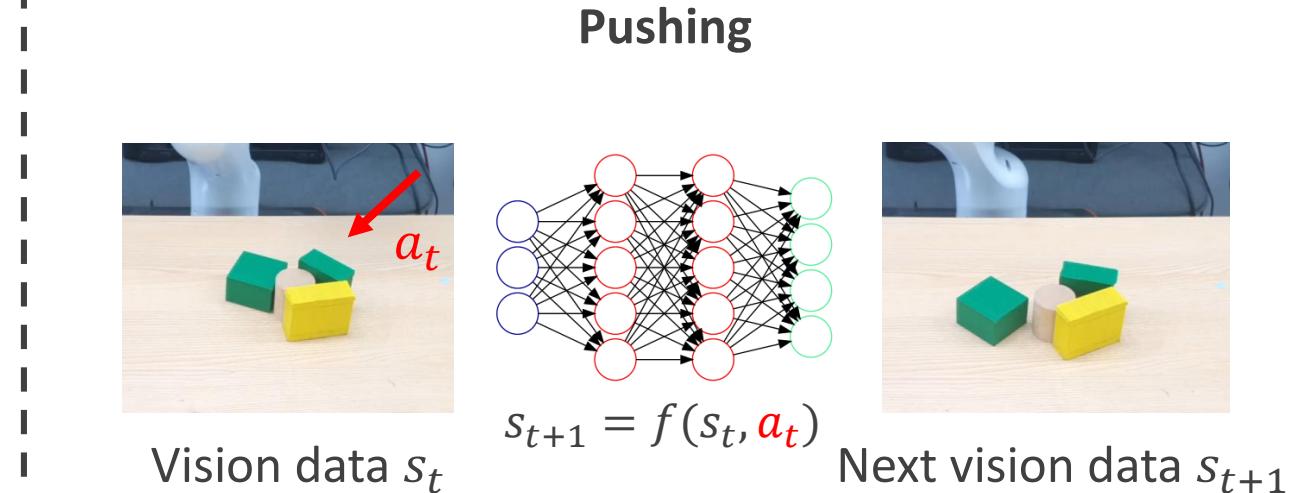
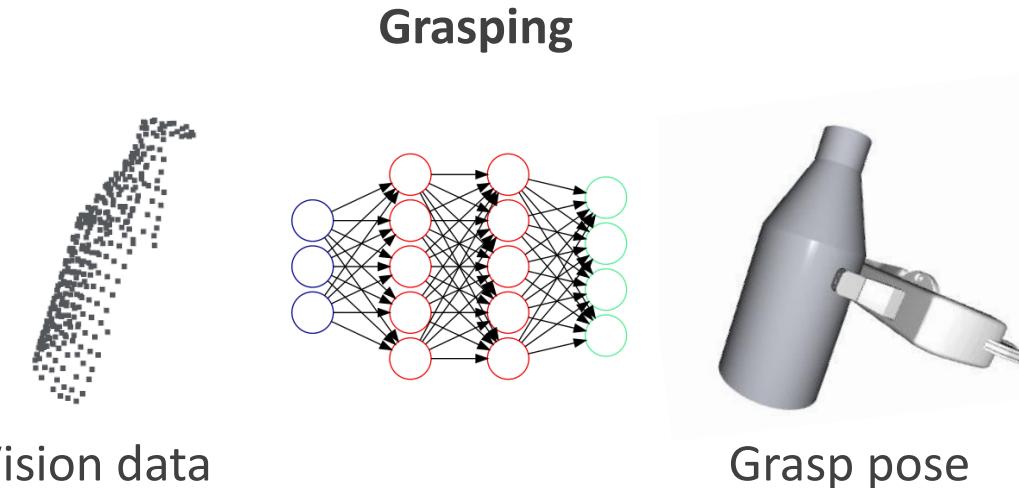
End-to-end Approaches



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

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

End-to-end Approaches



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

- 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)



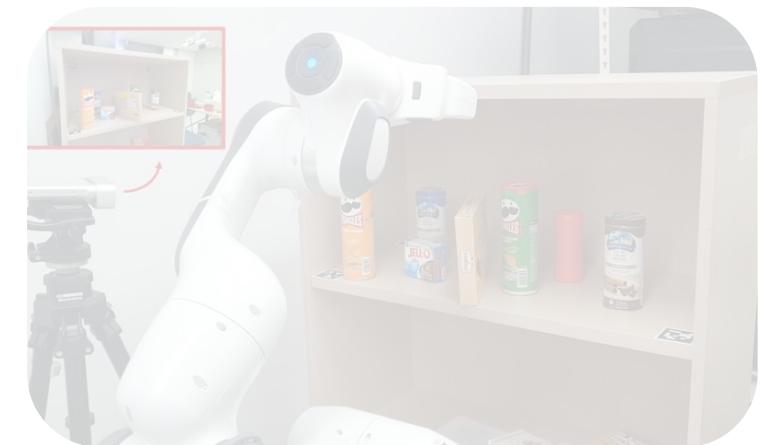
Shape Recognition-based Approaches



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



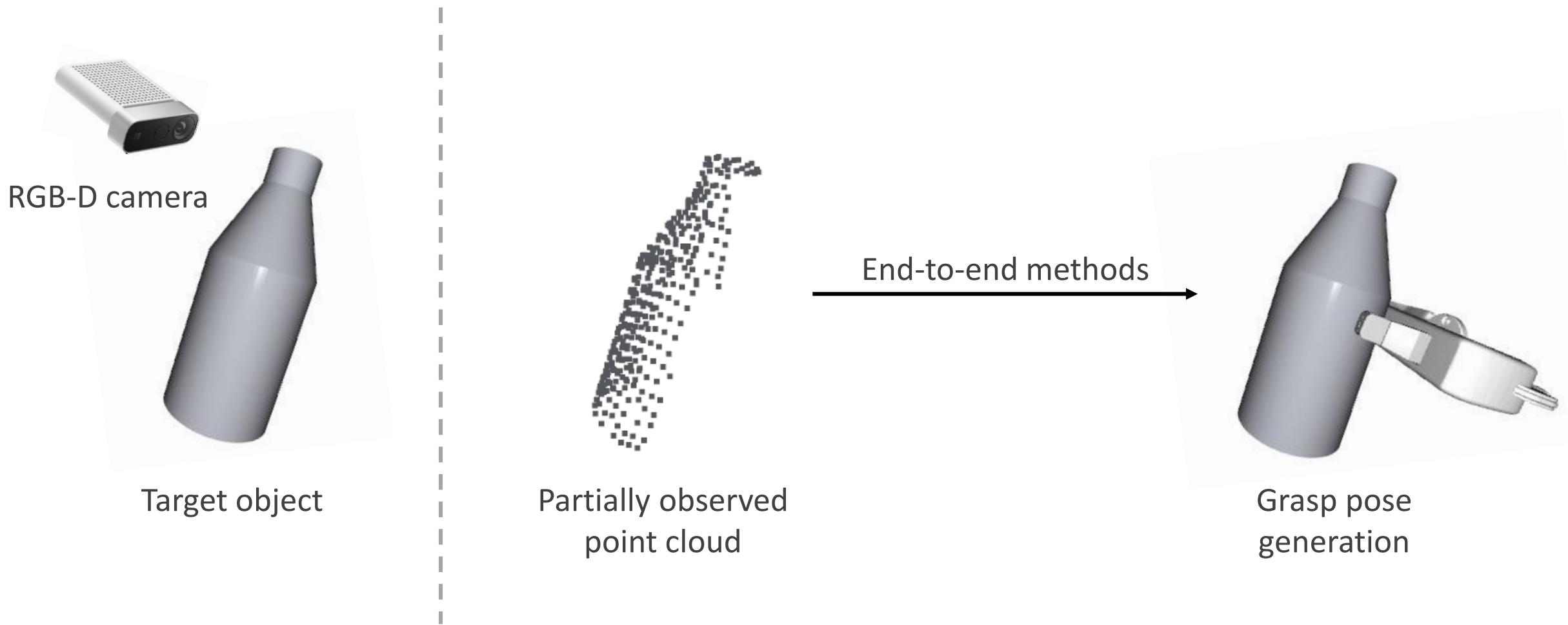
SQPDNet
(S. Kim, et al., CoRL'22)



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

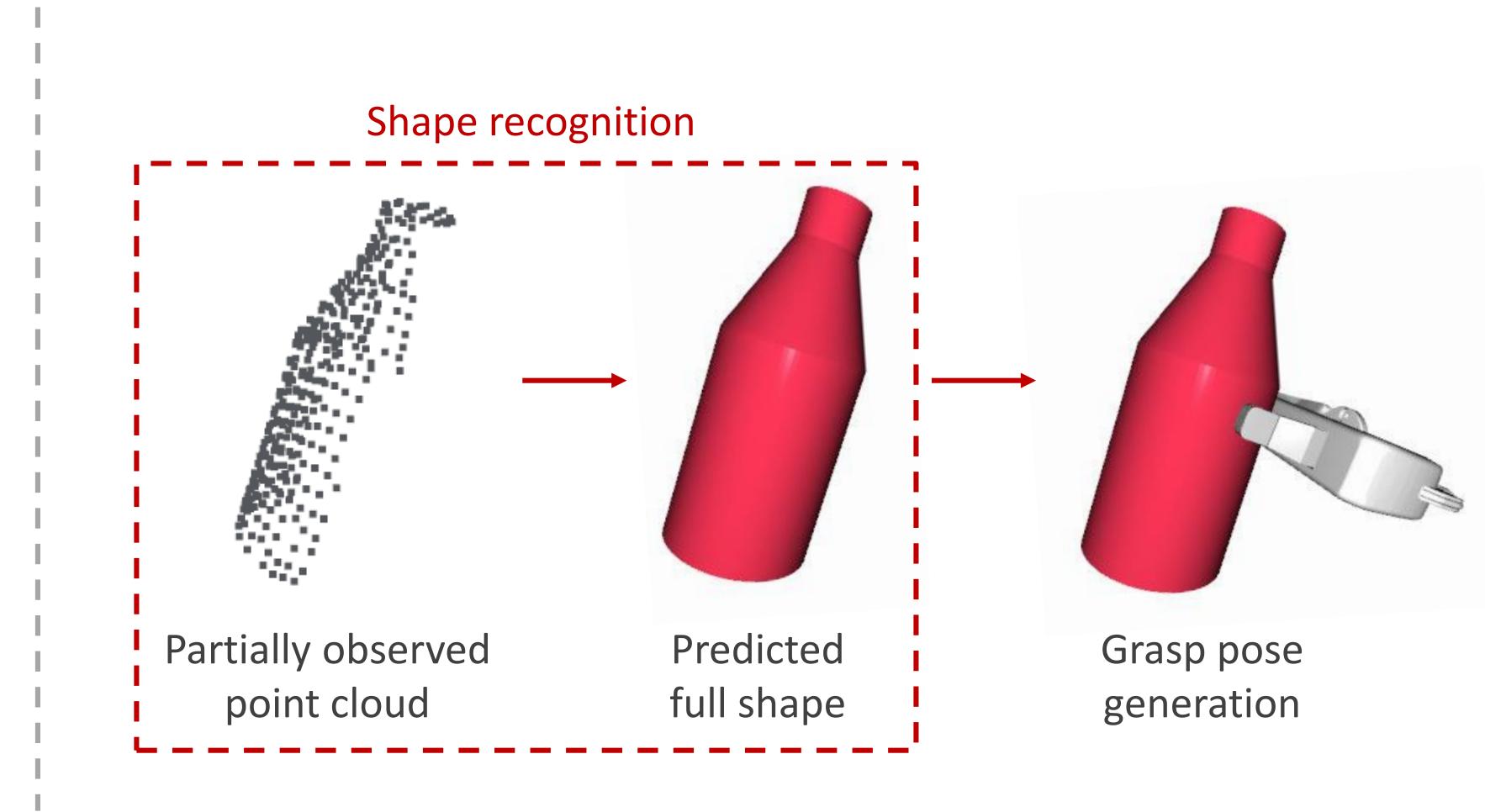


Vision-based Grasping



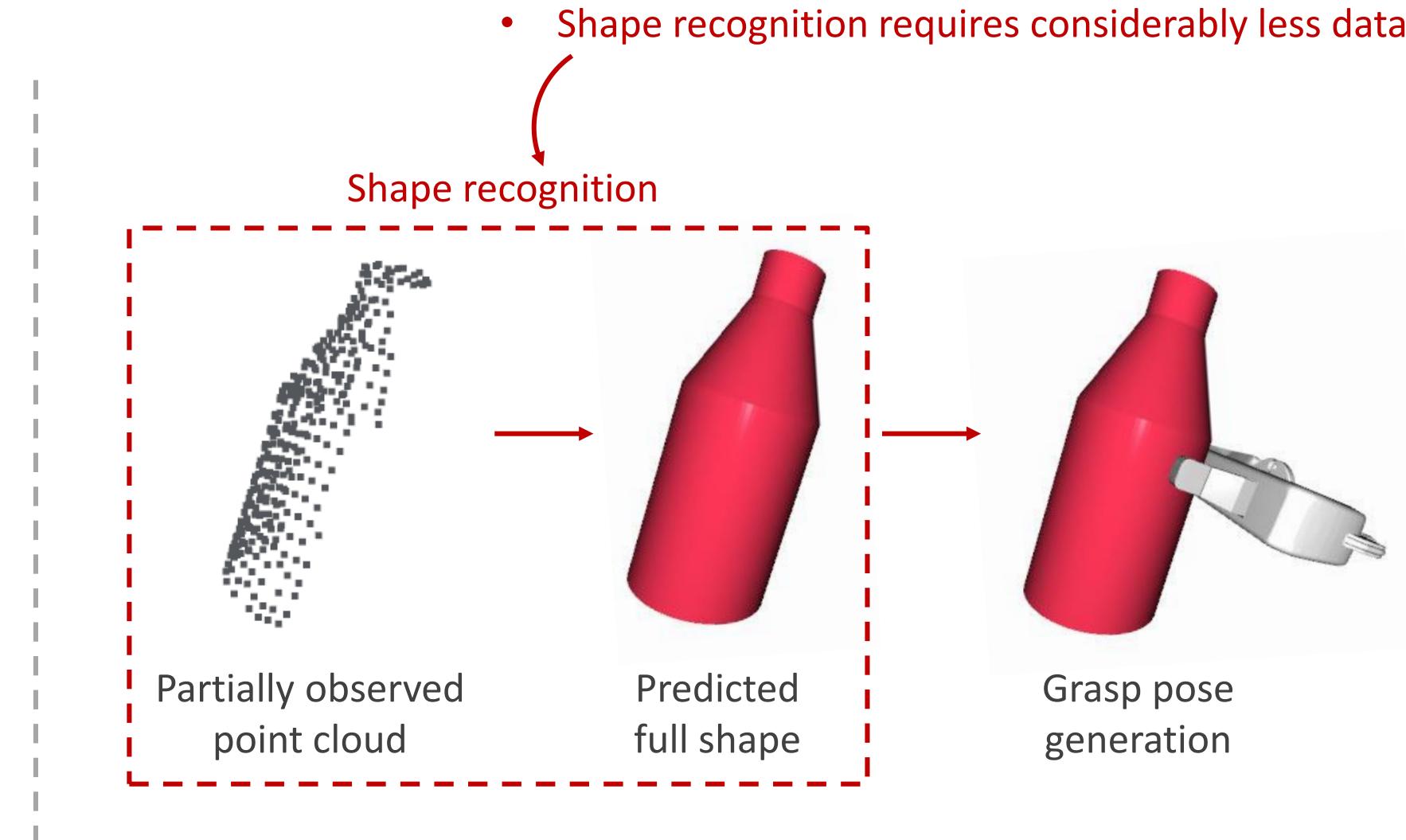


Shape Recognition-based Grasping





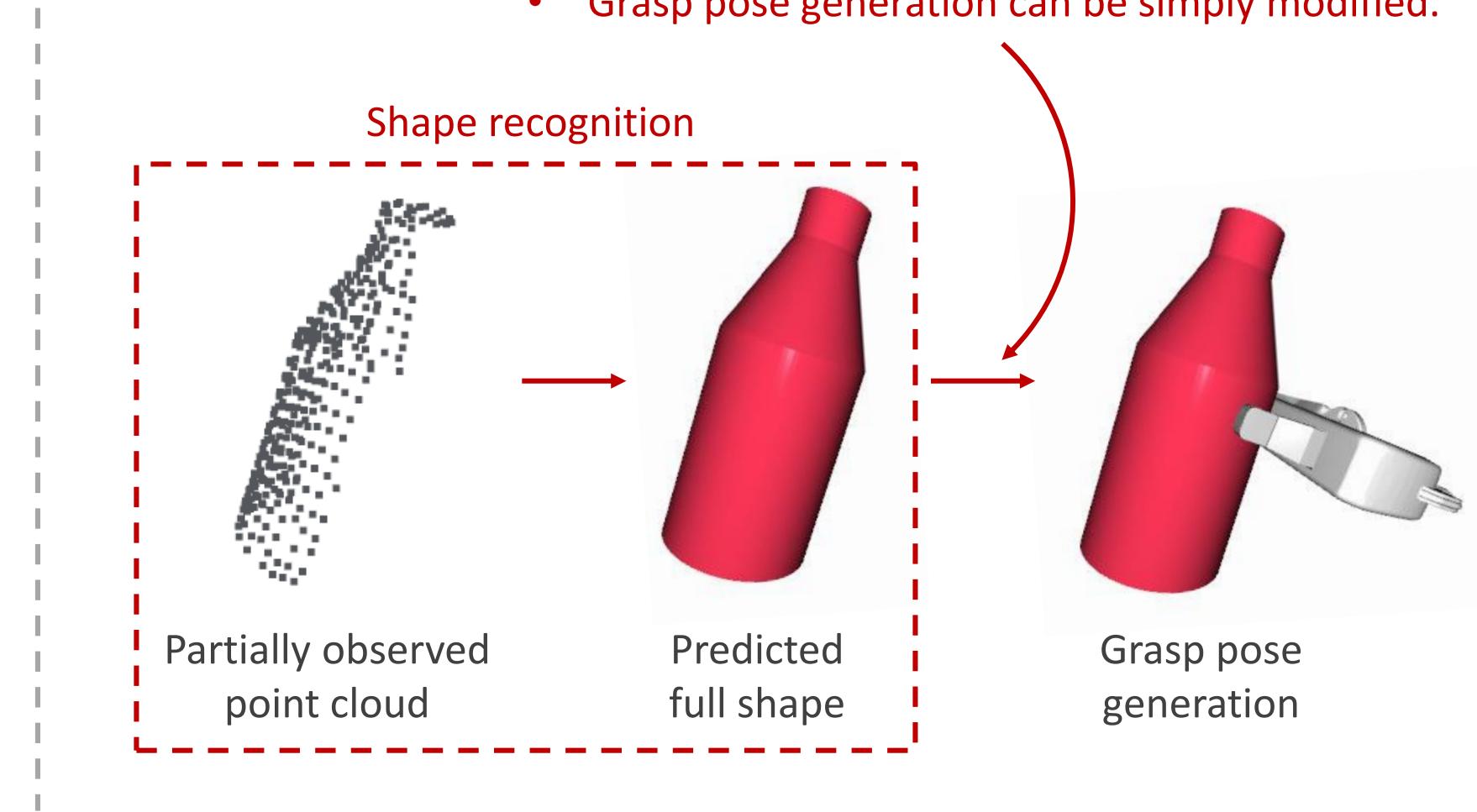
Shape Recognition-based Grasping





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



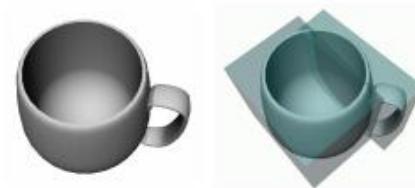
Bounding box

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



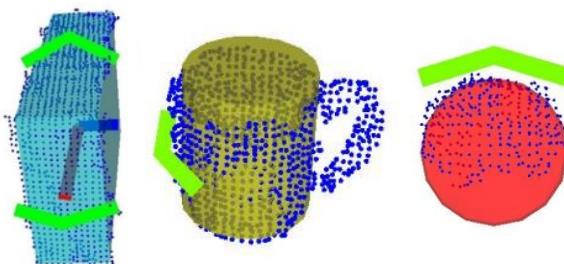
Shape Recognition-based Grasping

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

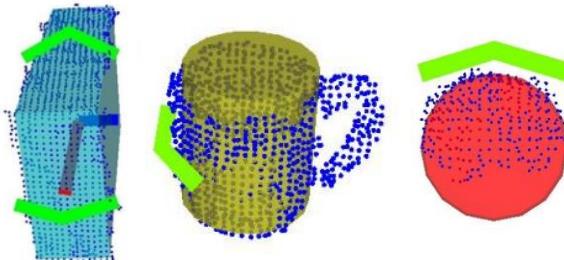


Bounding box

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

Superquadrics

(G. Vezzani, ICRA'17)



Box, cylinder, sphere

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



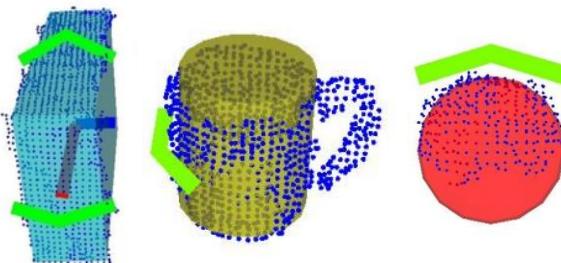
Shape Recognition-based Grasping

Shape expressiveness



Bounding box

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



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

Superquadrics

(G. Vezzani, ICRA'17)



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

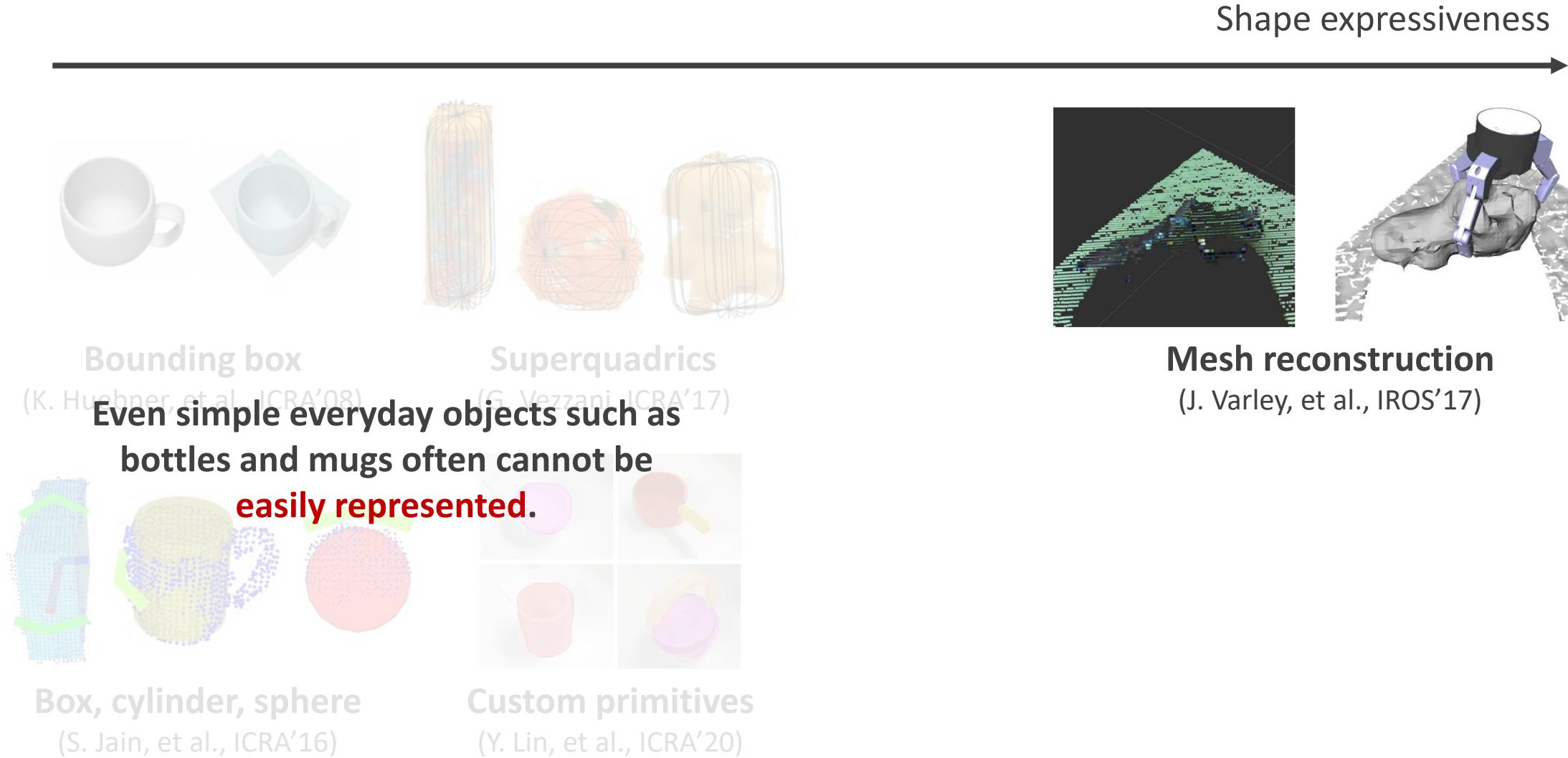


Shape Recognition-based Grasping



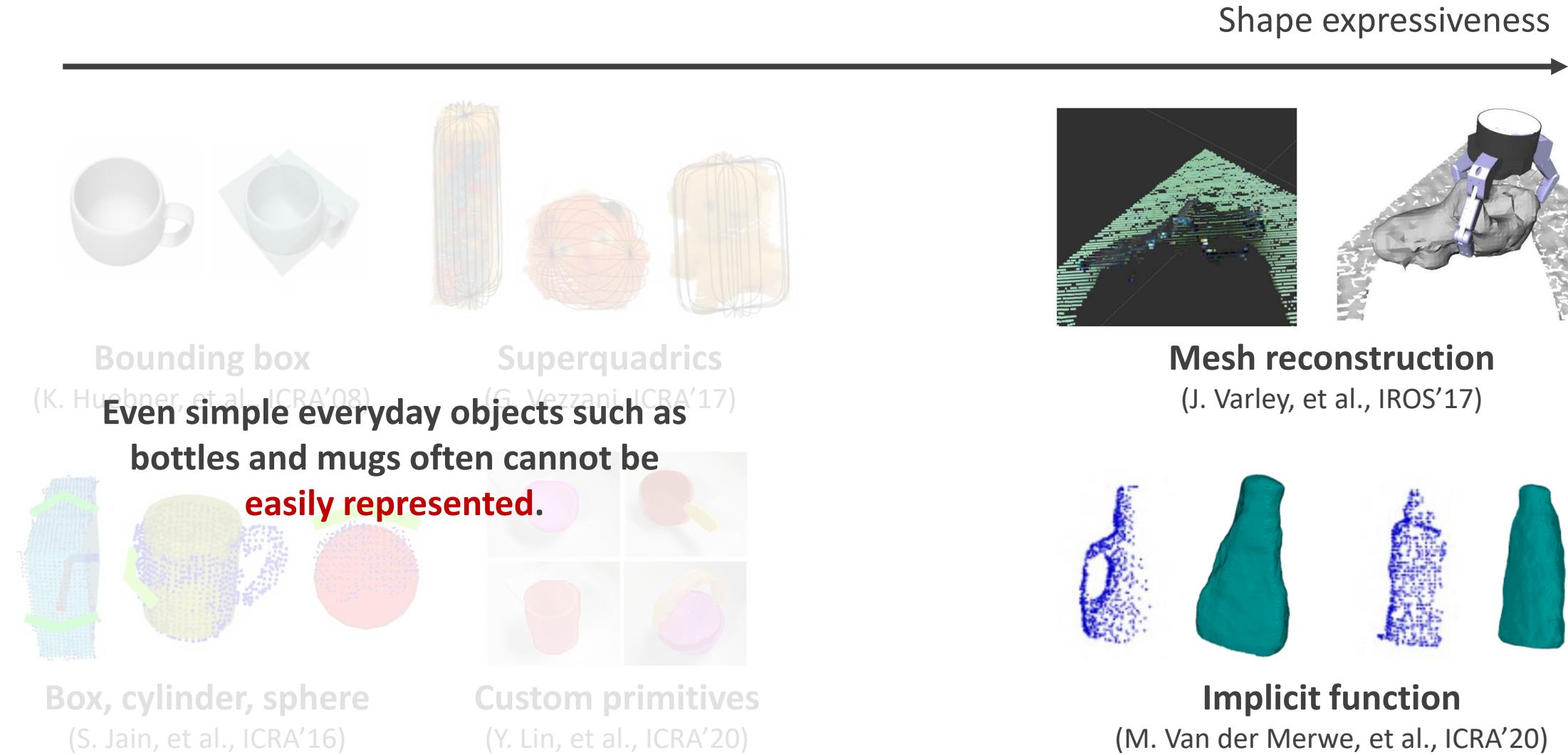


Shape Recognition-based Grasping



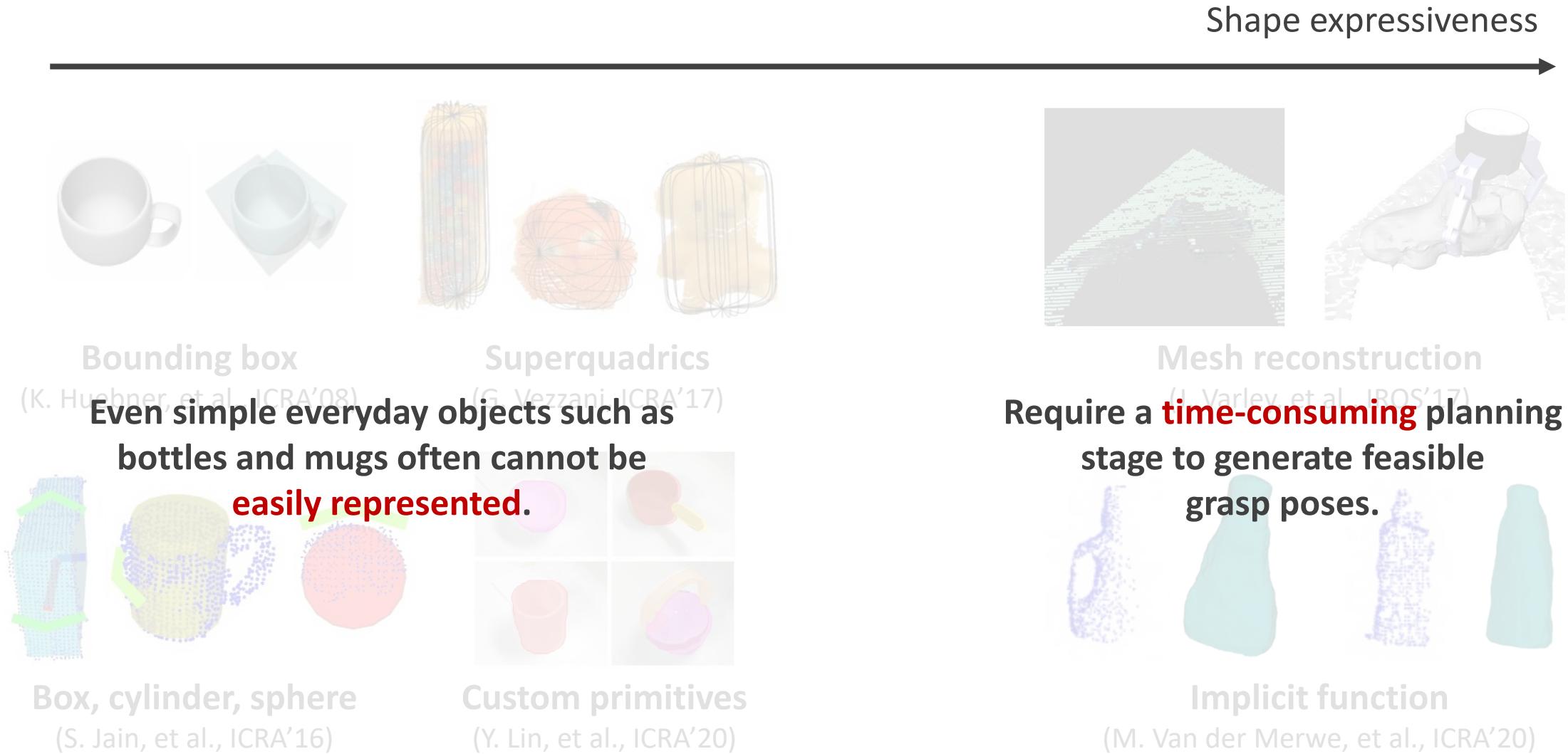


Shape Recognition-based Grasping



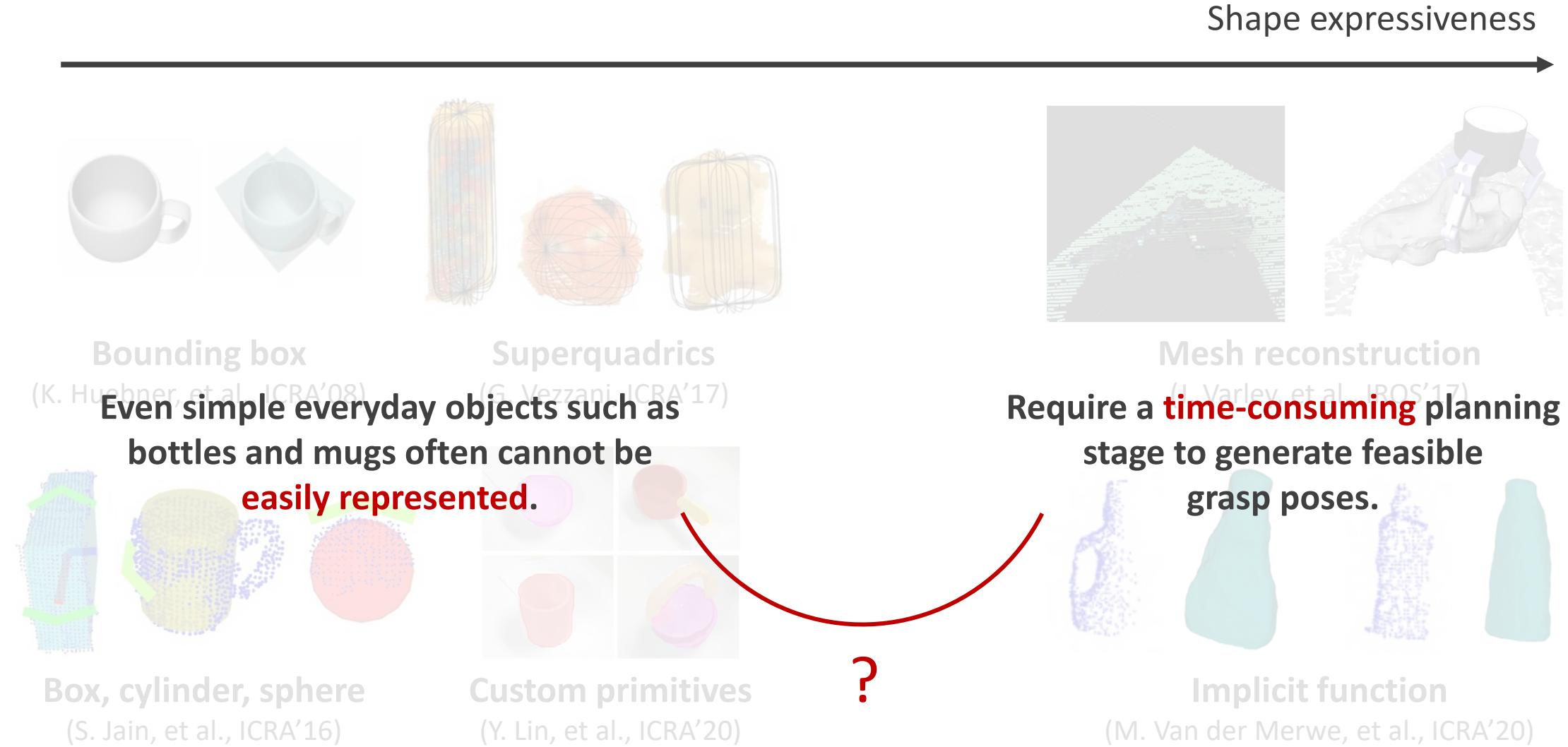


Shape Recognition-based Grasping





Shape Recognition-based Grasping





Deformable Superquadrics

Superquadrics

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

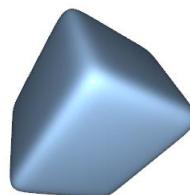


Deformable Superquadrics

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



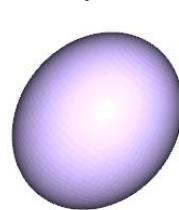
$e_1 = 0.2$

$e_2 = 0.2$



$e_1 = 0.2$

$e_2 = 1.0$



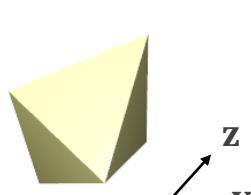
$e_1 = 1.0$

$e_2 = 1.0$



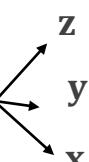
$e_1 = 2.0$

$e_2 = 0.2$



$e_1 = 2.0$

$e_2 = 1.0$



a = (a_1, a_2, a_3) : size parameters

e = (e_1, e_2) : shape parameters

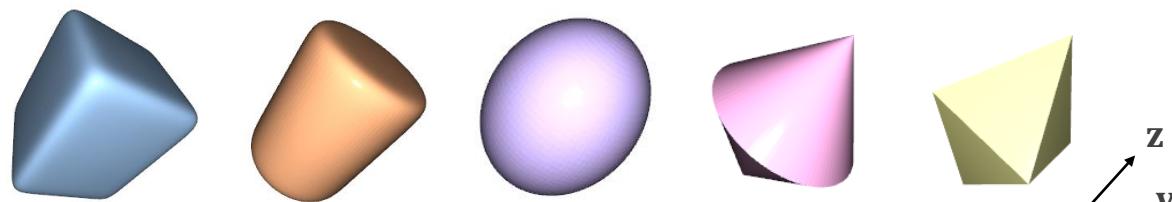


Deformable Superquadrics

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



$$e_1 = 0.2$$

$$e_2 = 0.2$$

$$e_1 = 0.2$$

$$e_2 = 1.0$$

$$e_1 = 1.0$$

$$e_2 = 1.0$$

$$e_1 = 2.0$$

$$e_2 = 0.2$$

$$e_1 = 2.0$$

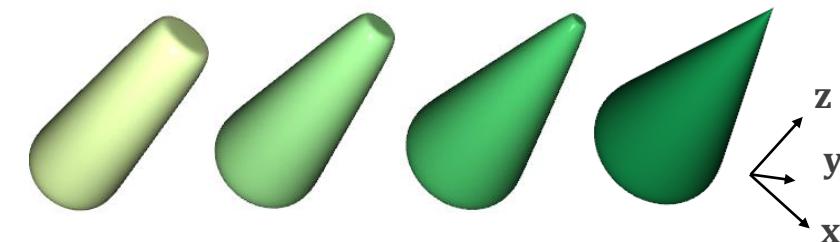
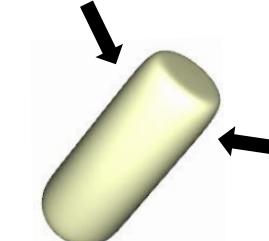
$$e_2 = 1.0$$

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

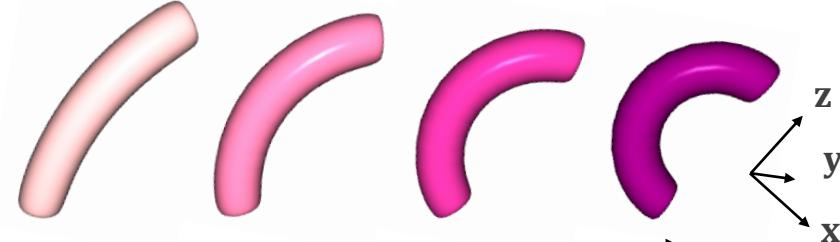
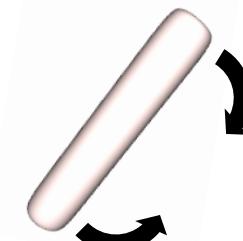
$\mathbf{e} = (e_1, e_2)$: shape parameters

Deformable Superquadrics

Tapering



Bending



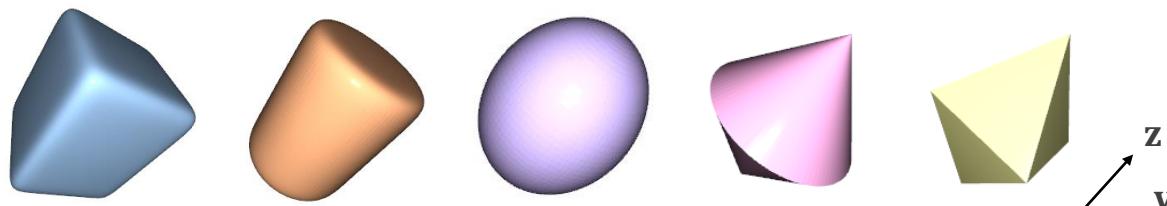
Deformable Superquadrics



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



$$e_1 = 0.2$$

$$e_2 = 0.2$$

$$e_1 = 0.2$$

$$e_2 = 1.0$$

$$e_1 = 1.0$$

$$e_2 = 1.0$$

$$e_1 = 2.0$$

$$e_2 = 0.2$$

$$e_1 = 2.0$$

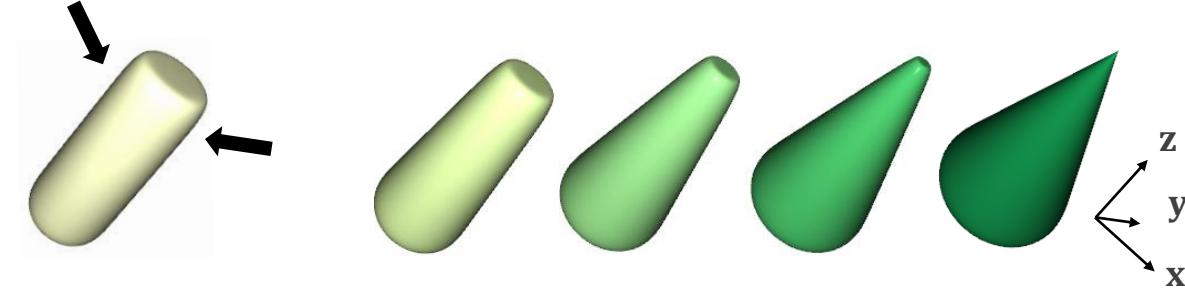
$$e_2 = 1.0$$

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

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

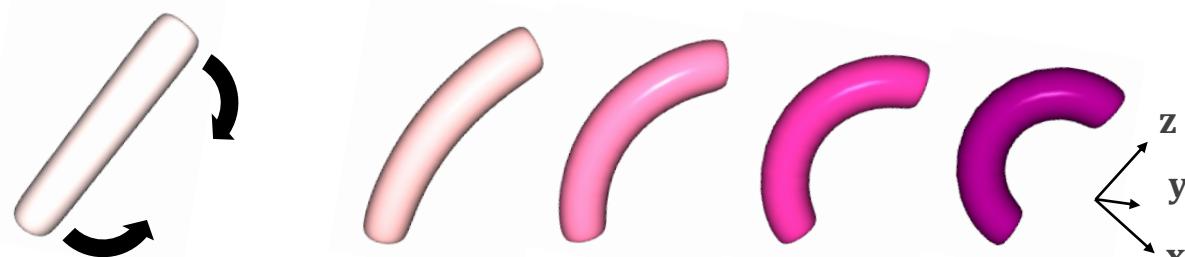
Deformable Superquadrics

Tapering with **tapering parameter k**



$|k|$ increase

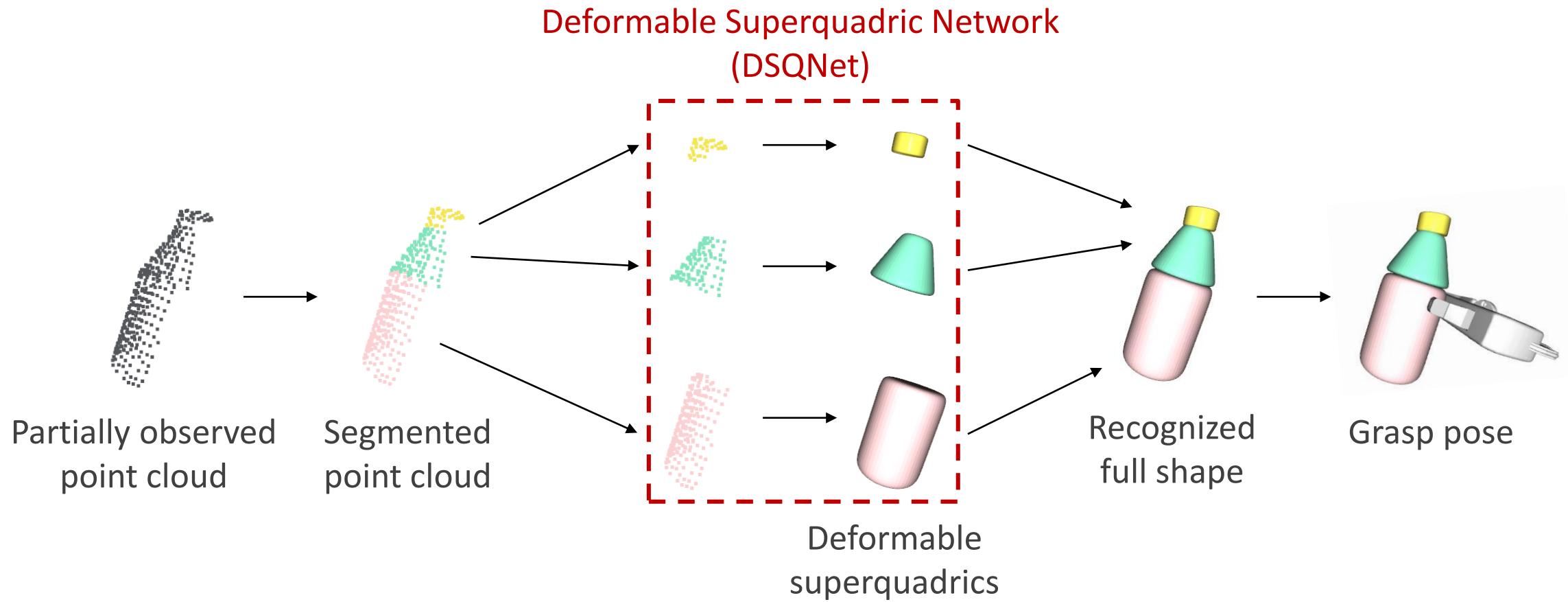
Bending with **bending parameters (b, α)**



b increase

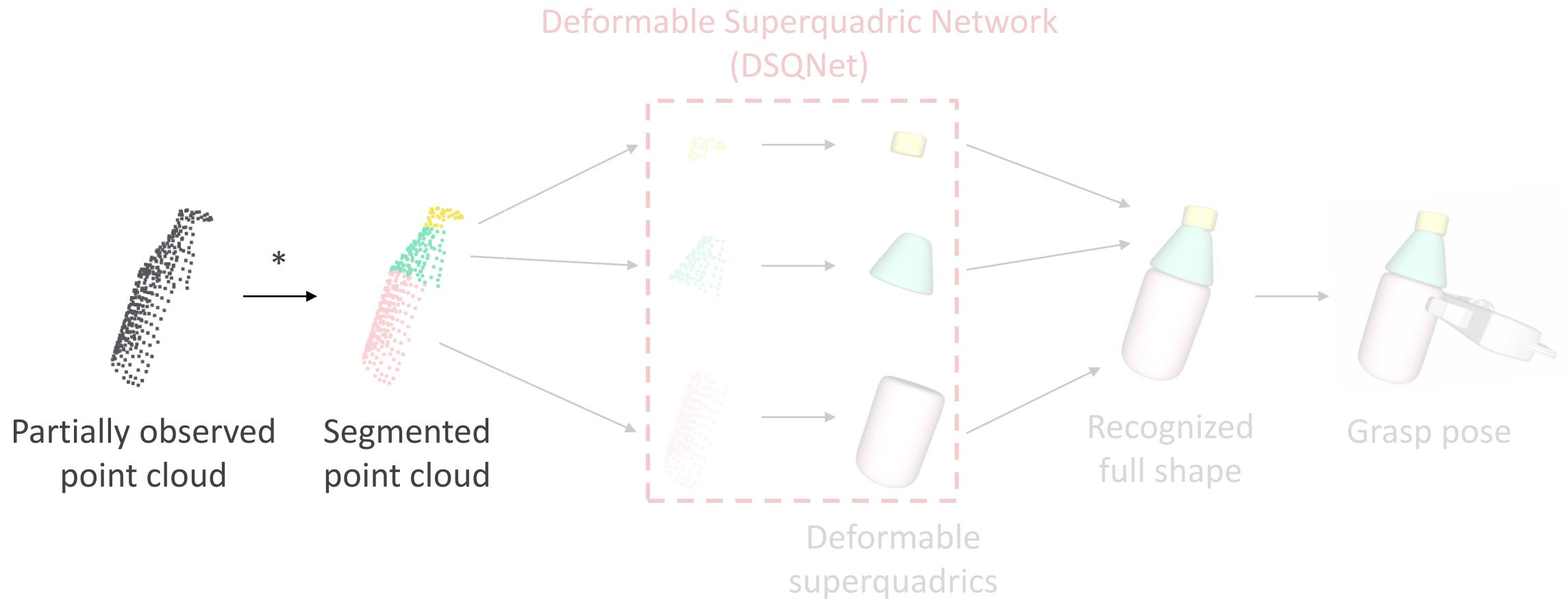


Our Method



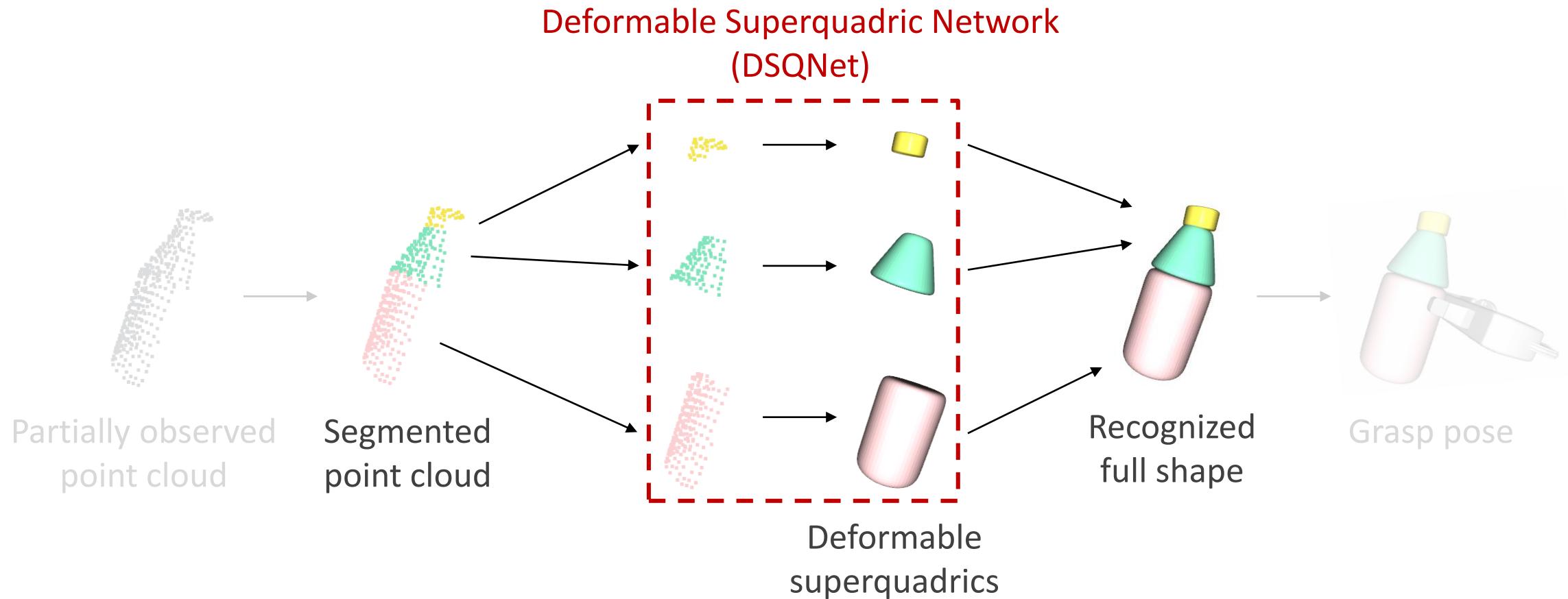


Our Method



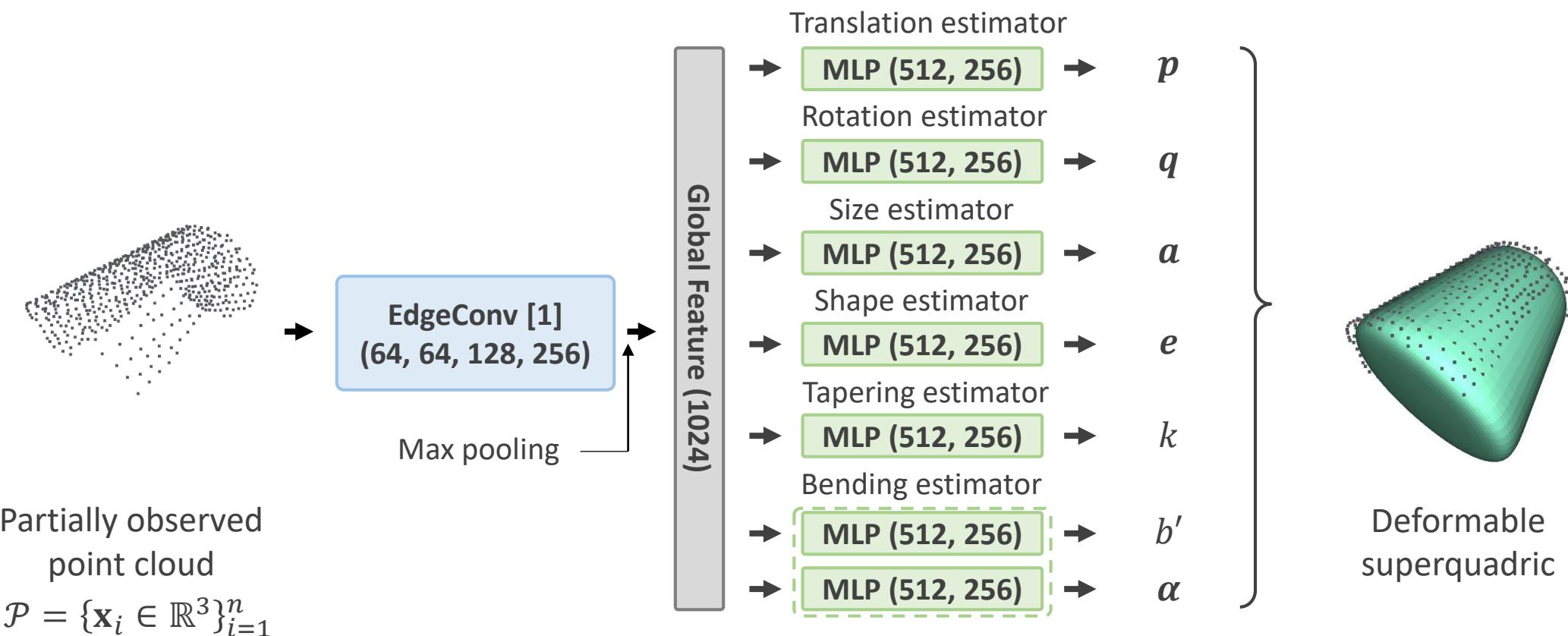


Our Method

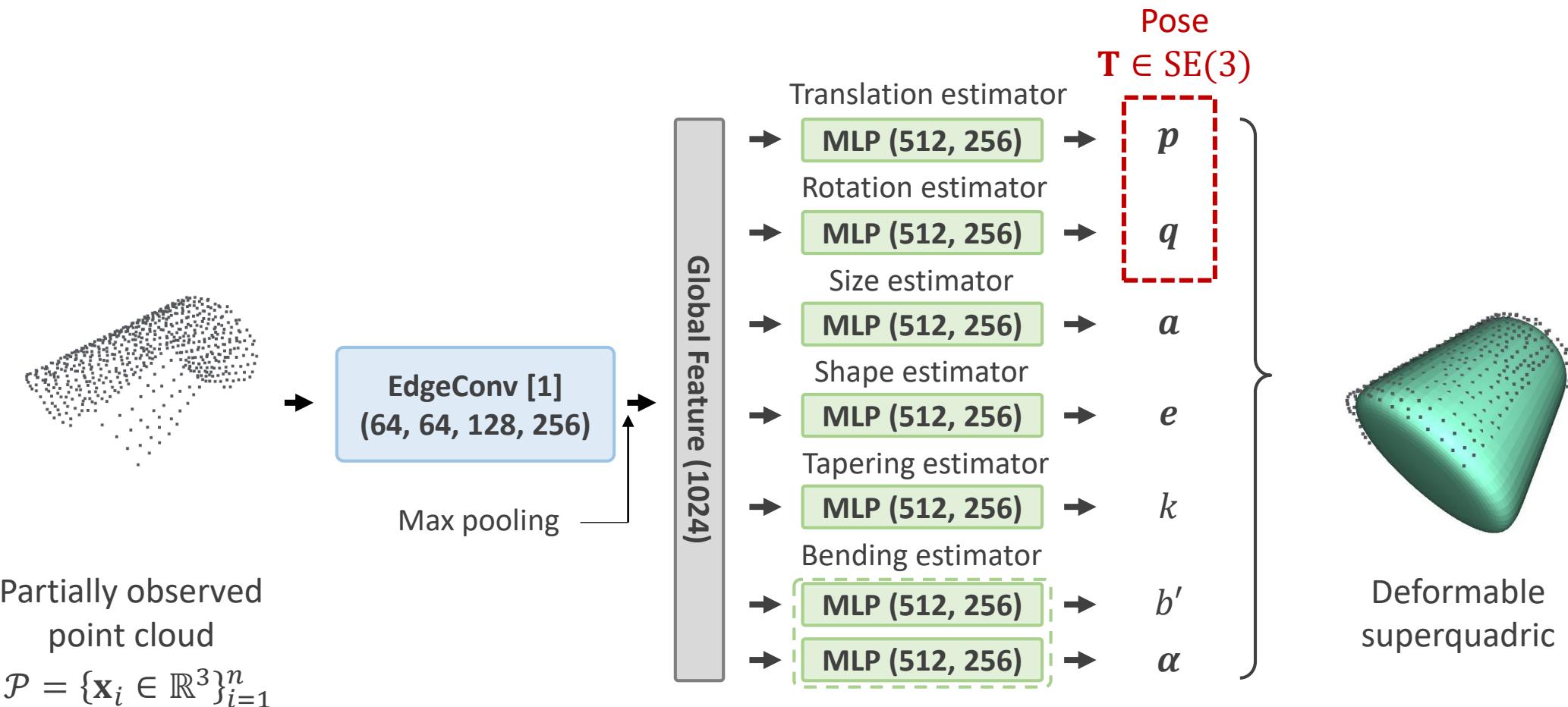




Deformable Superquadric Network

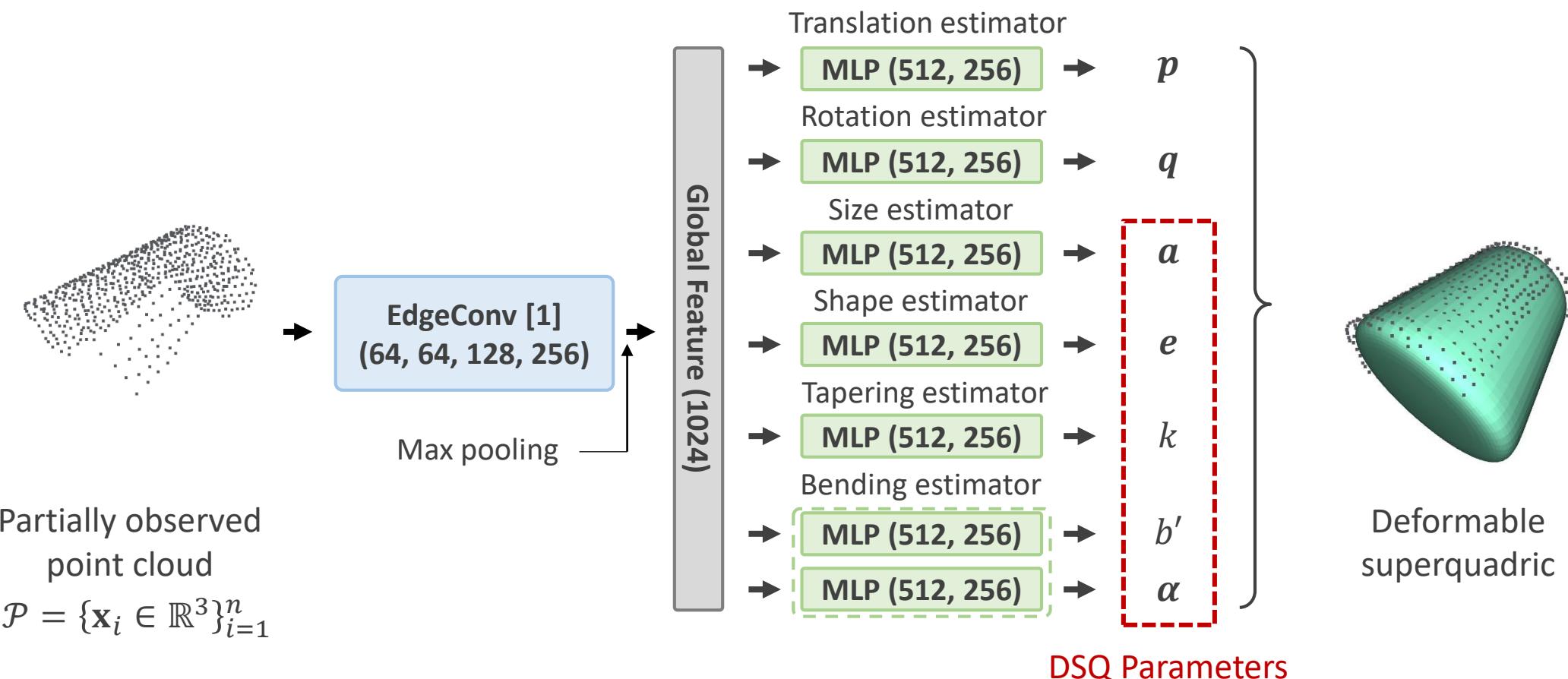


Deformable Superquadric Network





Deformable Superquadric Network



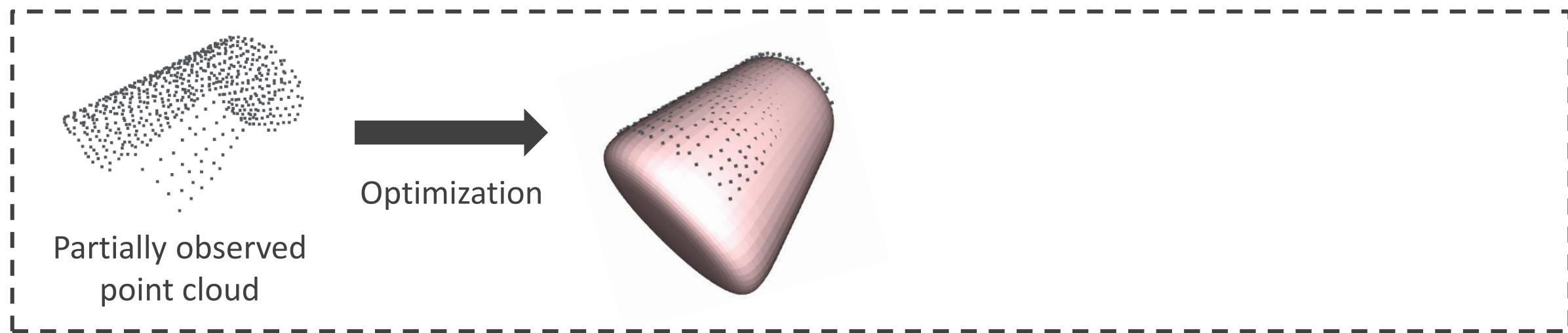


Deformable Superquadric Network



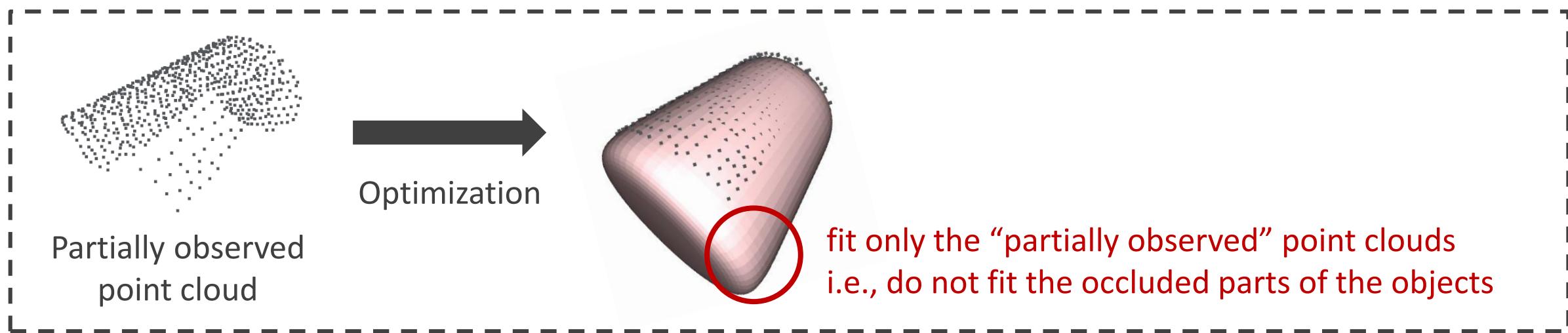


Deformable Superquadric Network



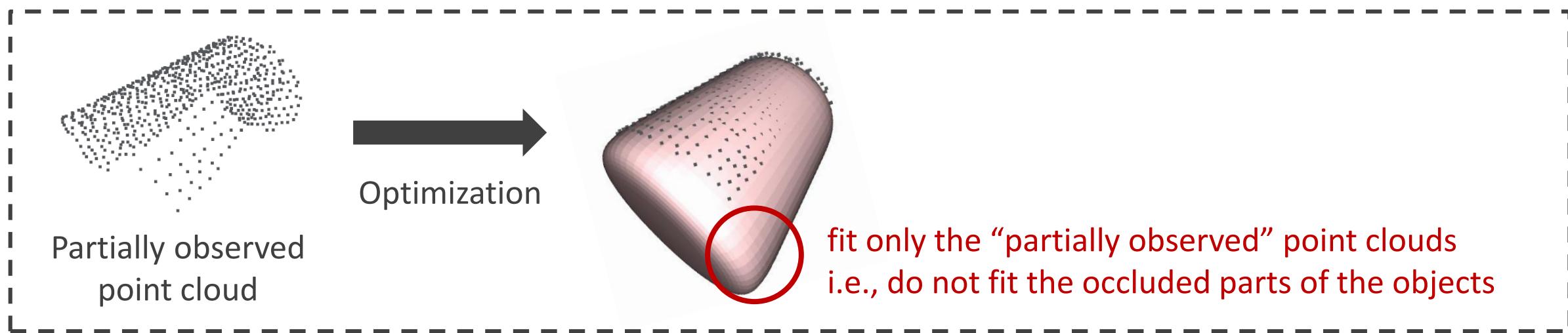


Deformable Superquadric Network

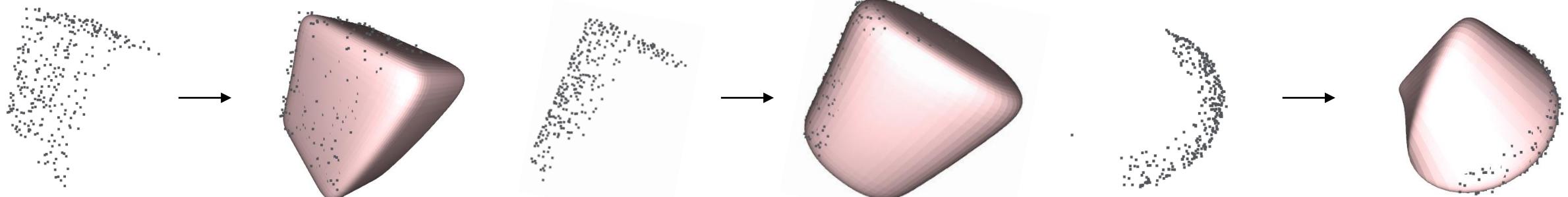




Deformable Superquadric Network

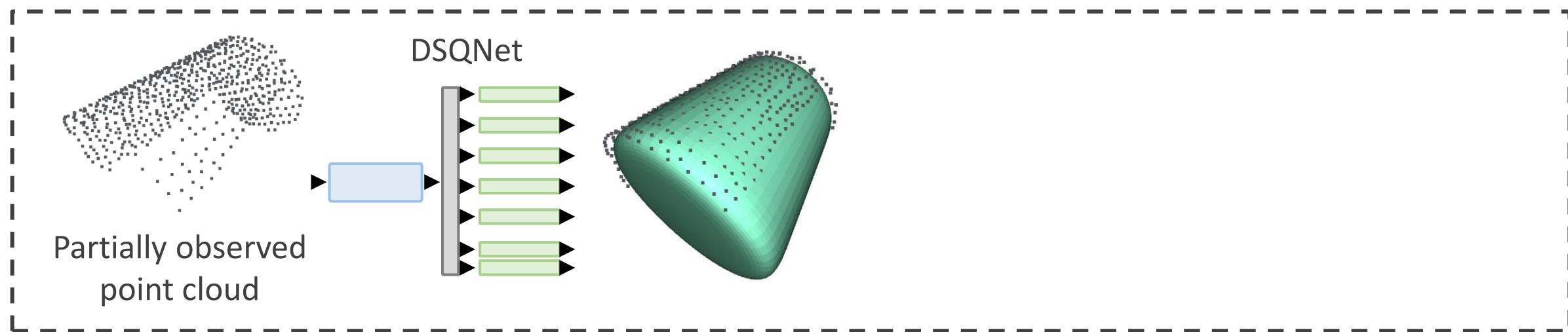


Examples of **optimization-based method's results**



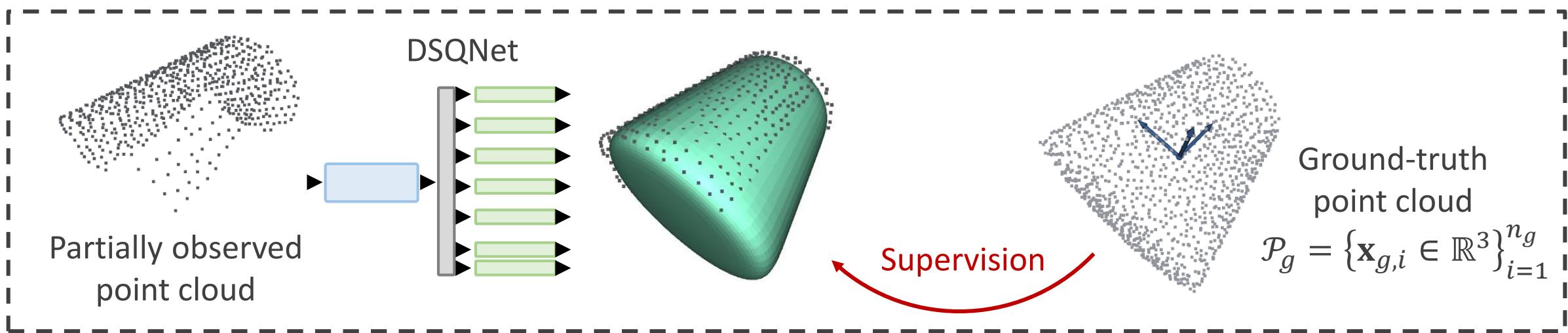


Deformable Superquadric Network



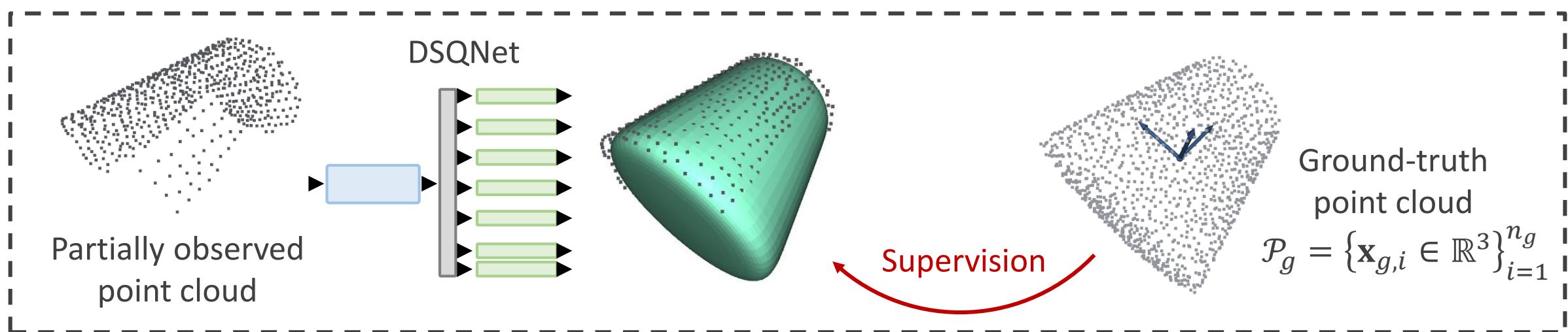


Deformable Superquadric Network

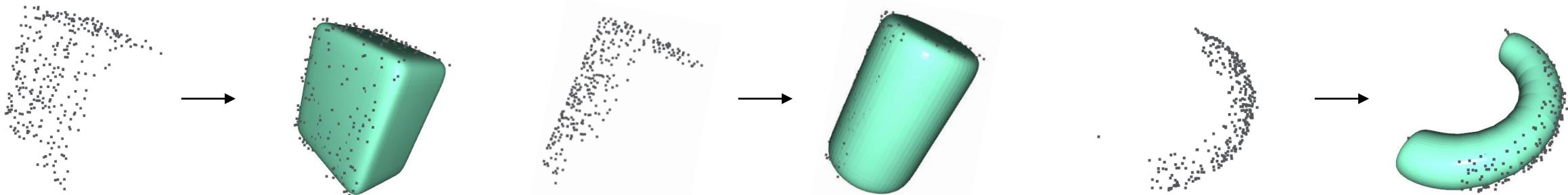




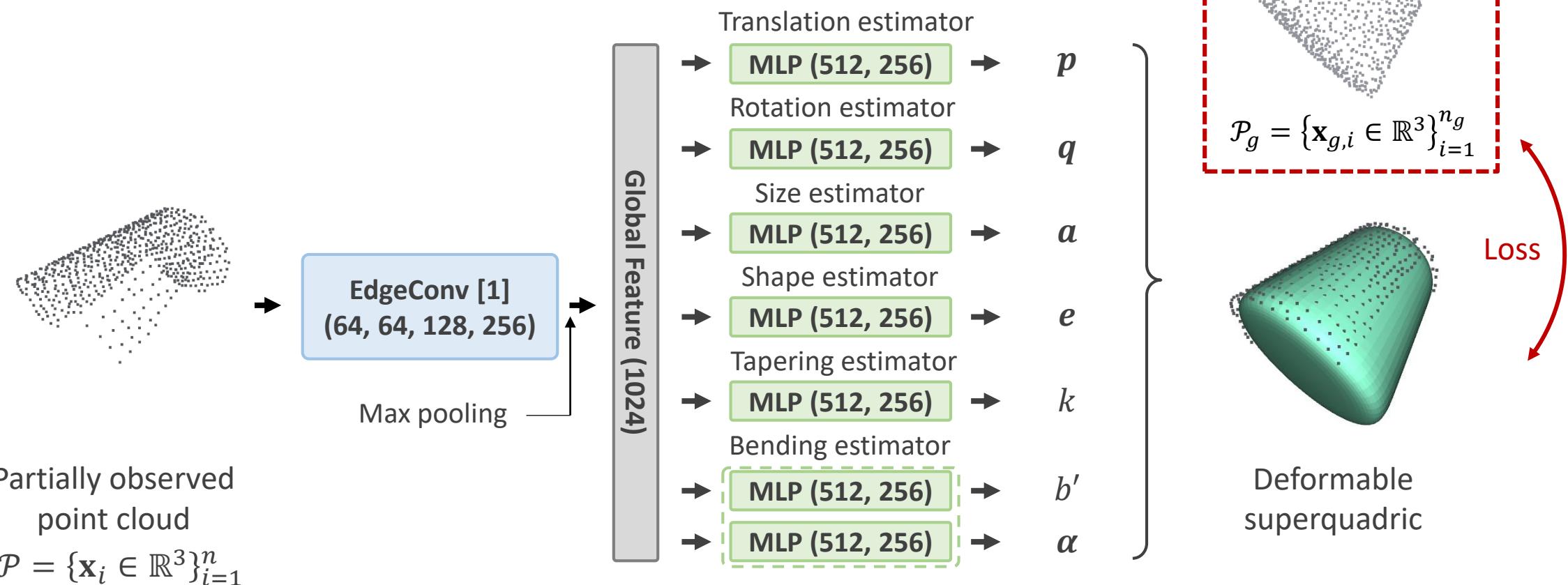
Deformable Superquadric Network



Examples of **DSQNet's** results

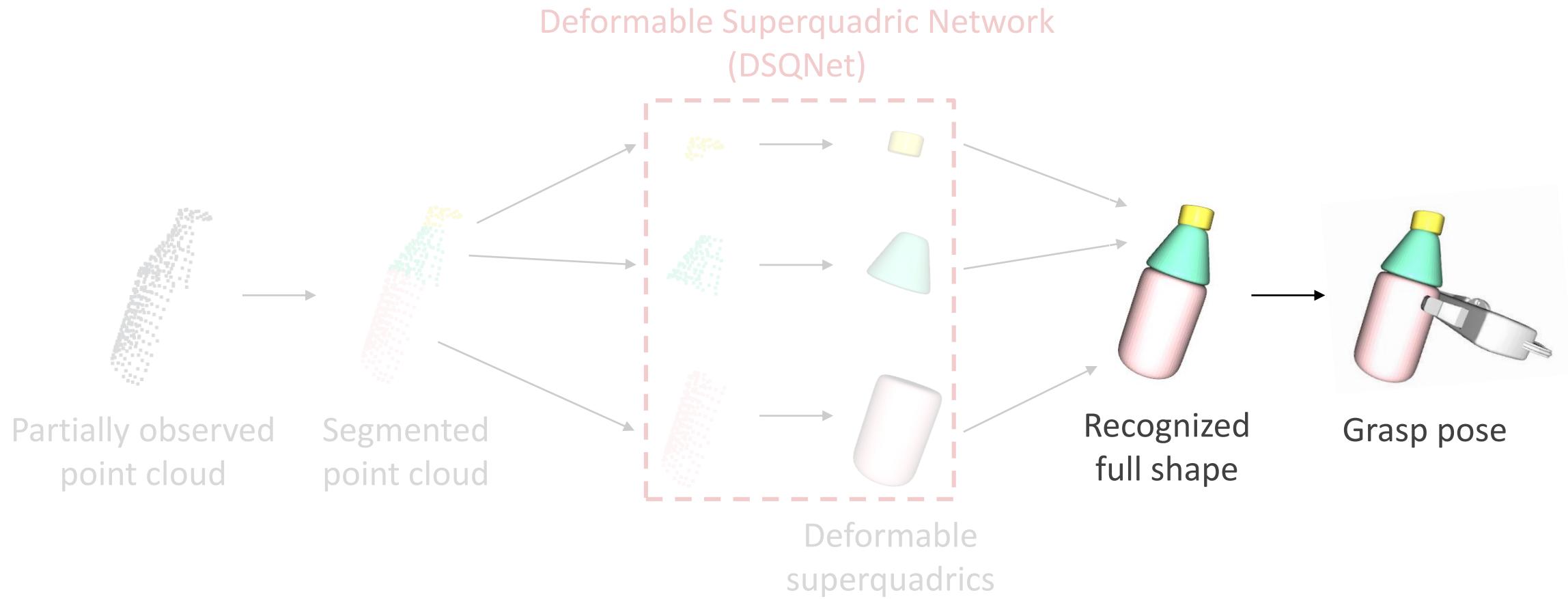


Deformable Superquadric Network



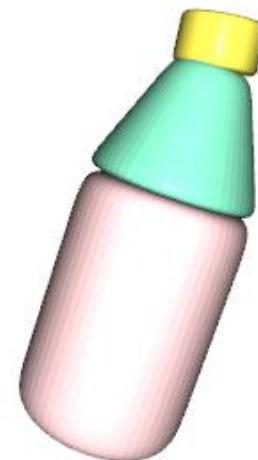


Our Method



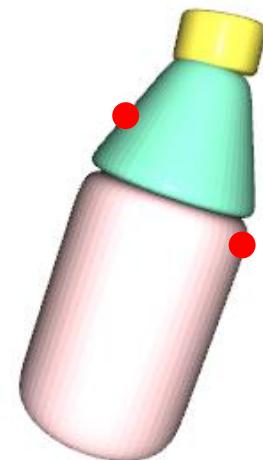


Grasp Pose Generation





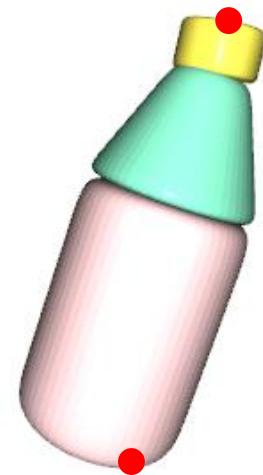
Grasp Pose Generation



Sample antipodal points



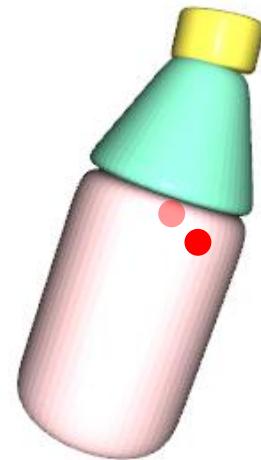
Grasp Pose Generation



Sample antipodal points



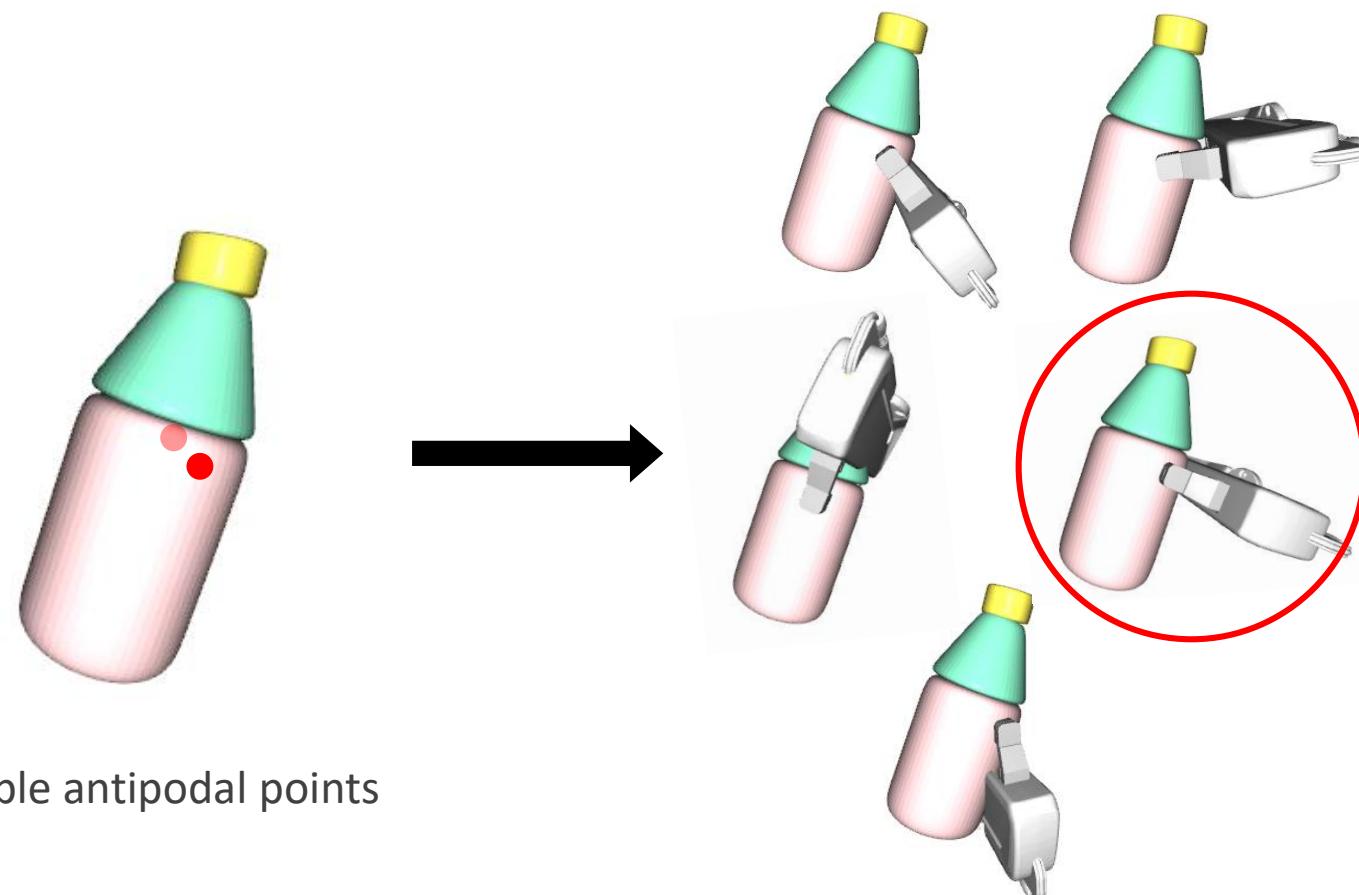
Grasp Pose Generation



Sample antipodal points



Grasp Pose Generation

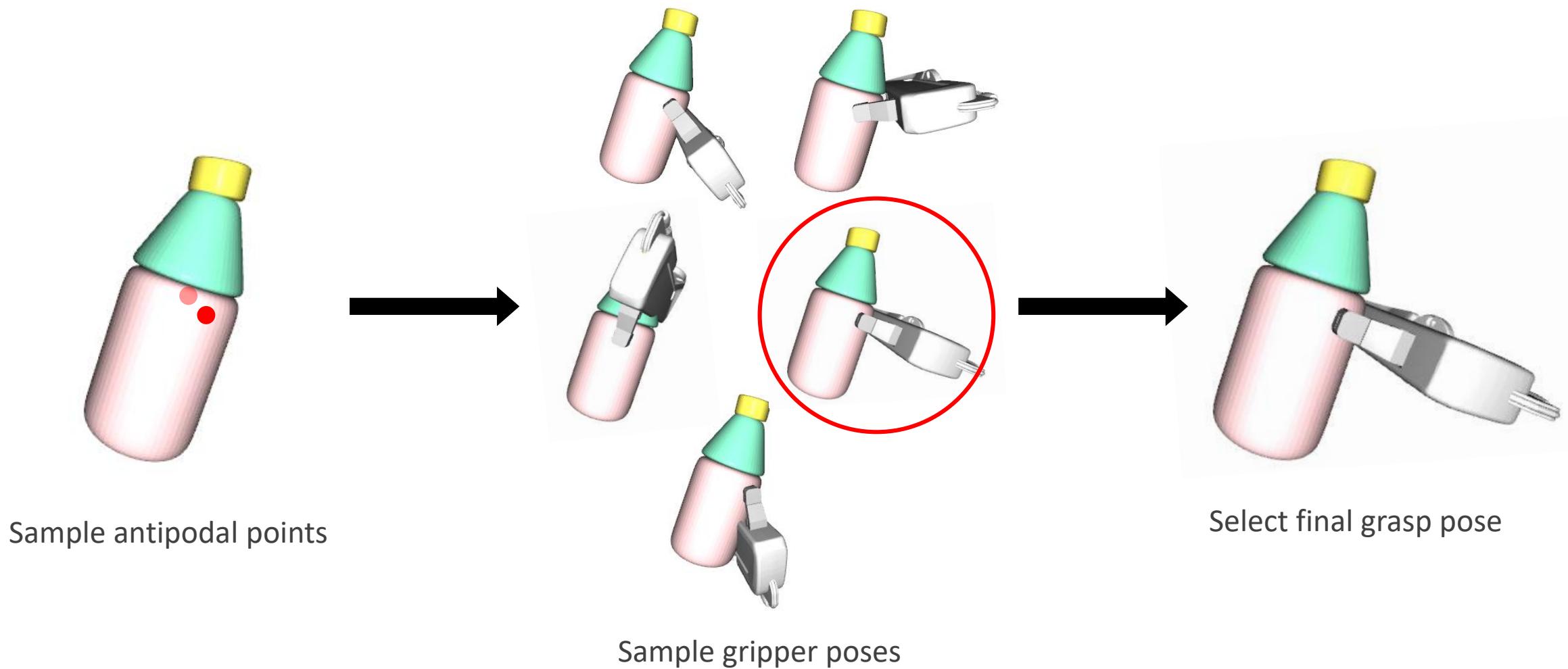


Sample antipodal points

Sample gripper poses



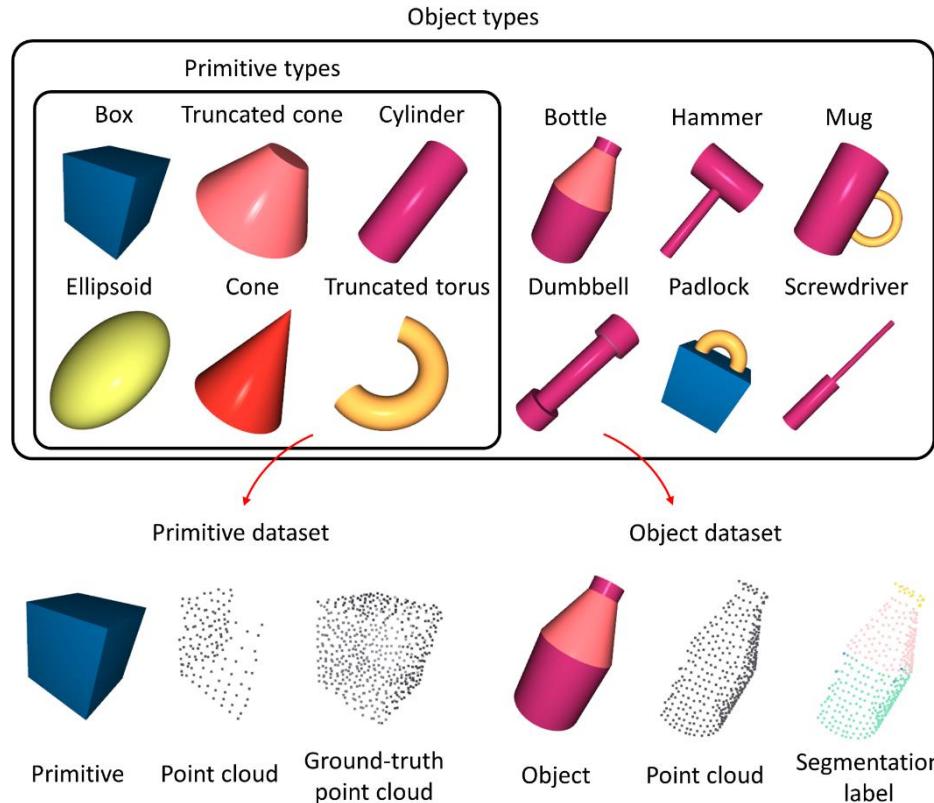
Grasp Pose Generation



Experimental Results



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



Real-world objects



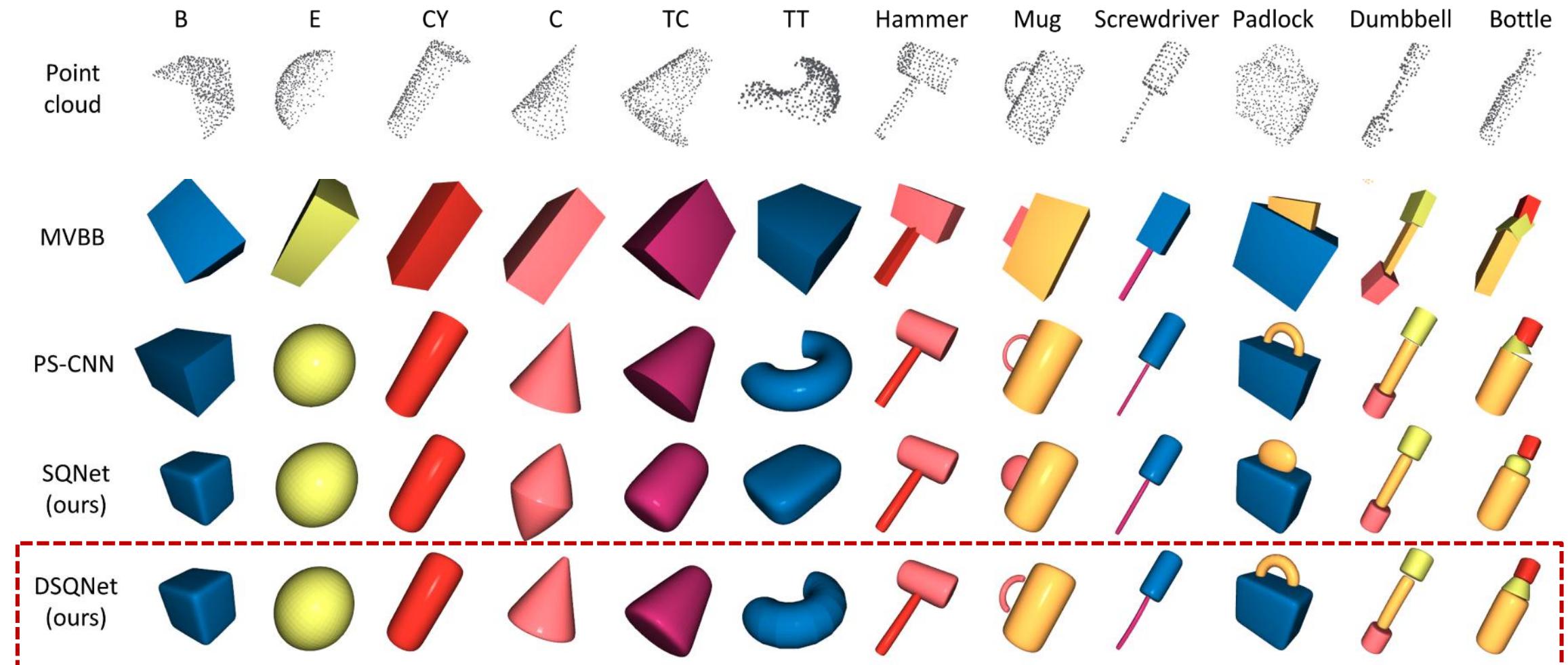


Experimental Results

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

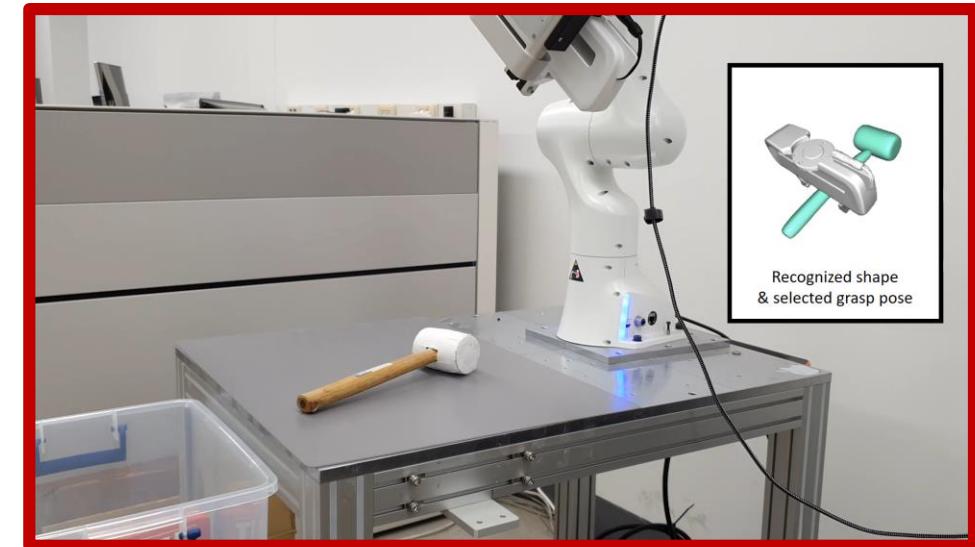


Grasping single-part shapes



- Recognize the object using single deformable superquadric.

Grasping multi-part shapes



- Recognize the object using multiple deformable superquadrics.



Grasping Dinnerware Objects



Bowl



Dish



Mug



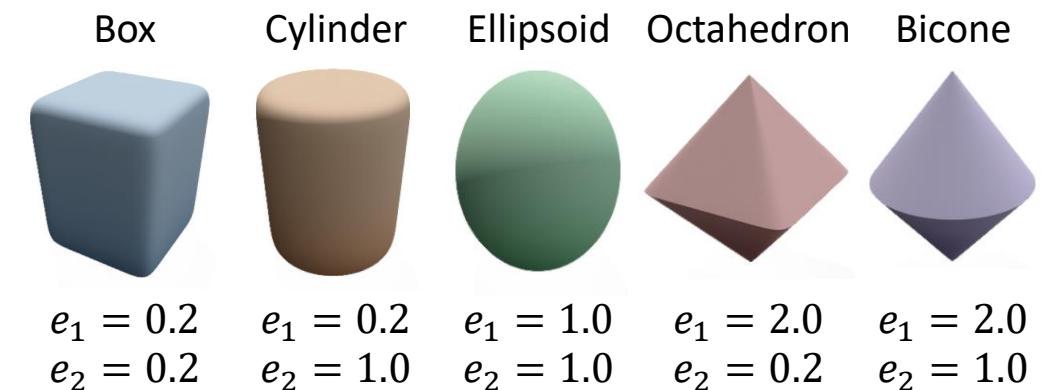
Spoon



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

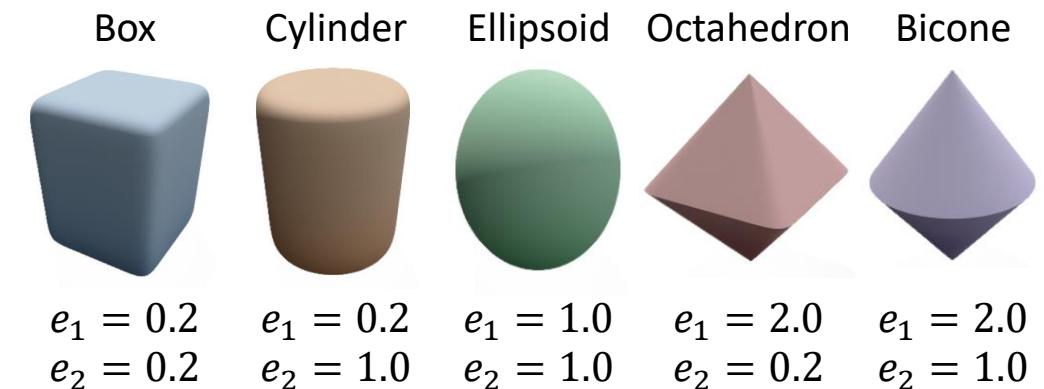
$$f(x, y, z) = \left(\left| \frac{x}{a_1} \right|^{2/e_2} + \left| \frac{y}{a_2} \right|^{2/e_2} \right)^{e_2/e_1} + \left| \frac{z}{a_3} \right|^{2/e_1} = 1$$

Superparaboloid

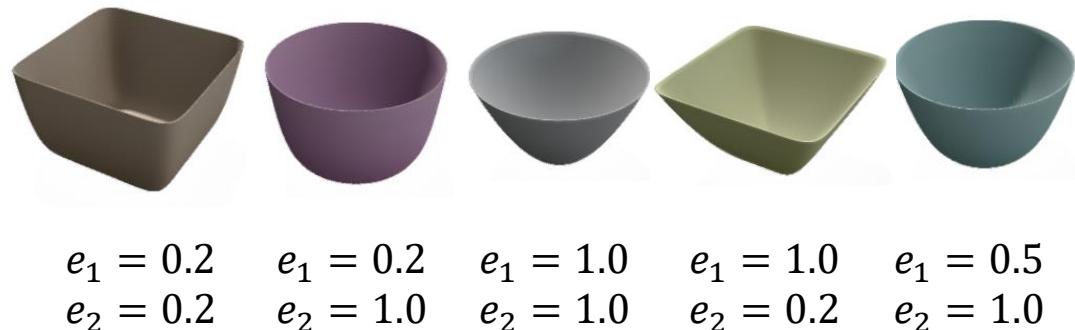
$$f(x, y, z) = \left(\left| \frac{x}{a_1} \right|^{2/e_2} + \left| \frac{y}{a_2} \right|^{2/e_2} \right)^{e_2/e_1} - \left(\frac{z}{a_3} \right) = 1$$

a = (a_1, a_2, a_3) : size parameters

e = (e_1, e_2) : shape parameters



Dishes and bowls





Experimental Results

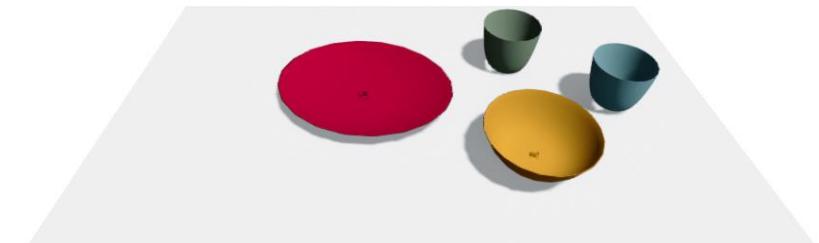
Scene



Partial observation



3D shape recognition



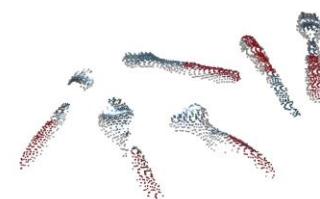


Experimental Results

Scene



Partial observation



3D shape recognition



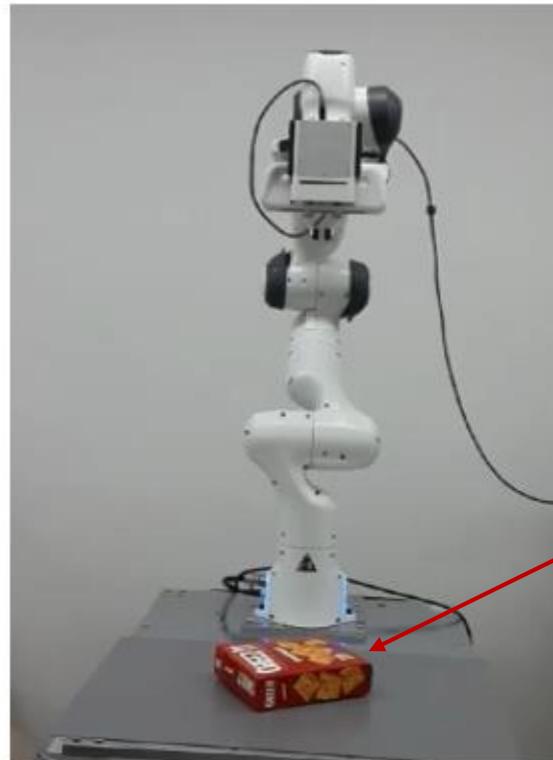


Robot Grasping



- Success rate = 92% (92/100)

Needs for Non-prehensile Manipulation



- Too large to grasp



- Too cluttered environment



Needs for Non-prehensile Manipulation



- Move the target object



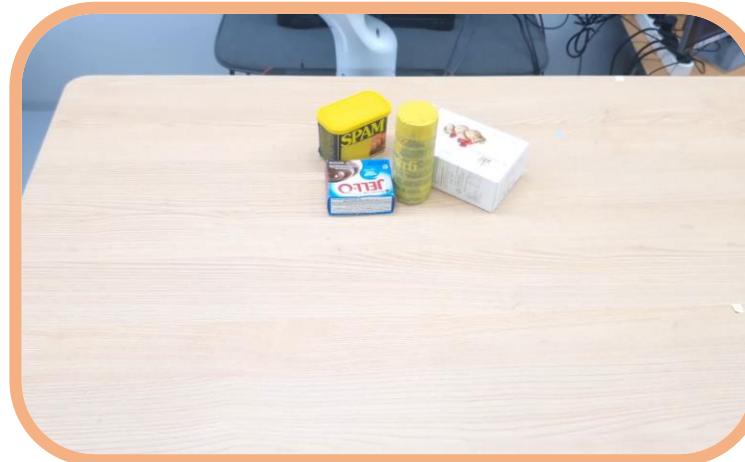
- Singulate the target object



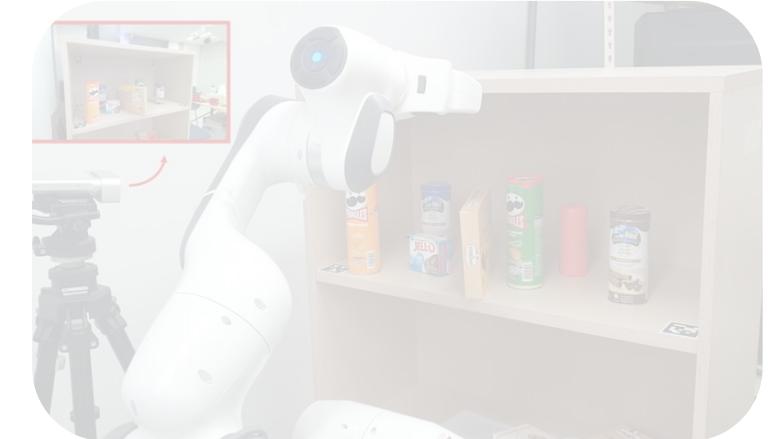
Shape Recognition-based Approaches



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



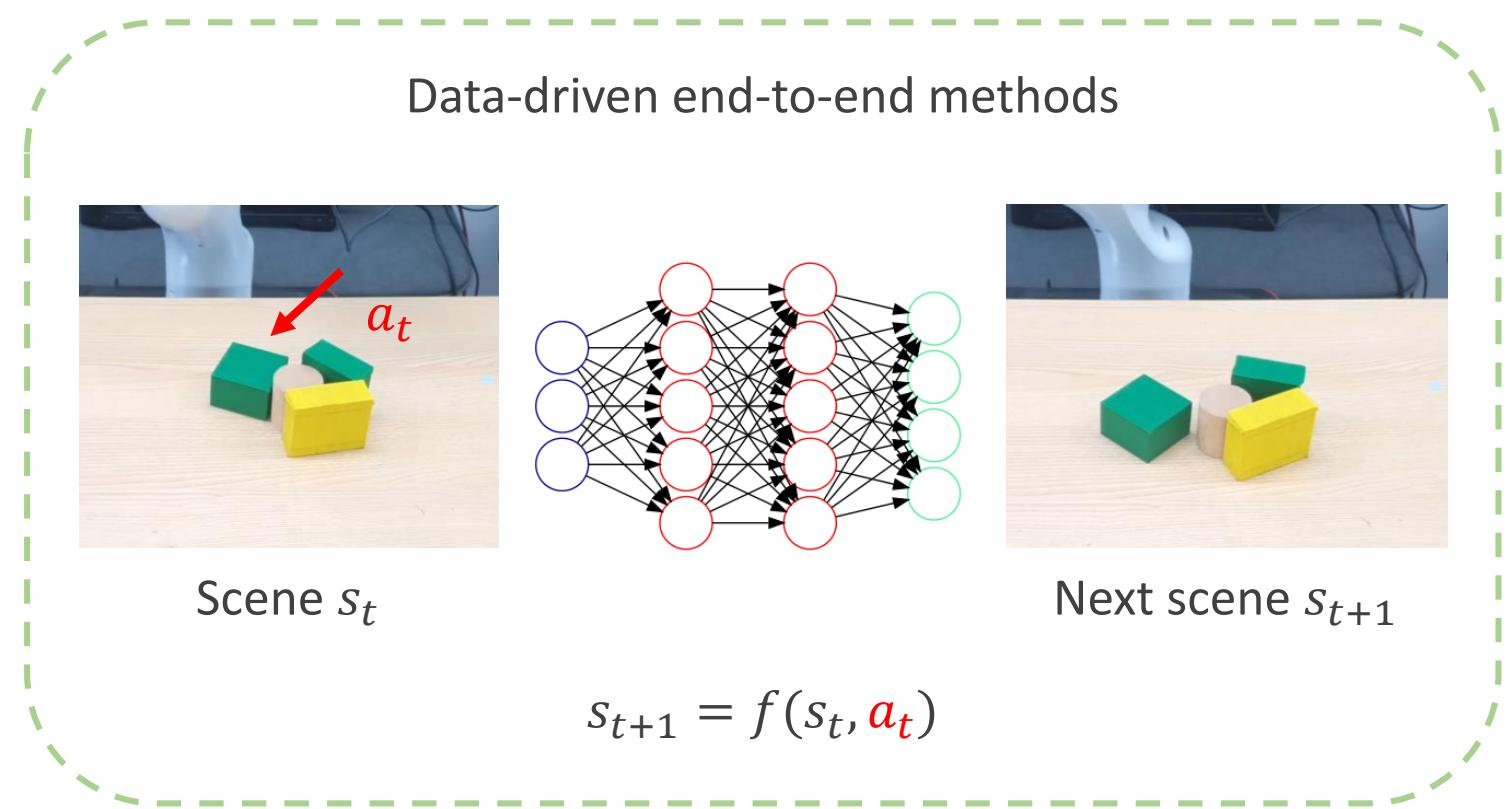
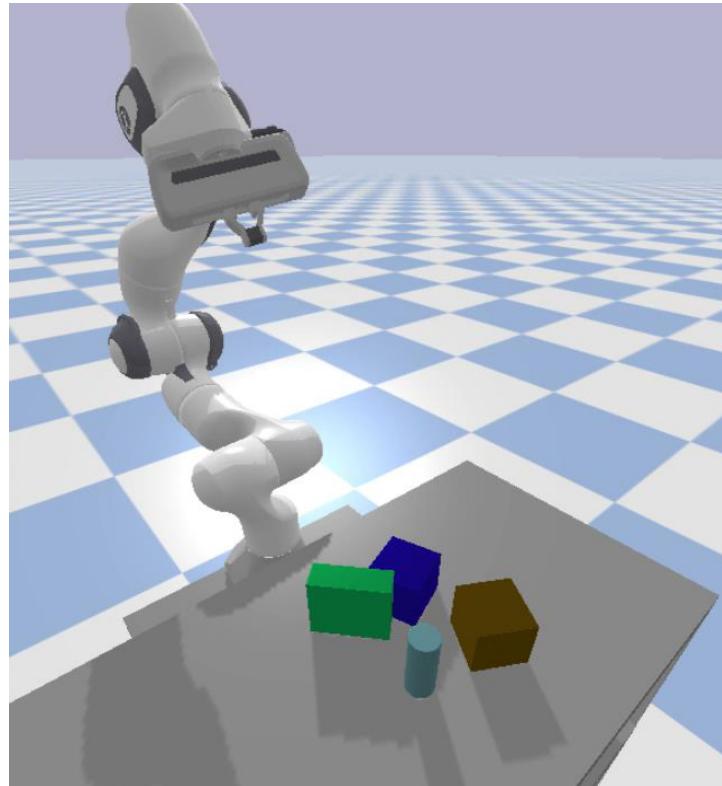
SQPDNet
(S. Kim, et al., CoRL'22)



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

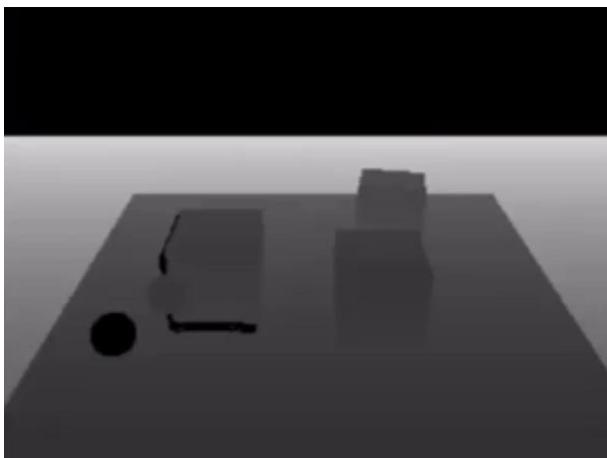


Vision-based Pushing Manipulation

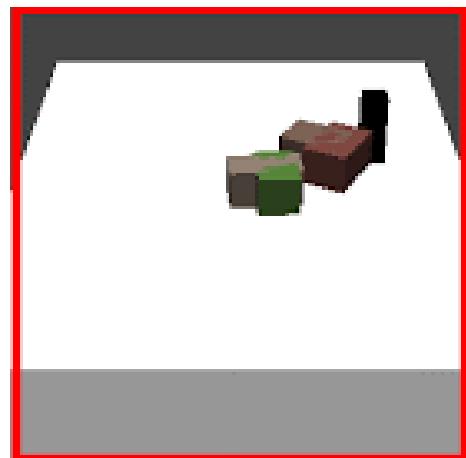




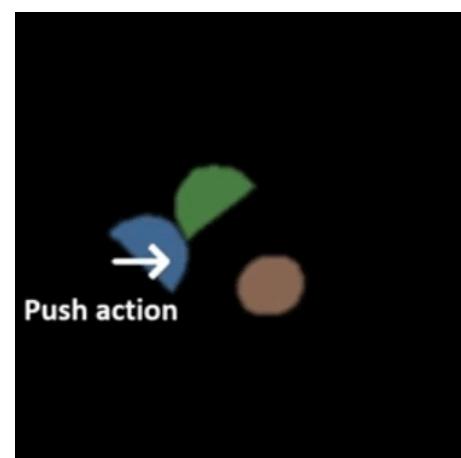
Data-Driven Methods



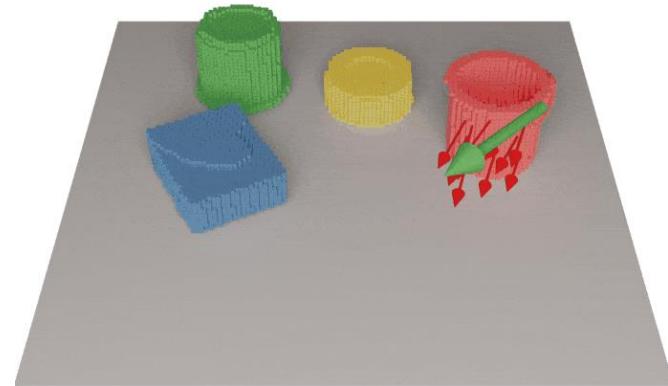
SE3-Net
(A. Byravan, et al., ICRA'17)



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



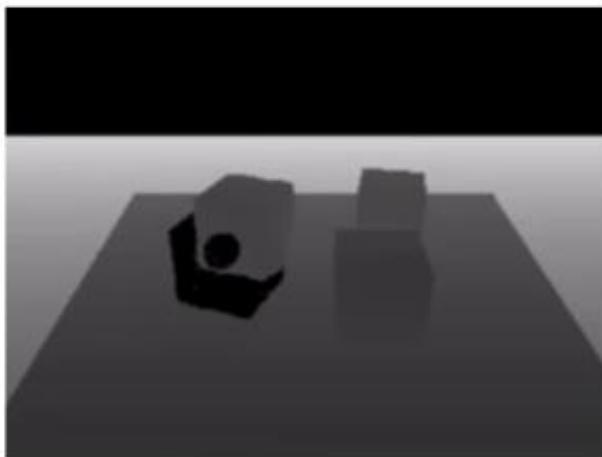
DIPN
(J. Wang, et al., ICRA'21)



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



Data-Driven Methods



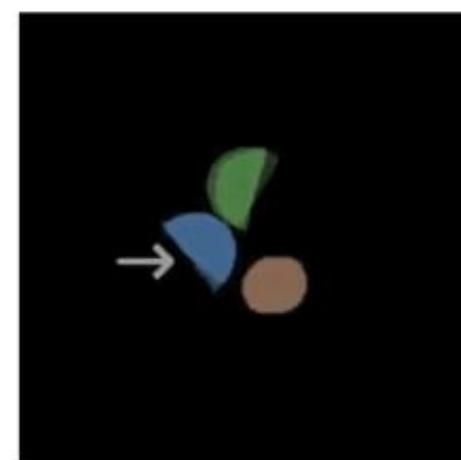
SE3-Net

(A. Byravan, et al., ICRA'17)



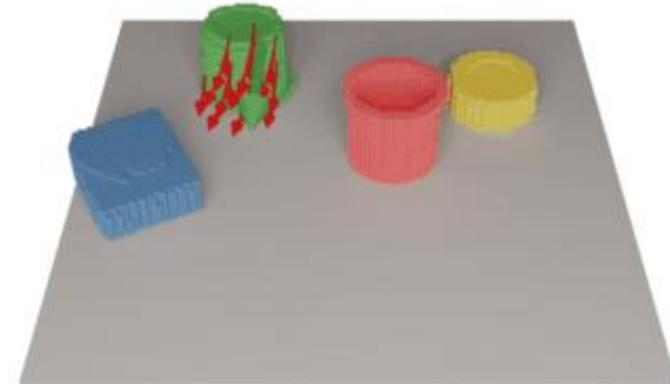
OC-MPC

(Y. Ye, et al., CoRL'19)



DIPN

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



DSR-Net

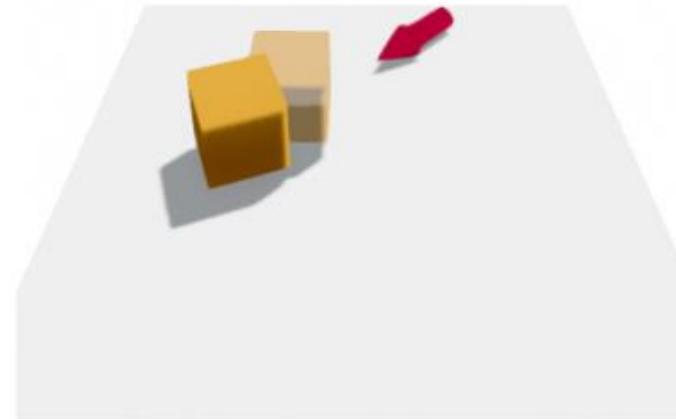
(Z. Xu, et al., CoRL'20)

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



Reducing Needed Training Data: Equivariance

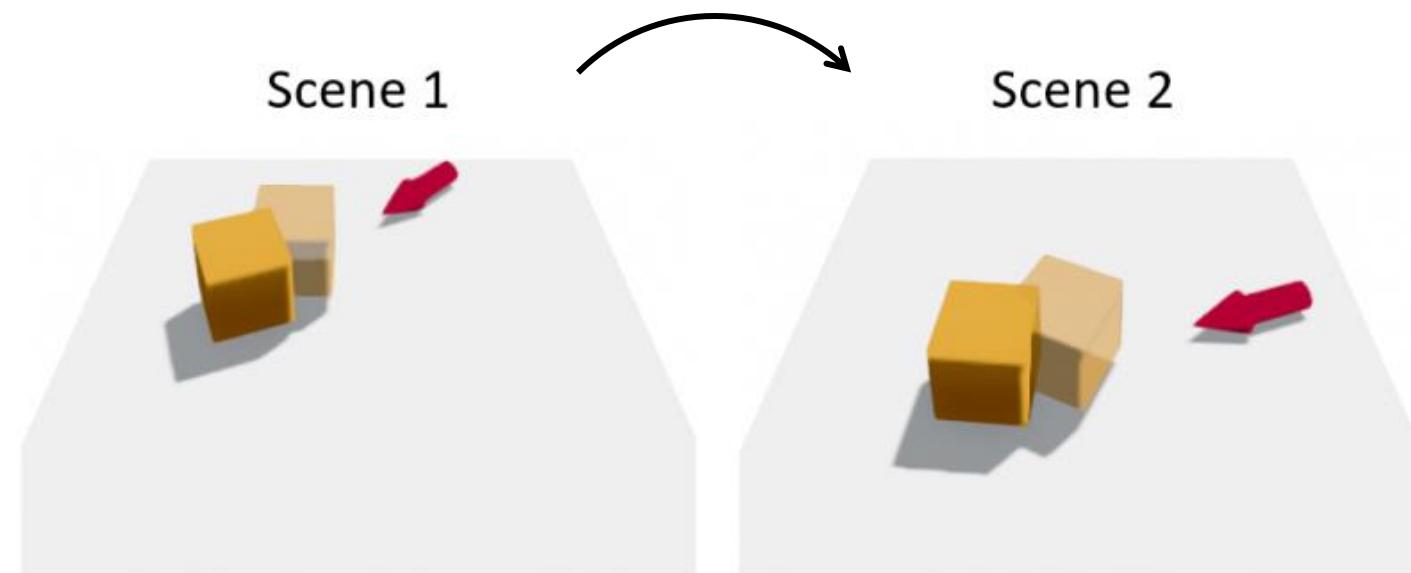
Scene 1





Reducing Needed Training Data: Equivariance

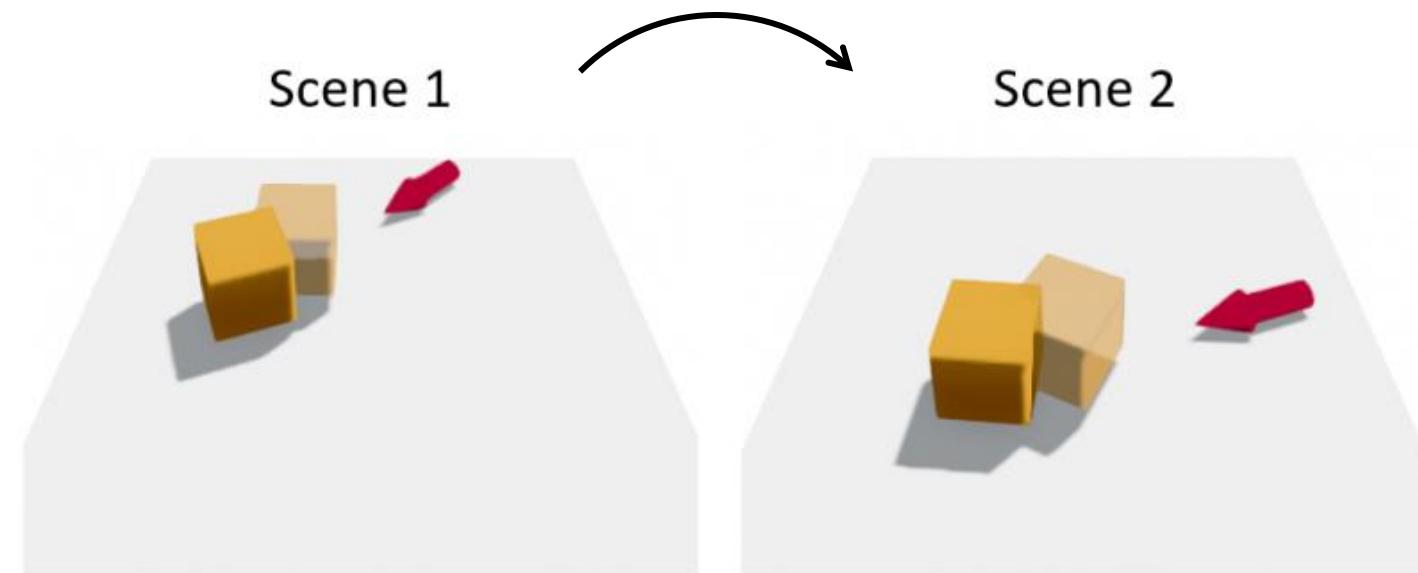
Objects and actions are translated and rotated.





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

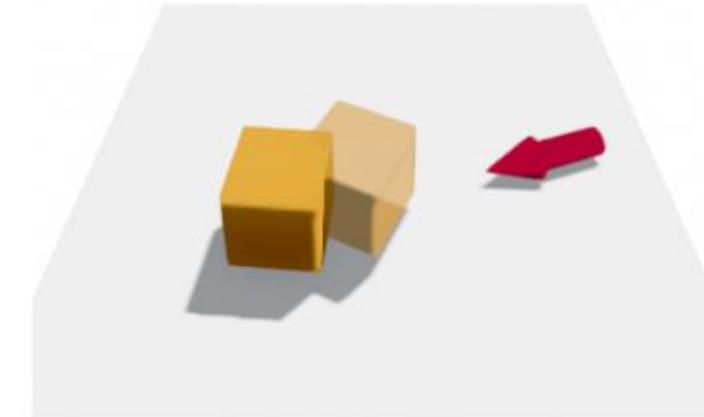
Objects and actions are translated and rotated.

$$\begin{bmatrix} \text{Rot}(\hat{\mathbf{z}}, \theta) & \mathbf{t}_{\mathbf{x}\mathbf{y}} \\ 0 & 1 \end{bmatrix} \in \text{SE}(2)$$

Scene 1



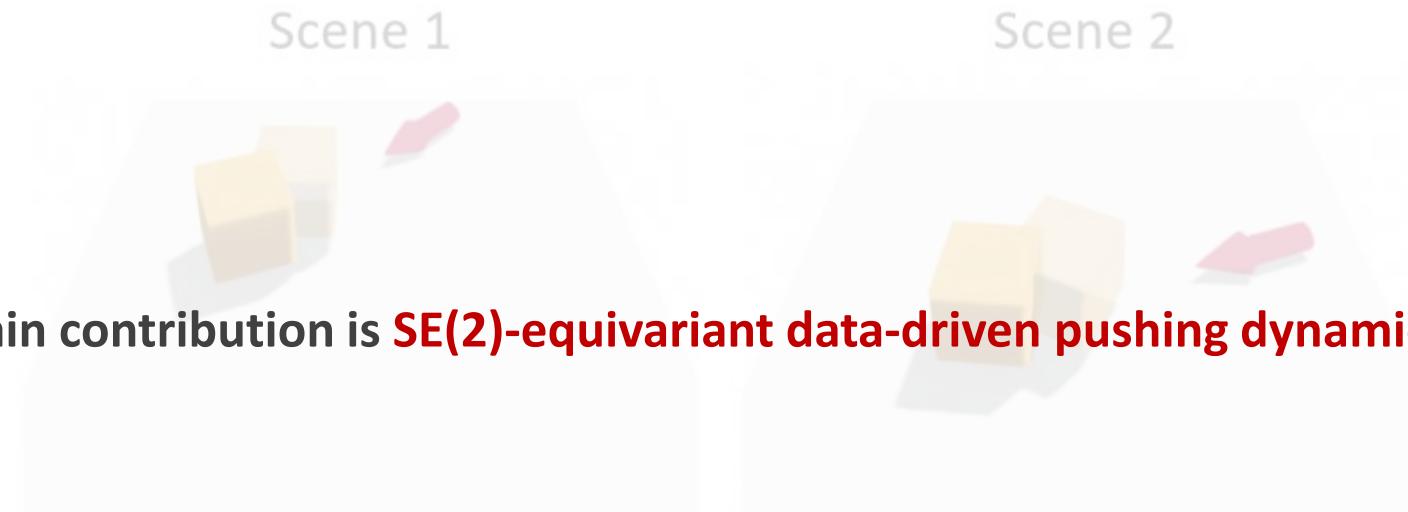
Scene 2



A network that possesses this property is said to be **equivariant** with respect to translations and rotations.
equivariant with respect to $\text{SE}(2)$.

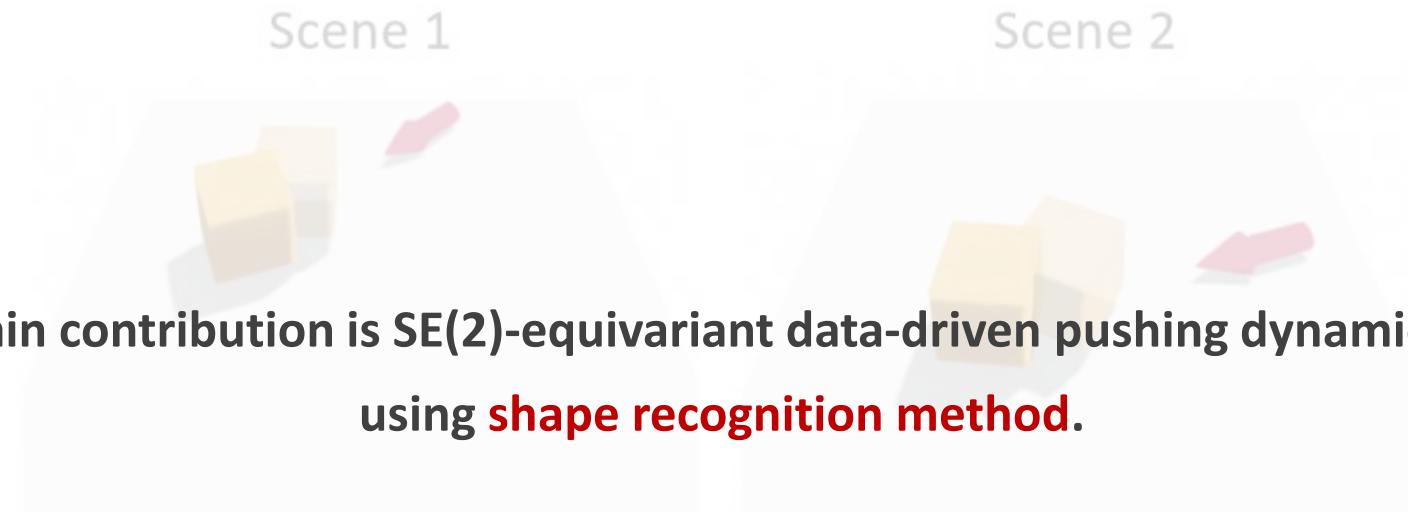


Reducing Needed Training Data: Equivariance





Reducing Needed Training Data: Equivariance





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
 $\mathbf{T}_i \in \text{SE}(3)$ 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

$\mathbf{T}_i \in \text{SE}(3)$ 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.

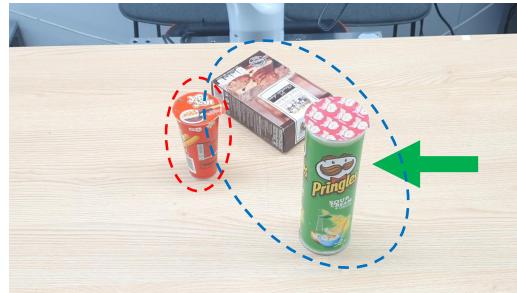
Assume that we know the objects

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

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



SE(2)-Equivariant Network Architecture



Assume that we know the objects

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

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

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



SE(2)-Equivariant Network Architecture



Assume that we know the objects

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

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

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

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

•
•
•

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

SE(2)-Equivariant Network Architecture



Assume that we know the objects

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

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

Apply $\mathbf{C} = \begin{bmatrix} \text{Rot}(\hat{\mathbf{z}}, \theta) & \mathbf{t}_{\mathbf{x}\mathbf{y}} \\ 0 & 1 \end{bmatrix} \in \text{SE}(2)$

↓

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

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

•
•
•

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

SE(2)-Equivariant Network Architecture



Assume that we know the objects

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

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

Apply $\mathbf{C} = \begin{bmatrix} \text{Rot}(\hat{\mathbf{z}}, \theta) & \mathbf{t}_{\mathbf{x}\mathbf{y}} \\ 0 & 1 \end{bmatrix} \in \text{SE}(2)$



$$f(\{\mathbf{CT}_1^t, \mathbf{q}_1\}, \{\mathbf{CT}_2^t, \dots, \mathbf{CT}_N^t, \mathbf{q}_2, \dots, \mathbf{q}_N\}, \mathbf{Ca}^t)$$

$$f(\{\mathbf{CT}_2^t, \mathbf{q}_2\}, \{\mathbf{CT}_1^t, \dots, \mathbf{CT}_N^t, \mathbf{q}_1, \dots, \mathbf{q}_N\}, \mathbf{Ca}^t)$$

•
•
•

$$f(\{\mathbf{CT}_N^t, \mathbf{q}_N\}, \{\mathbf{CT}_1^t, \dots, \mathbf{CT}_{N-1}^t, \mathbf{q}_1, \dots, \mathbf{q}_{N-1}\}, \mathbf{Ca}^t)$$

SE(2)-Equivariant Network Architecture



Assume that we know the objects

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

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

Apply $\mathbf{C} = \begin{bmatrix} \text{Rot}(\hat{\mathbf{z}}, \theta) & \mathbf{t}_{\mathbf{x}\mathbf{y}} \\ 0 & 1 \end{bmatrix} \in \text{SE}(2)$

↓

$$\mathbf{CT}_1^{t+1} = f(\{\mathbf{CT}_1^t, \mathbf{q}_1\}, \{\mathbf{CT}_2^t, \dots, \mathbf{CT}_N^t, \mathbf{q}_2, \dots, \mathbf{q}_N\}, \mathbf{Ca}^t)$$

$$\mathbf{CT}_2^{t+1} = f(\{\mathbf{CT}_2^t, \mathbf{q}_2\}, \{\mathbf{CT}_1^t, \dots, \mathbf{CT}_N^t, \mathbf{q}_1, \dots, \mathbf{q}_N\}, \mathbf{Ca}^t)$$

•
•
•

$$\mathbf{CT}_N^{t+1} = f(\{\mathbf{CT}_N^t, \mathbf{q}_N\}, \{\mathbf{CT}_1^t, \dots, \mathbf{CT}_{N-1}^t, \mathbf{q}_1, \dots, \mathbf{q}_{N-1}\}, \mathbf{Ca}^t)$$

SE(2)-Equivariant Network Architecture



Assume that we know the objects

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

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

Definition 1 A pushing dynamics model f is SE(2)-equivariant if

$$\mathbf{CT}_1^{t+1} = f(\{\mathbf{CT}_1^t, \mathbf{q}_1\}, \{\mathbf{CT}_2^t, \dots, \mathbf{CT}_N^t, \mathbf{q}_2, \dots, \mathbf{q}_N\}, \mathbf{Ca}^t)$$

$$\mathbf{CT}_2^{t+1} = f(\{\mathbf{CT}_2^t, \mathbf{q}_2\}, \{\mathbf{CT}_1^t, \dots, \mathbf{CT}_N^t, \mathbf{q}_1, \dots, \mathbf{q}_N\}, \mathbf{Ca}^t)$$

•
•
•

$$\mathbf{CT}_N^{t+1} = f(\{\mathbf{CT}_N^t, \mathbf{q}_N\}, \{\mathbf{CT}_1^t, \dots, \mathbf{CT}_{N-1}^t, \mathbf{q}_1, \dots, \mathbf{q}_{N-1}\}, \mathbf{Ca}^t)$$

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



SE(2)-Equivariant Network Architecture

Visual observation





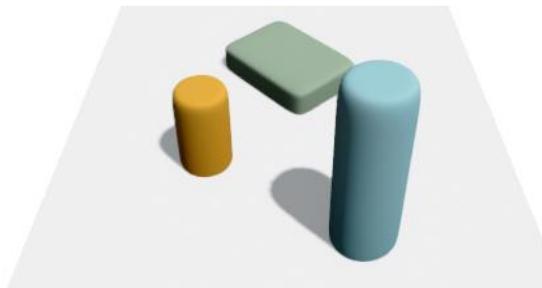
SE(2)-Equivariant Network Architecture

Visual observation

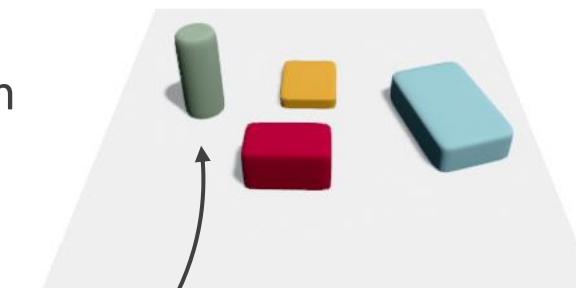


Shape
Recognition
→

Recognized 3D objects



Shape
Recognition
→



$\mathbf{T}_i \in \text{SE}(3)$ Object poses
 \mathbf{q}_i Shape parameters

SE(2)-Equivariant Network Architecture



Visual observation

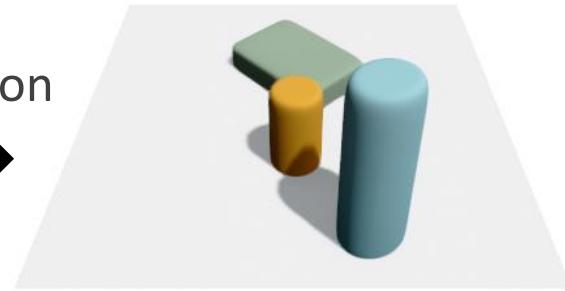
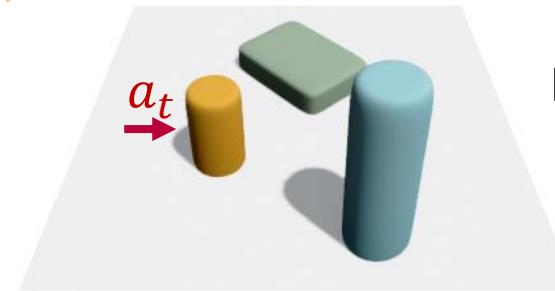


$\mathbf{T}_i \in \text{SE}(3)$ Object poses
 \mathbf{q}_i Shape parameters

Recognized 3D objects

a_t

Prediction



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

SE(2)-Equivariant Network Architecture



Visual observation

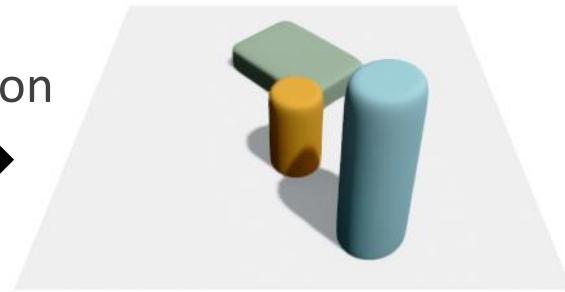
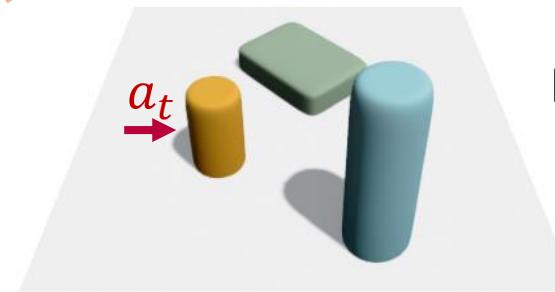


$\mathbf{T}_i \in \text{SE}(3)$ Object poses
 \mathbf{q}_i Shape parameters

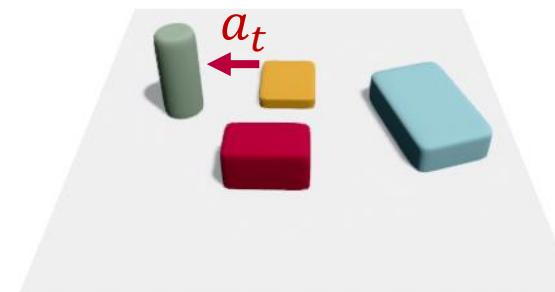
Recognized 3D objects

a_t

Prediction



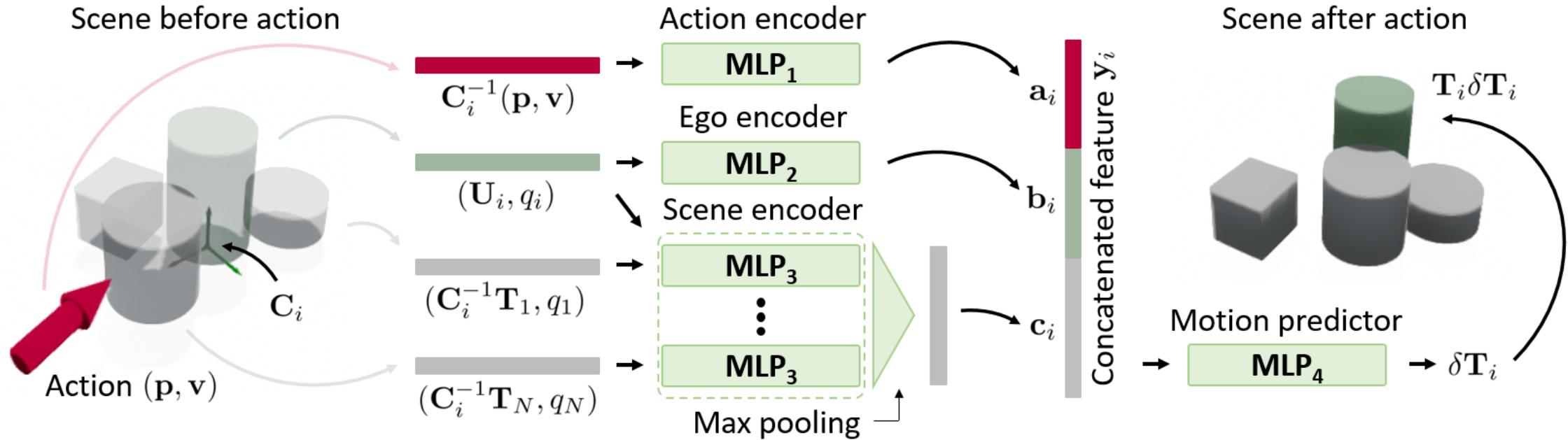
a_t



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

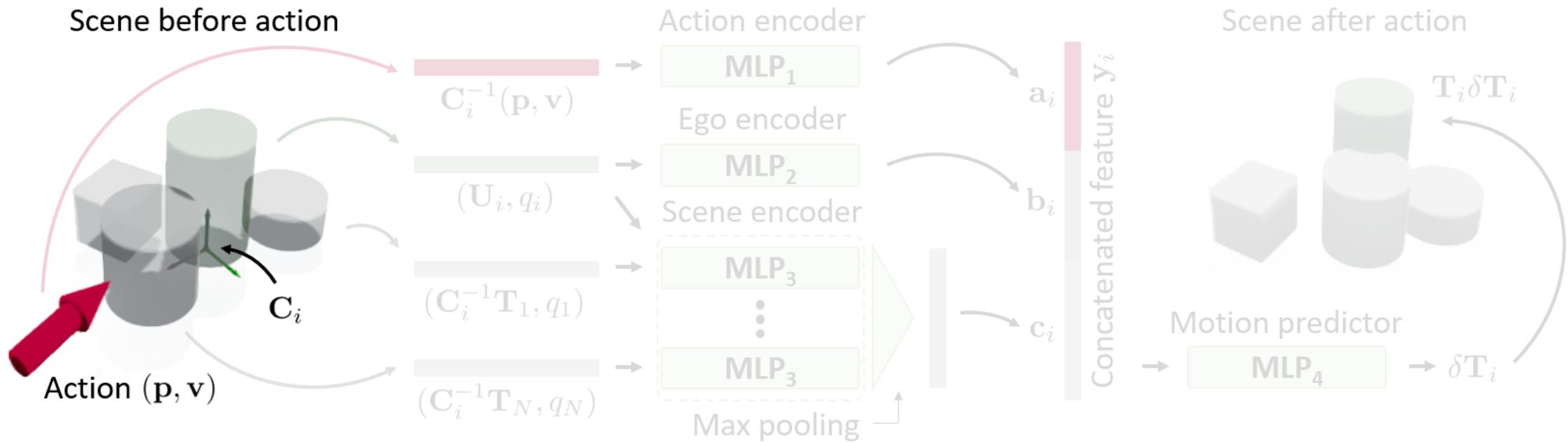
Superquadric Pushing Dynamics Model (SQPDNet)

SE(2)-Equivariant Network Architecture

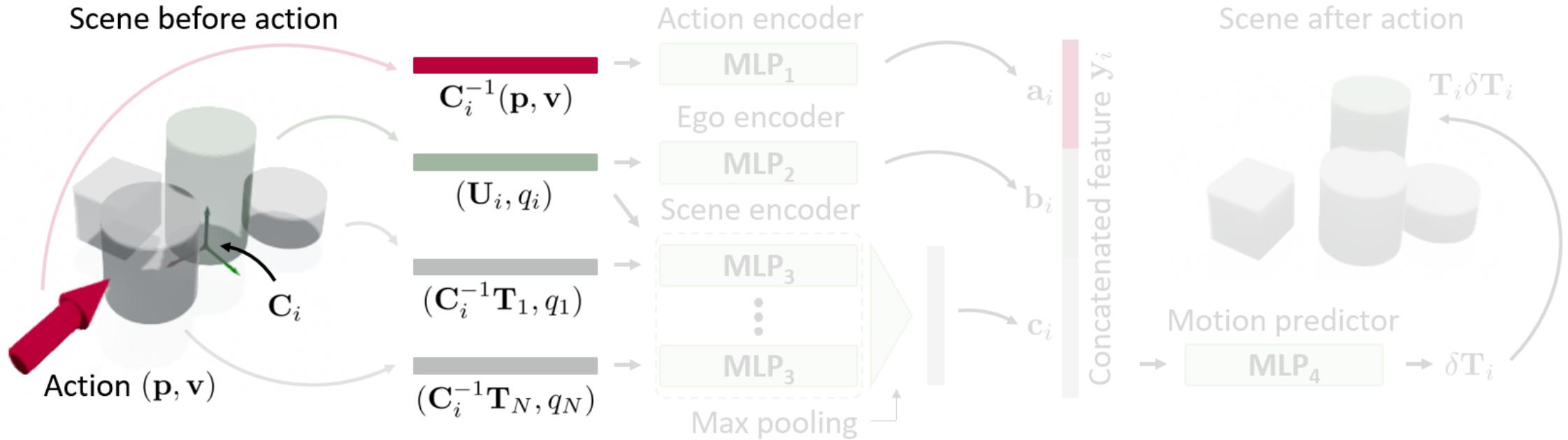




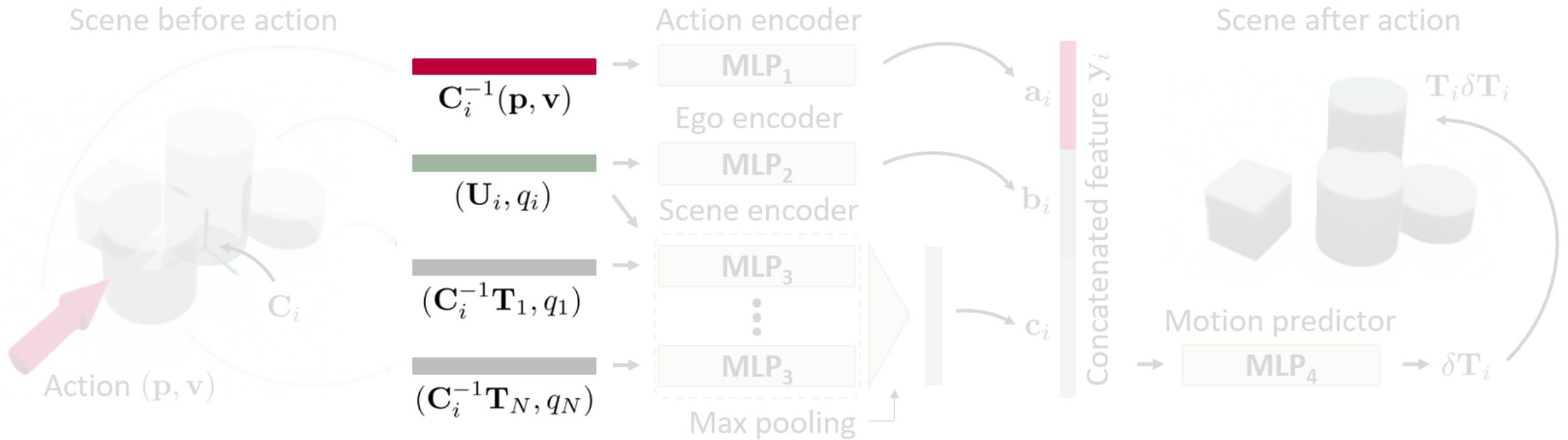
SE(2)-Equivariant Network Architecture



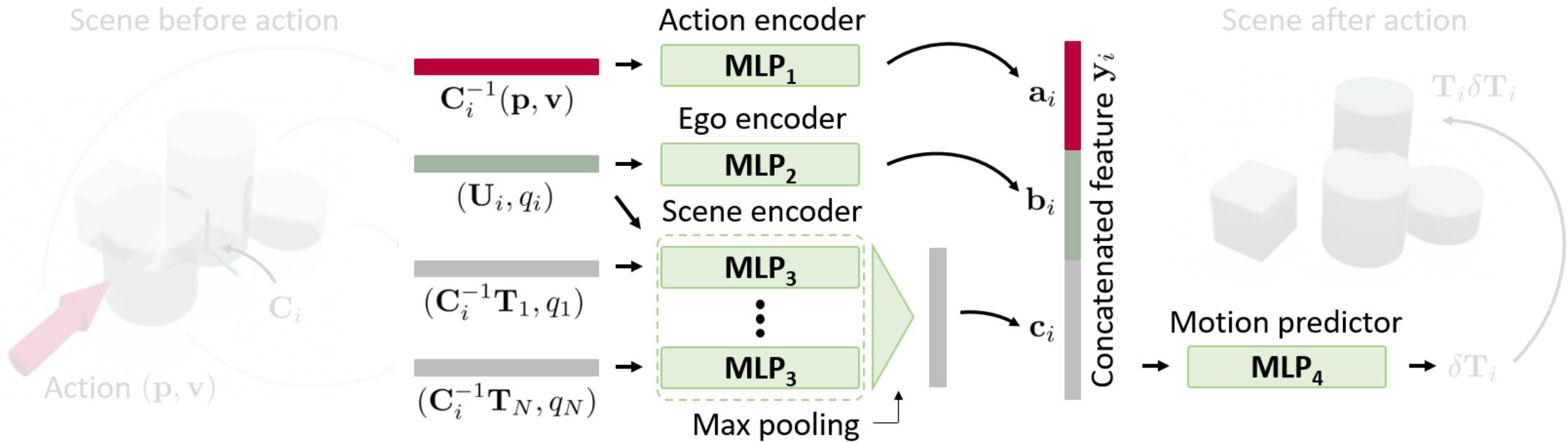
SE(2)-Equivariant Network Architecture



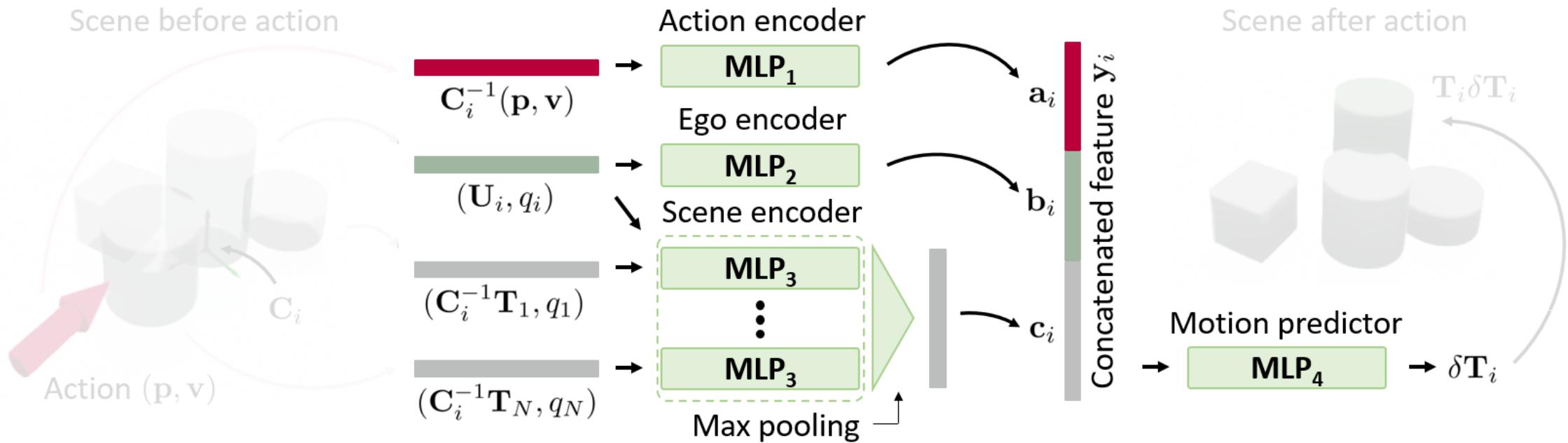
SE(2)-Equivariant Network Architecture



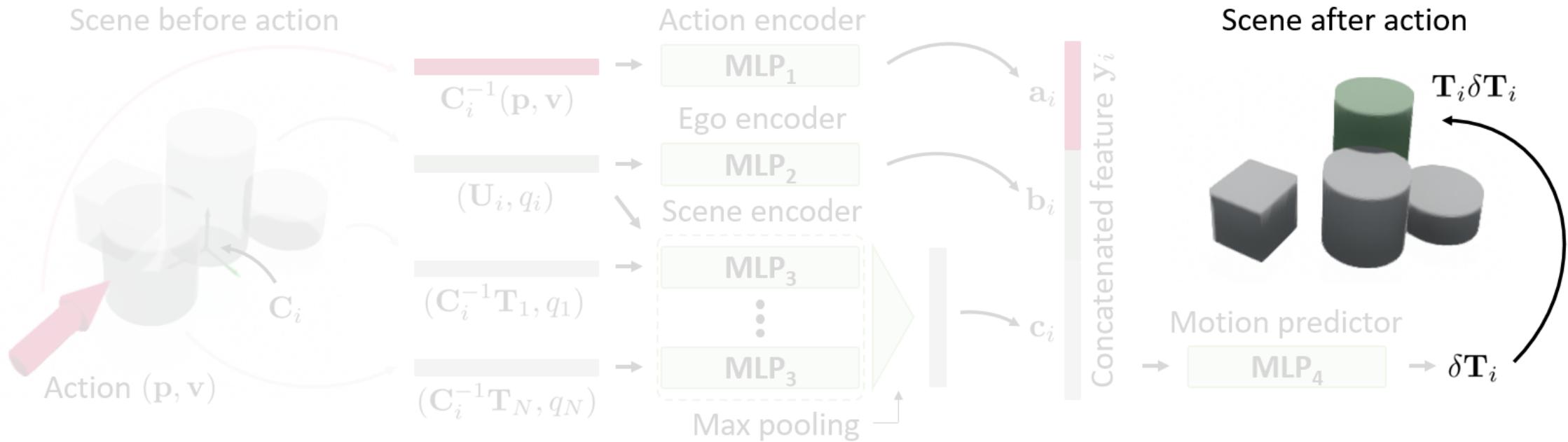
SE(2)-Equivariant Network Architecture



SE(2)-Equivariant Network Architecture



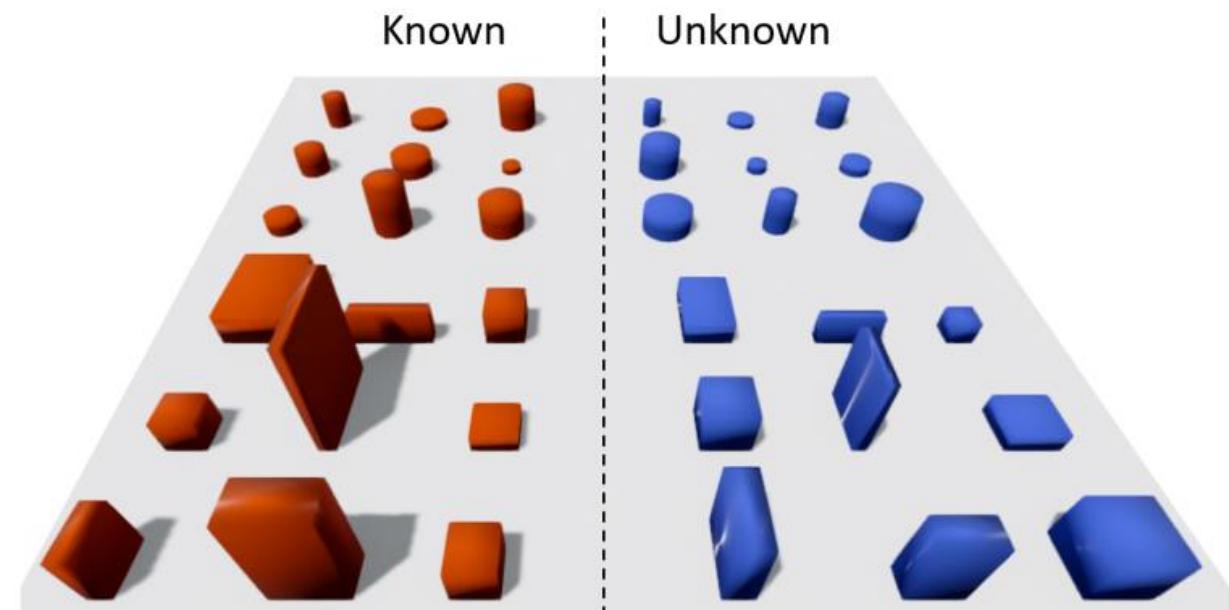
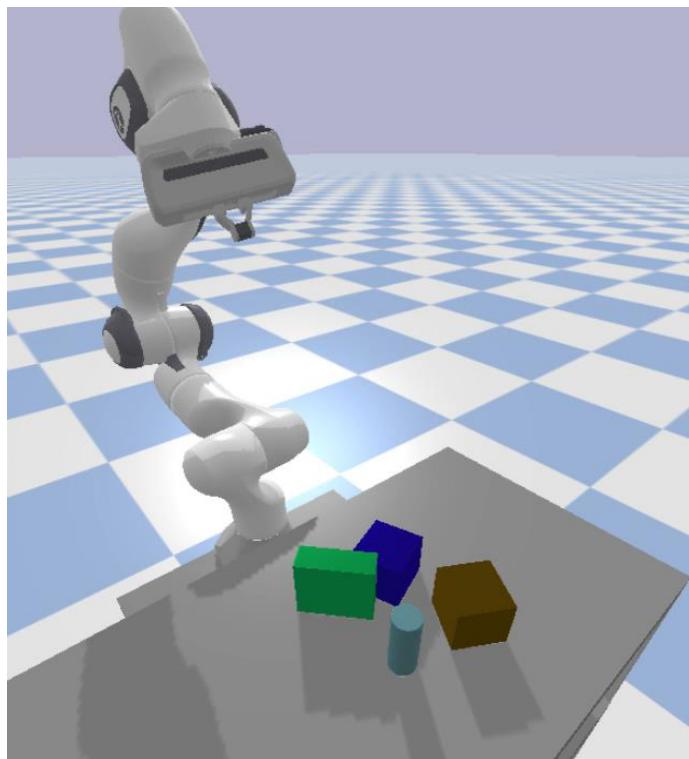
SE(2)-Equivariant Network Architecture





Experimental Results

Pushing manipulation dataset





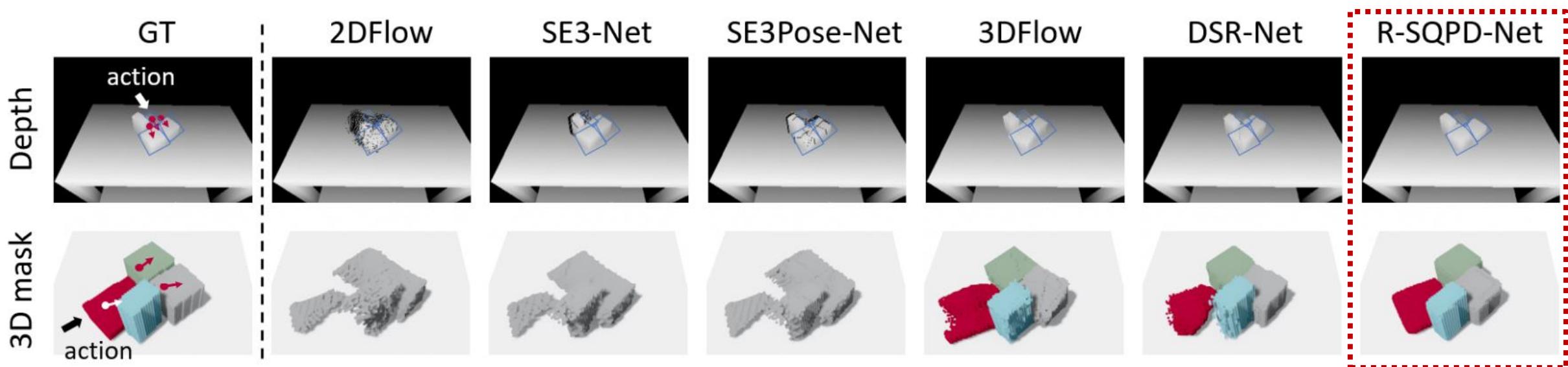
Experimental Results

METHOD	Known				Unknown				
	Flow error (↓)		Mask IoU (↑)		Flow error (↓)		Mask IoU (↑)		
	visible	full	2D	3D		visible	full	2D	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



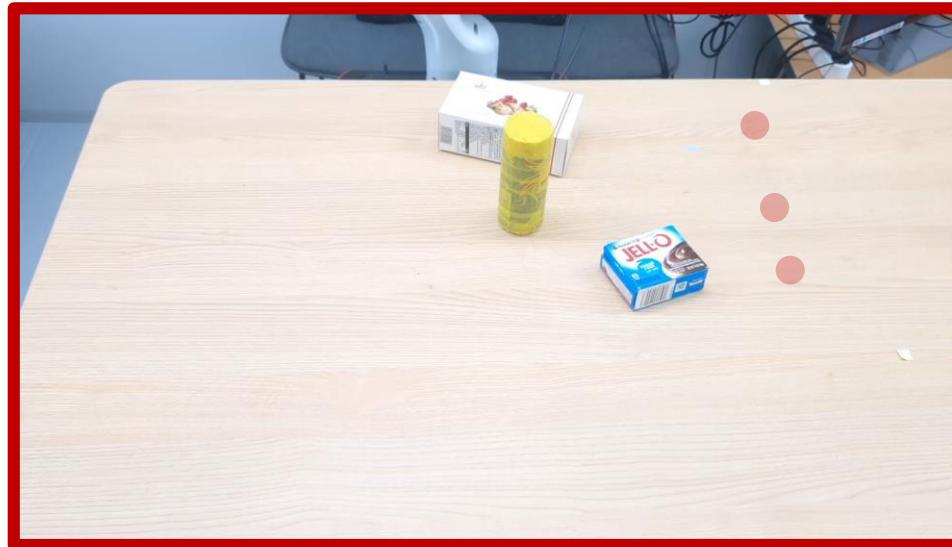
Object singulation task





Robot Pushing Manipulation

Object moving task



Object singulation task



- Move the objects to their desired poses.

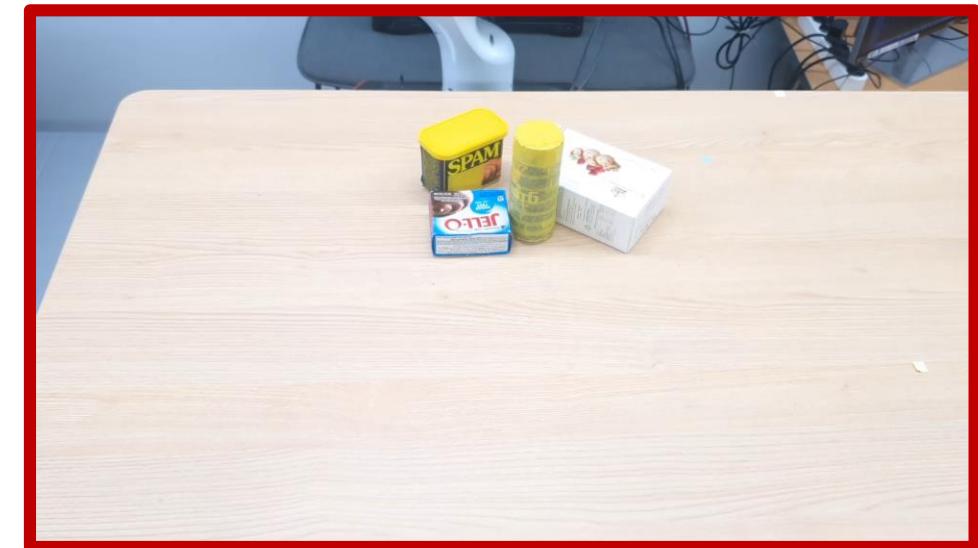
Robot Pushing Manipulation



Object moving task



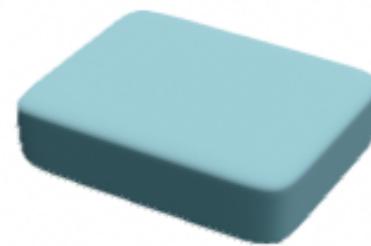
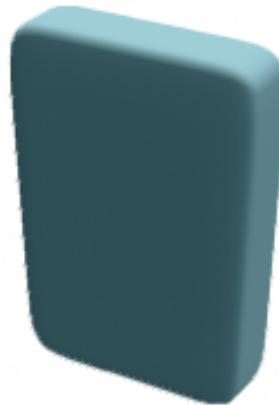
Object singulation task



- Separate the objects by more than a certain distance τ (e.g., $\tau = 20\text{cm}$).

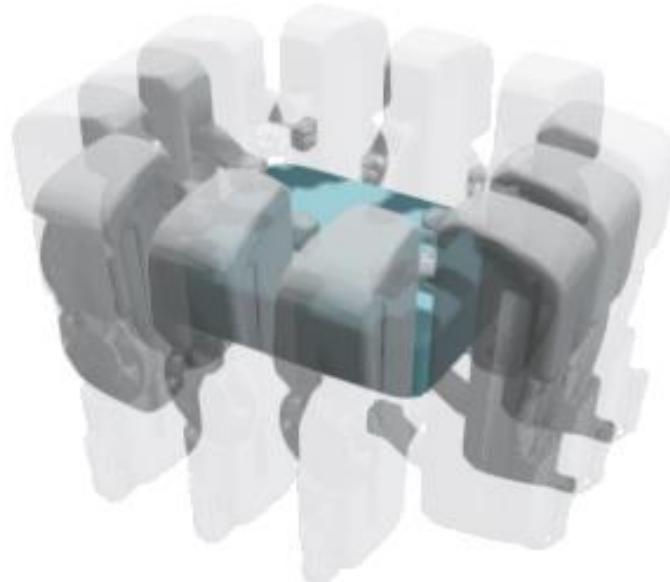
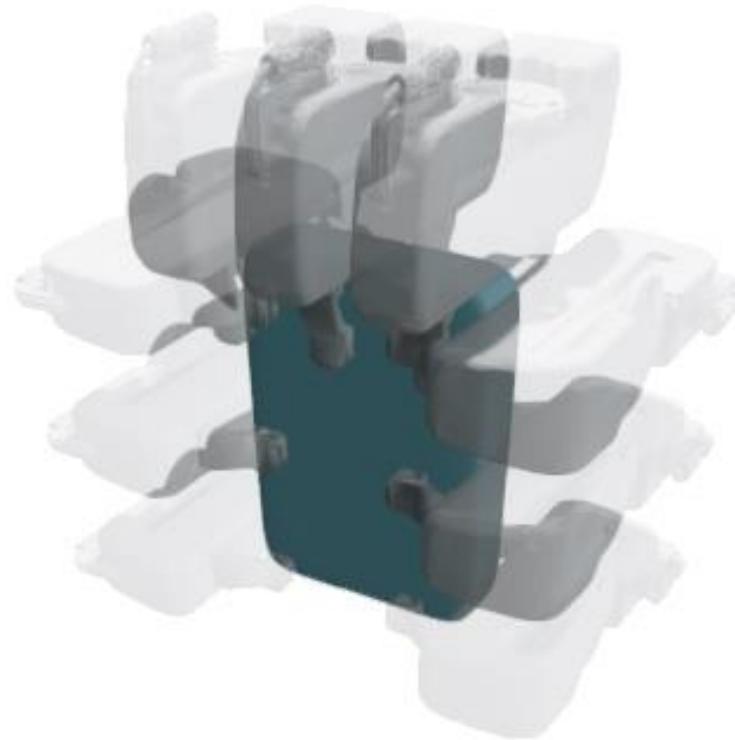


Robot Pushing Manipulation





Robot Pushing Manipulation

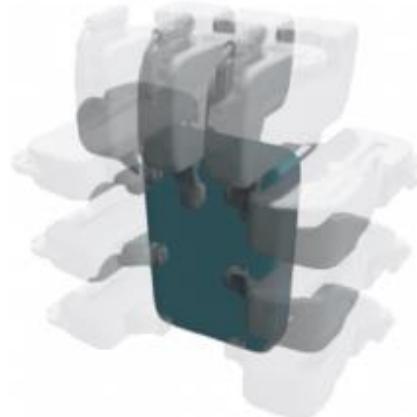


Given the pose and shape parameters of the object, generating grasp poses is easy.

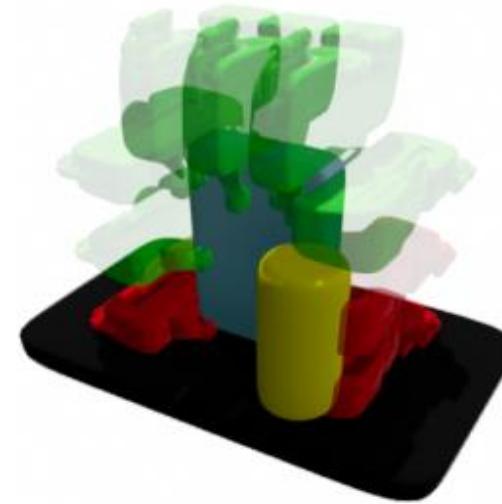


Robot Pushing Manipulation

Pre-defined
grasp poses



Collision-free
grasp poses

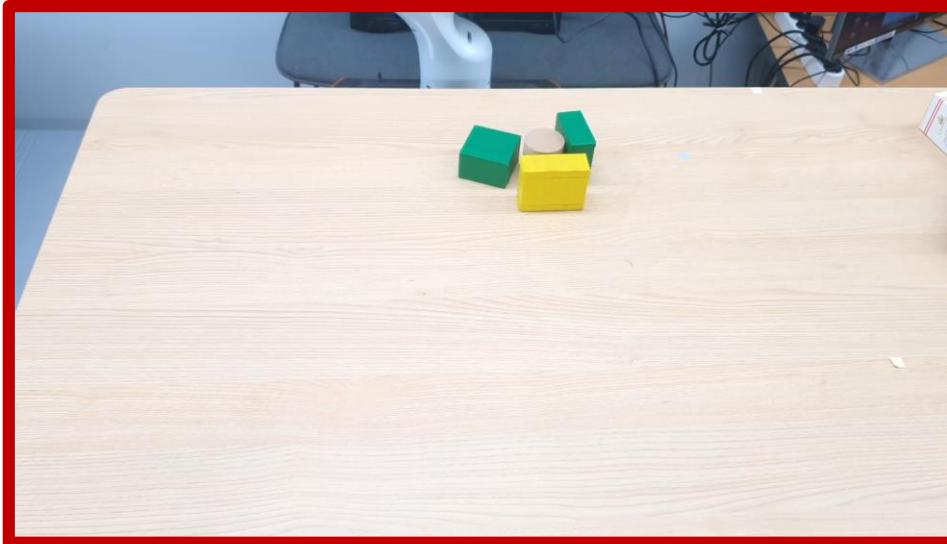


→ Grasp reward is 1 if valid grasp pose exists, 0 otherwise
Collision free from the table and other objects

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)



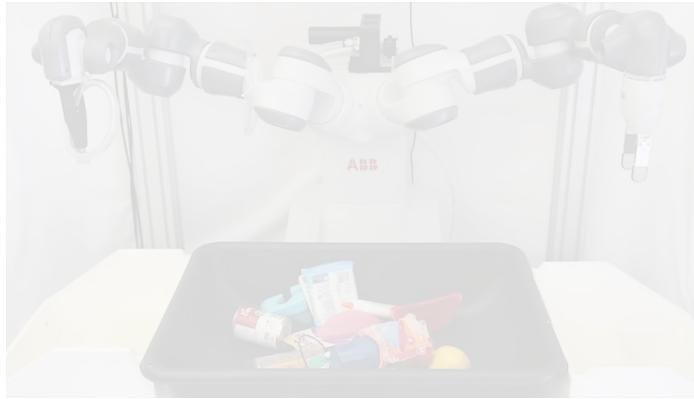
Grasping Invisible (Y. Yang, et al., RA-L'20)

Cannot be directly applied to the shelf environment!

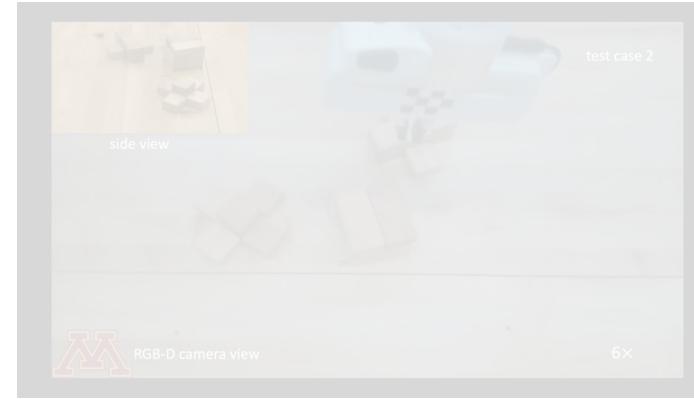
- **Limited action space of the manipulator**
- **Limited amount of visual information**



Mechanical Search Methods



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



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



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



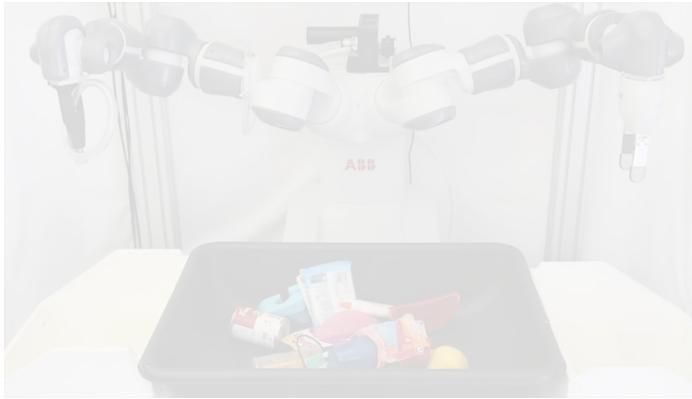
Bluction-DAR (H. Huang, et al., ICRA'22)

Cannot be directly applied to the shelf environment!

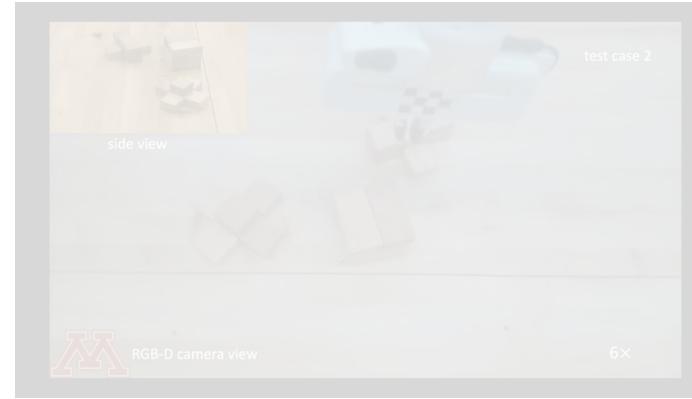
- Limited action space of the manipulator
- Limited amount of visual information is limited



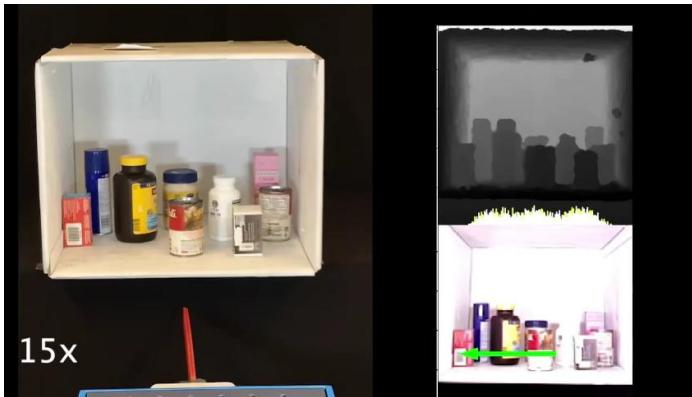
Mechanical Search Methods



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



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



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



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

Cannot be directly applied to the shelf environment!

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

- 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 $f: \text{SE}(3) \rightarrow \{0, 1\}$

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



A General Framework for Mechanical Search

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

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

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

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

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

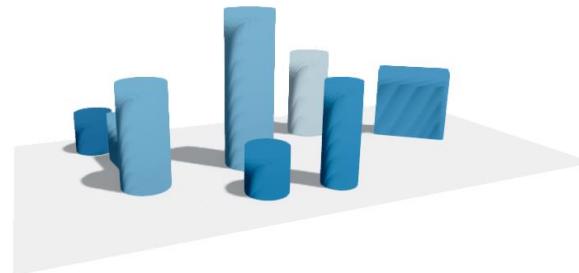


A General Framework for Mechanical Search

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

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

Given Observation

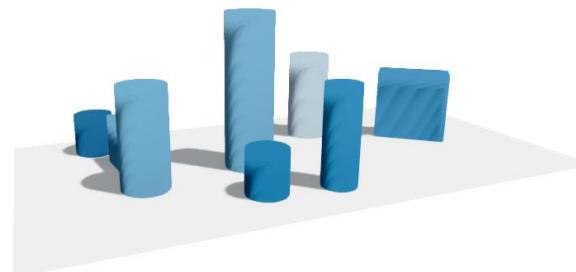


A General Framework for Mechanical Search

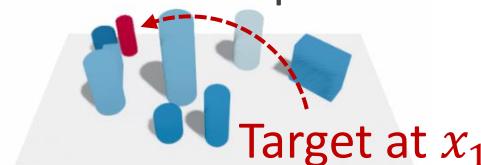
Find and grasp the desired target object on a cluttered shelf!

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

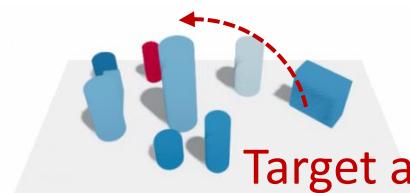
Given Observation



Candidate poses



Target at x_2

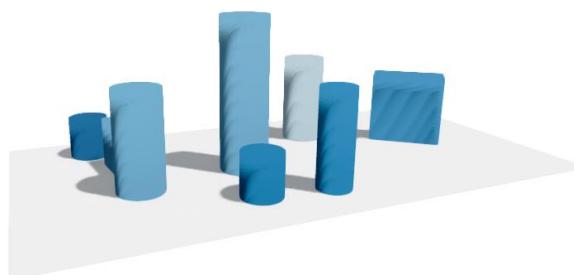


A General Framework for Mechanical Search

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

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

Given Observation



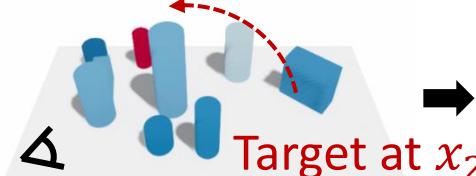
Candidate poses



Observation



Camera

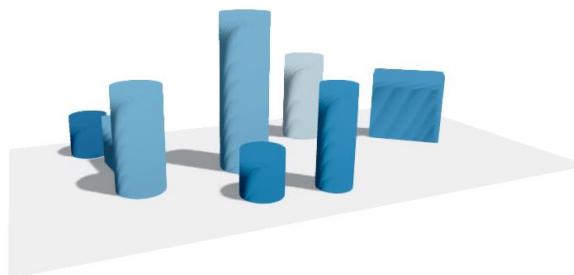


A General Framework for Mechanical Search

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

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

Given Observation



Candidate poses

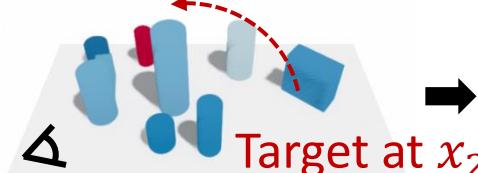


Observation



$$f(x_1) = 1$$

Camera

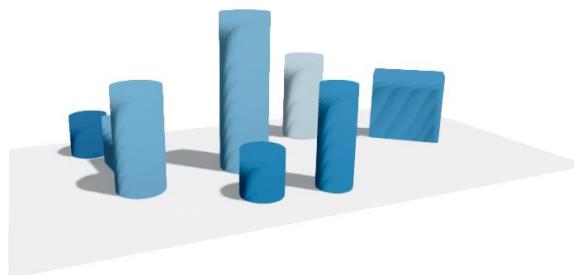


A General Framework for Mechanical Search

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

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

Given Observation



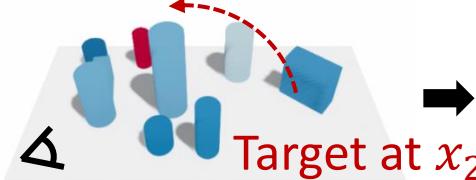
Candidate poses



Observation

$$f(x_1) = 1$$

Camera



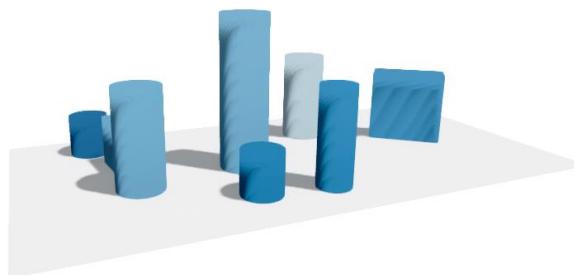
$$f(x_2) = 0$$

A General Framework for Mechanical Search

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

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

Given Observation



Candidate poses

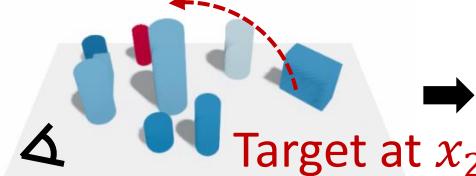


Observation



$$f(x_1) = 1$$

Camera



$$f(x_2) = 0$$

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

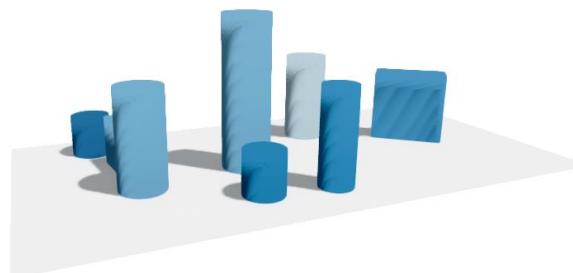
uncertainty of actual target pose \uparrow

A General Framework for Mechanical Search

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

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

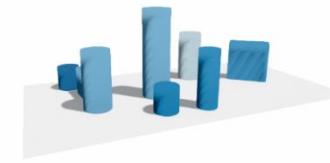
Given Observation



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 \uparrow

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

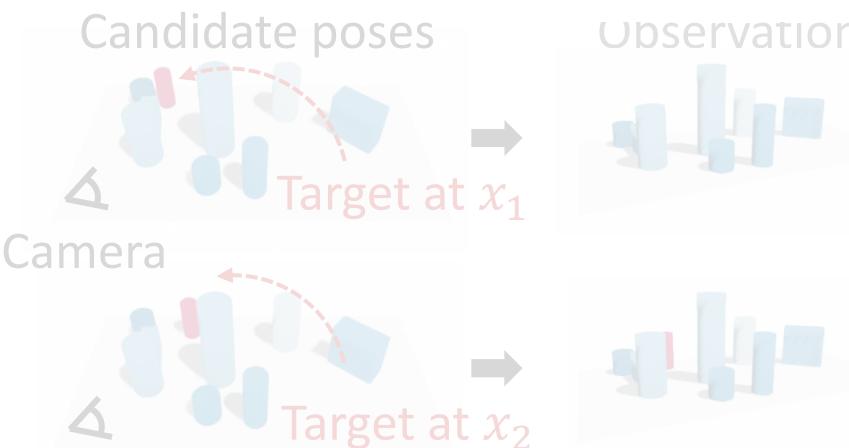
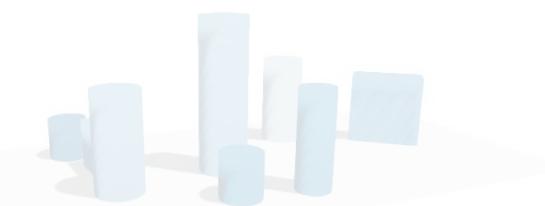
A General Framework for Mechanical Search

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

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

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

Given Observation



$$f(x_1) = 1$$

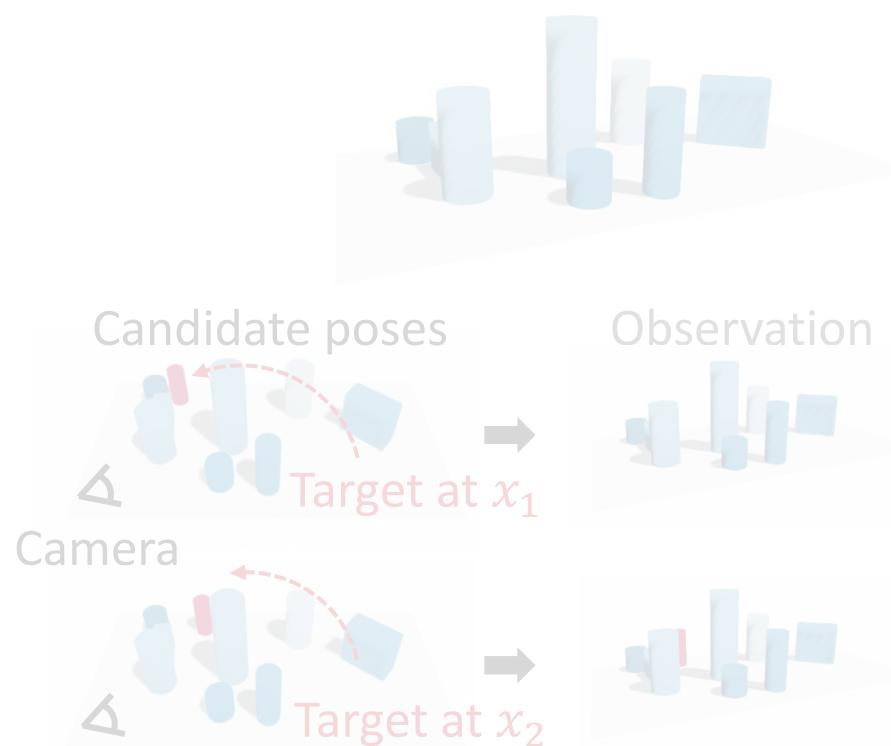
$$f(x_2) = 0$$

A General Framework for Mechanical Search

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

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

Given Observation



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



$$f(x_1) = 1$$

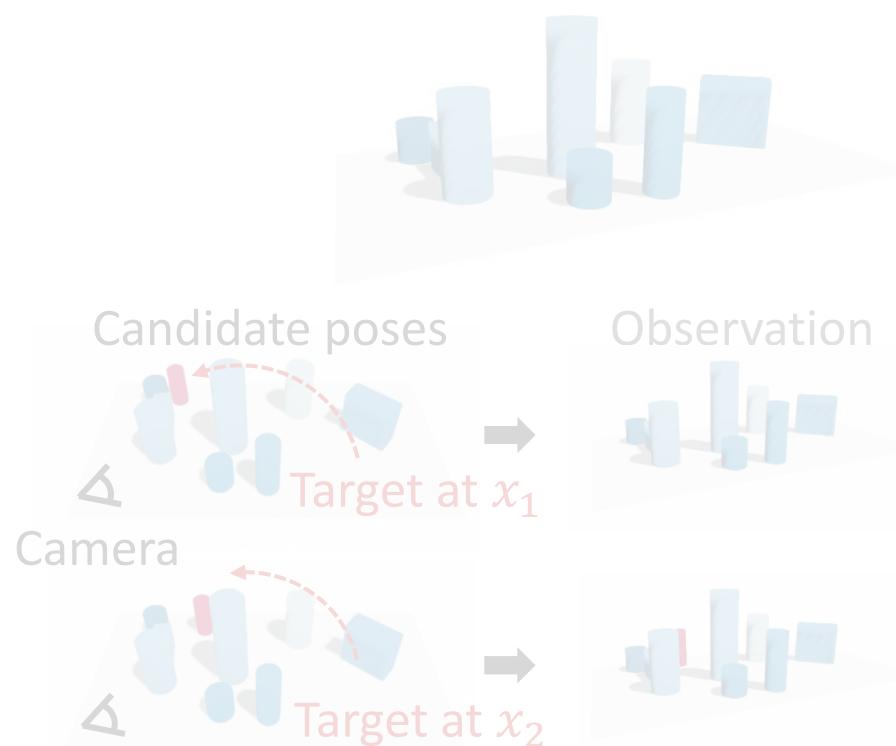
$$f(x_2) = 0$$

A General Framework for Mechanical Search

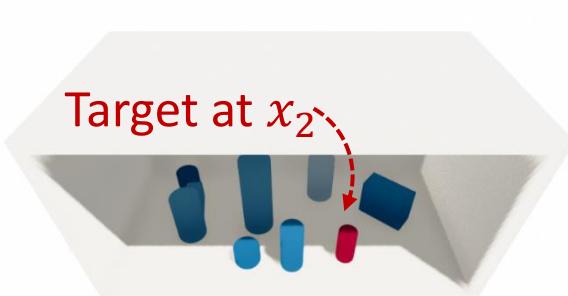
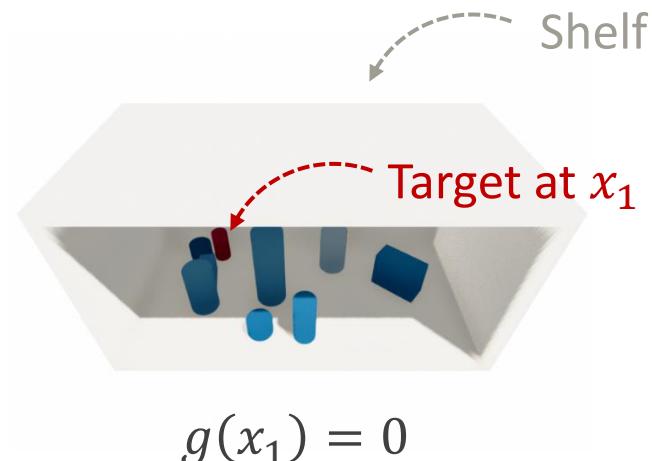
Find and grasp the desired target object on a cluttered shelf!

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

Given Observation



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

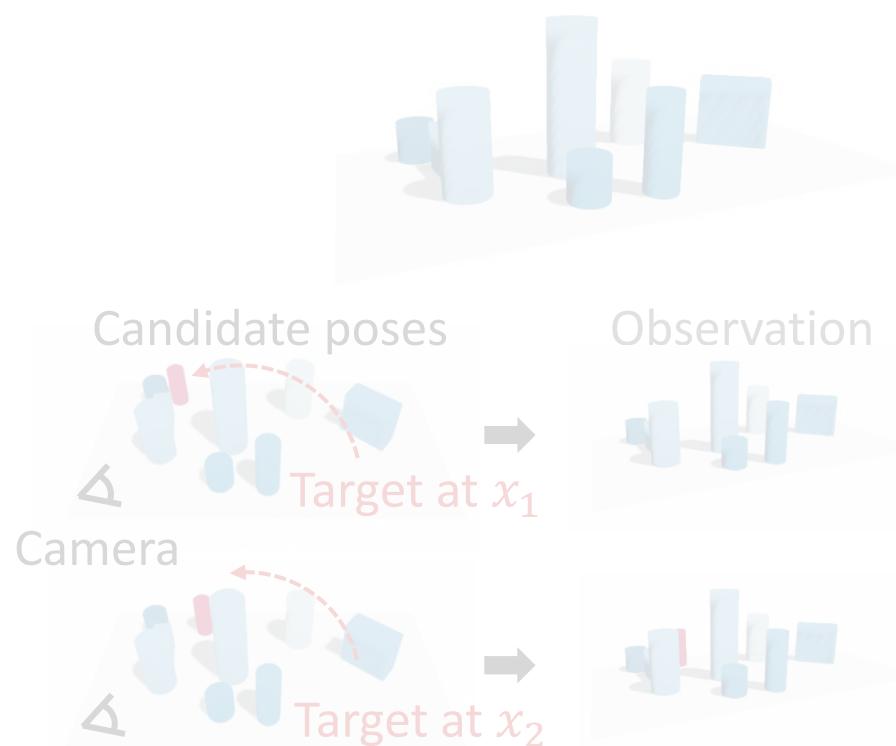


A General Framework for Mechanical Search

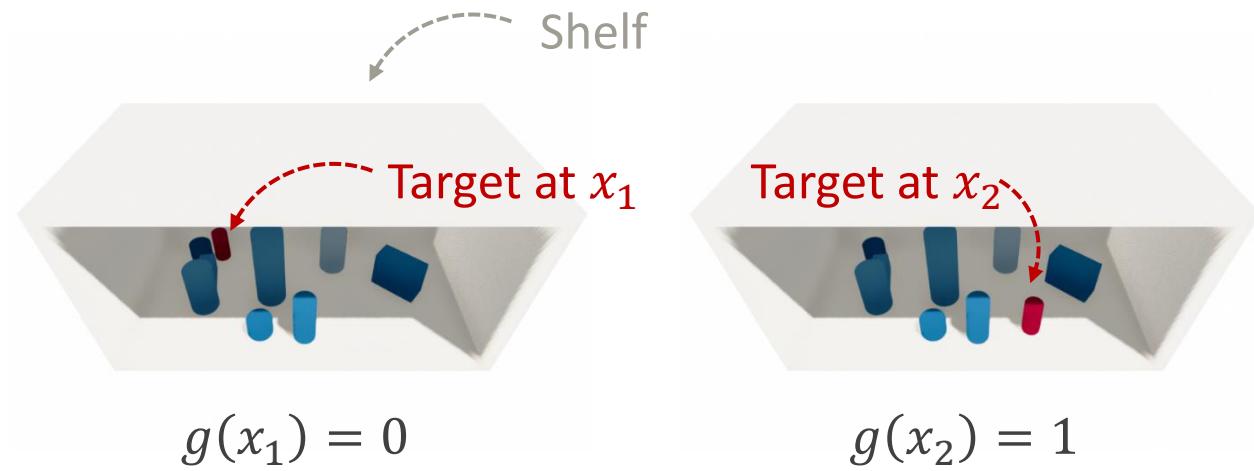
Find and grasp the desired target object on a cluttered shelf!

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

Given Observation



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



$$f(x_1) = 1$$

$$f(x_2) = 0$$

$$g(x_1) = 0$$

$$g(x_2) = 1$$



A General Framework for Mechanical Search

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

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

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

A General Framework for Mechanical Search

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

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

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

Optimal control formulation

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



Leveraging Shape Recognition

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

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

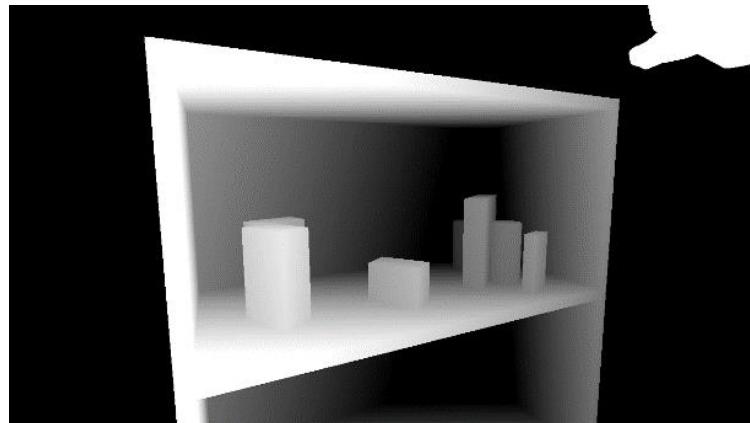
Graspability Function $g: \text{SE}(3) \rightarrow \{0, 1\}$

Leveraging Shape Recognition

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

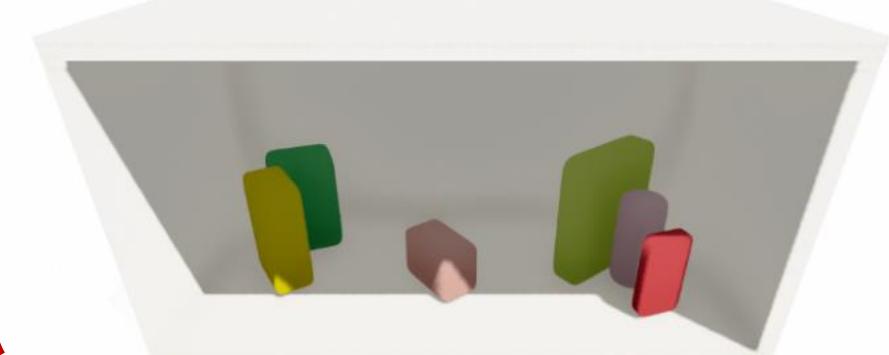
Existence Function $\hat{f}: \text{SE}(3) \rightarrow \{0, 1\}$

Observed Depth image



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

Recognized 3D objects





Leveraging Shape Recognition

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

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

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

Optimal control formulation

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

Leveraging Shape Recognition

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

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

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

Optimal control formulation

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

Tractable optimal control formulation

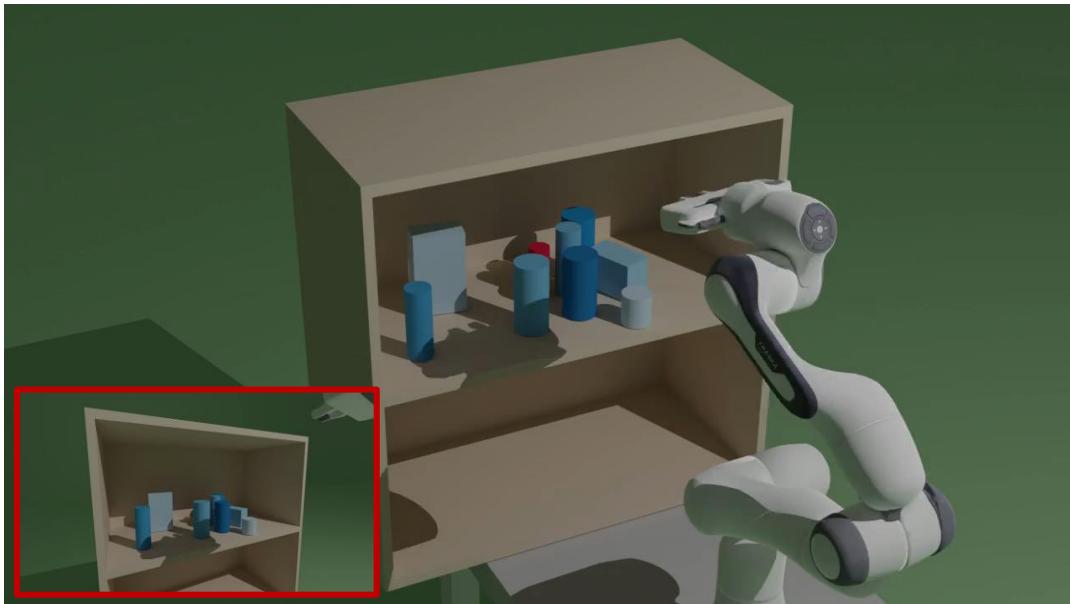
$$\min_{\{a_i\}_{i=1}^T} \sum_{x \in \mathcal{X}} \hat{f}_T(x) + \alpha \hat{f}_T(x)(1 - \hat{g}_T(x))$$

Approximate by leveraging
3D shape recognition

Experimental Results



Simulation environment



Real-world environment



Experimental Results

METHOD		The number of objects							
		2		4		6		8	
		Find	Grasp	Find	Grasp	Find	Grasp	Find	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>