OP-Align:
Object-level and Part-level Alignment for Self-supervised Category-level Articulated Object Pose Estimation

Yuchen Che[1], Ryo Furukawa[2], Asako Kanezaki[1]

[1]Tokyo Institute of Technology, [2]Accenture Japan Ltd.

東京工業大学
Tokyo Institute of Technology
Automation & Knowledge Lab.

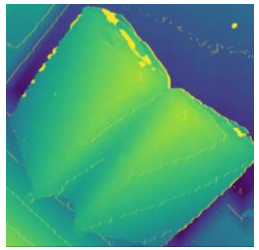
Category-level Articulated Object Pose Estimation

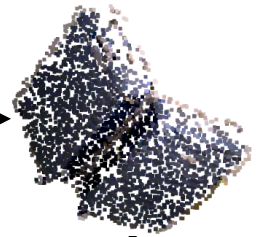
Task Overview

RGB

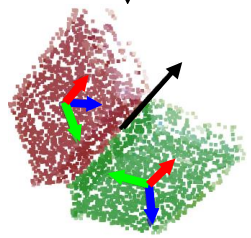
Depth

Mask





Model



Single-frame
Point Cloud

Part Segmentation
Part Pose
Joint Pivot
Joint Direction

Challenge

Huge variance in object pose, joint state

Object Pose

Joint States

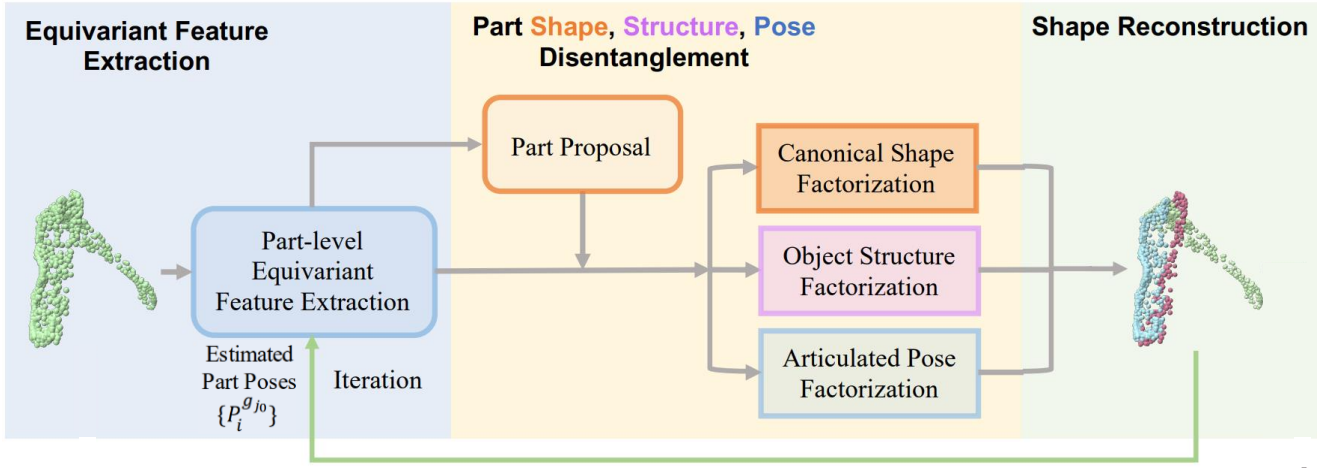
Previous Self-supervised Approaches

- Require **additional supervision / input**
- Cannot achieve **real-time inference speed**

Self-supervised Approaches

Method	w/o Pose Supervision	w/o Shape Supervision	Single Frame	Real-time Inference
PartMobility [35]	✓	✓		
UPPD [16]	✓		✓	✓
EAP [24]	✓	✓	✓	
Ours	✓	✓	✓	✓

Overview of EAP Strategy



Liu, Xueyi, et al. Self-supervised Category-level Articulated Object Pose Estimation with Part-level SE(3)-equivariance. ICLR, 2023.

- A new **state-of-the-art self-supervised model: OP-Align**
- A new **real-world dataset**

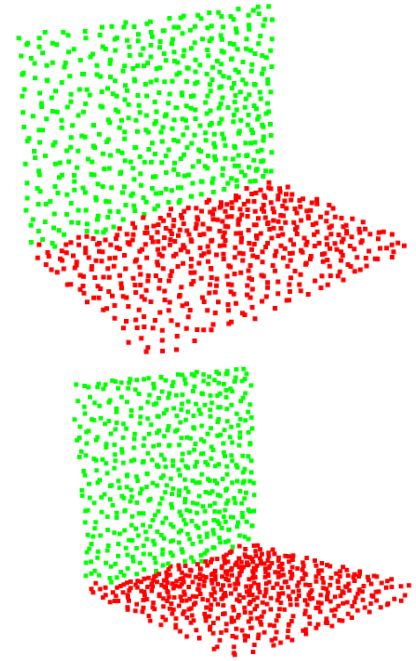
OP-Align: Object-level and Part-level Alignment & Reconstruction

- Construct a **canonical reconstruction** with a fixed pose and joint states among the category.
- Estimate **transformations** which aligns an input and this canonical reconstruction.

(A)

(B)

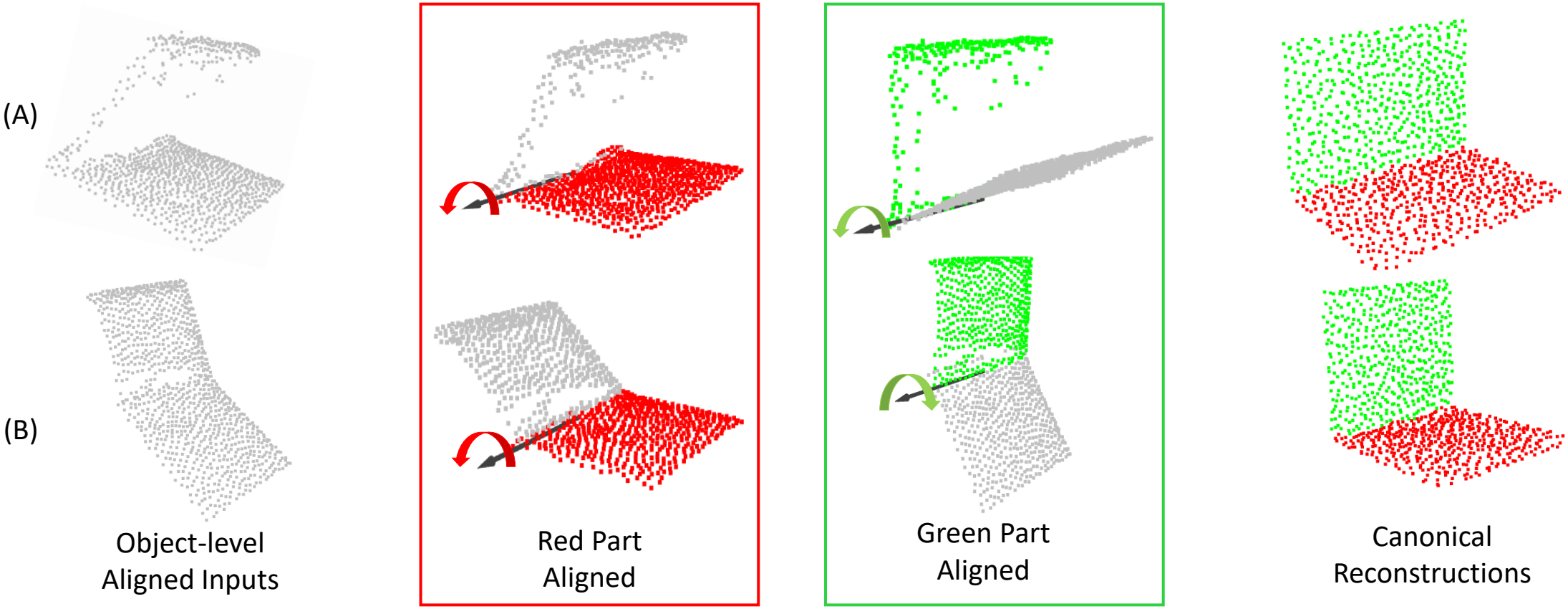
Arbitrary Inputs



Canonical Reconstructions

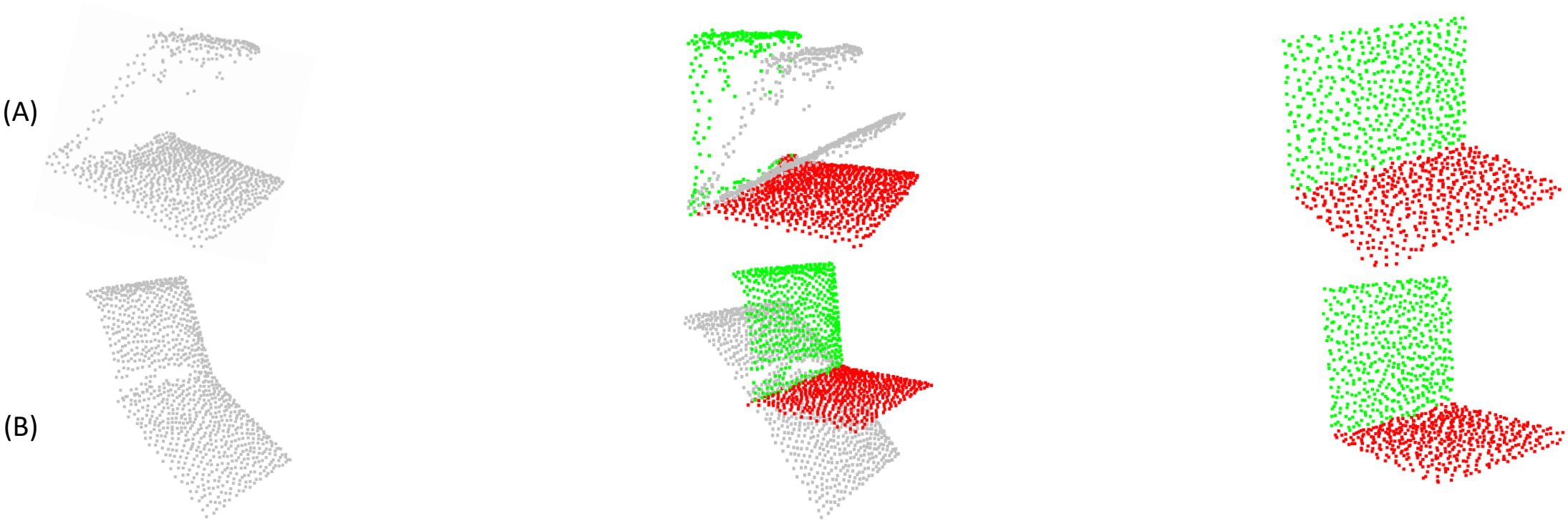
OP-Align: Object-level and Part-level Alignment & Reconstruction

- Construct a **canonical reconstruction** with a fixed pose and joint states among the category.
- Estimate **transformations** which aligns an input and this canonical reconstruction.



OP-Align: Object-level and Part-level Alignment & Reconstruction

- Construct a **canonical reconstruction** with a fixed pose and joint states among the category.
- Estimate **transformations** which aligns an input and this canonical reconstruction.

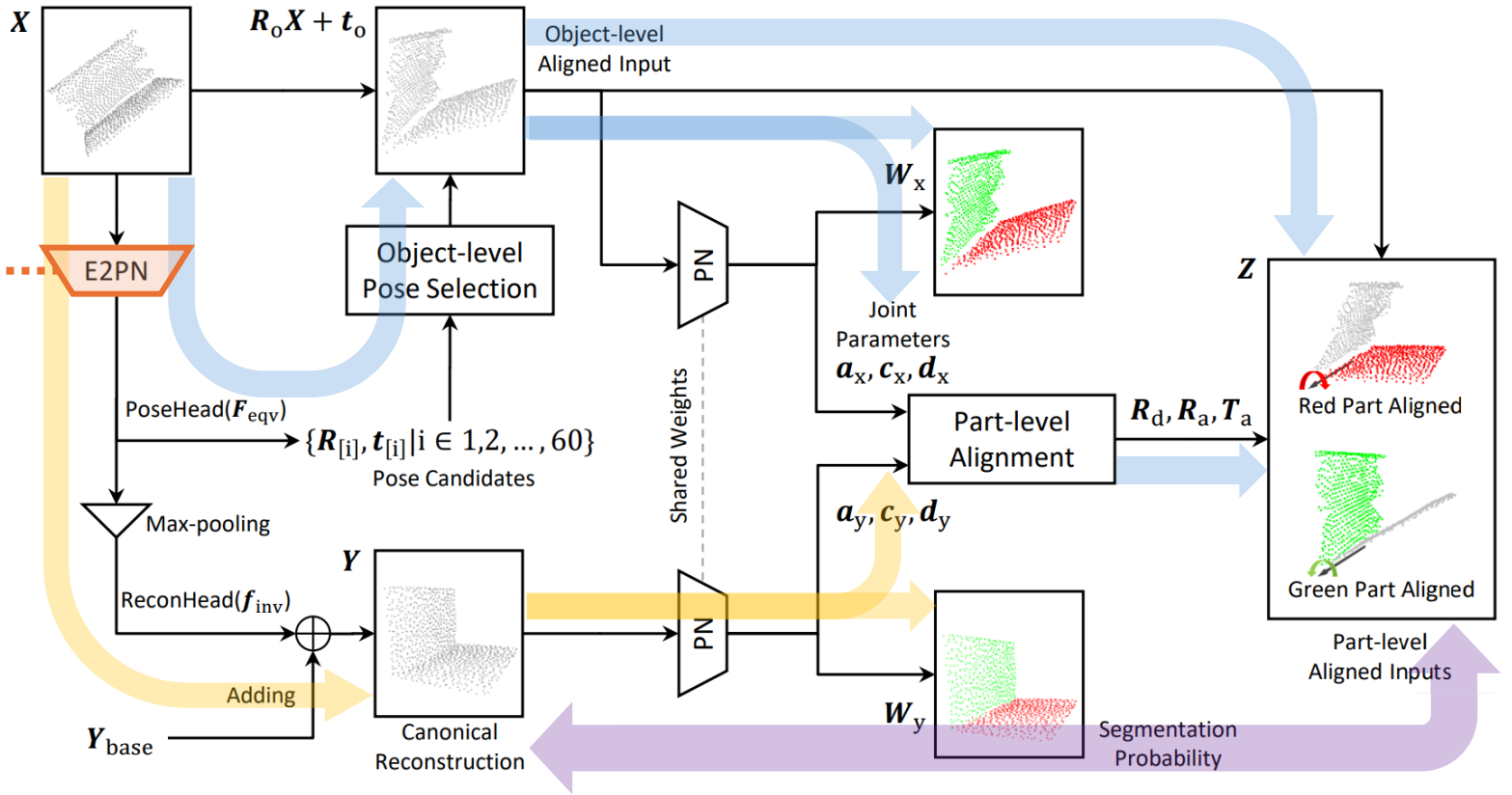


OP-Align: Network Architecture

- Two-phase network architecture
- Output relative transformation between the input and reconstruction

SE(3)-equivariance — **Object-level Phase** — **Part-level Phase**

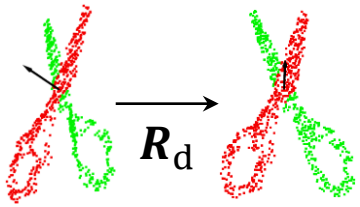
- **SE(3)-equivariant** pose estimation
- **SE(3)-invariant** reconstruction



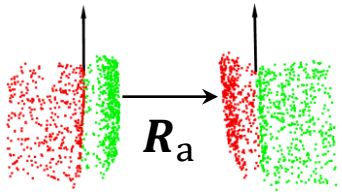
Part-level Alignment & Training

- Part-level alignment with predicted **joint parameters**
- Train such a model without annotation

Simulation of Joint Movement

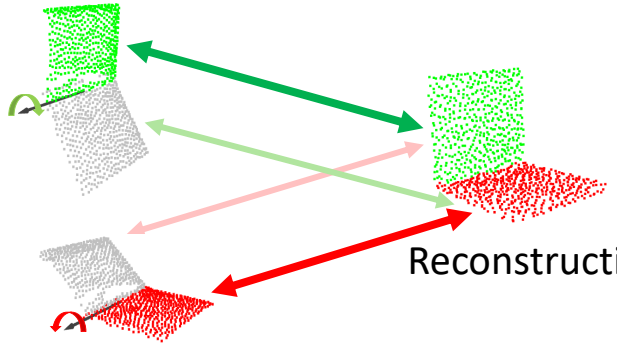


Joint Direction Alignment



Joint State Alignment (Revolute)

Weighted Chamfer Distance



Part-level aligned inputs

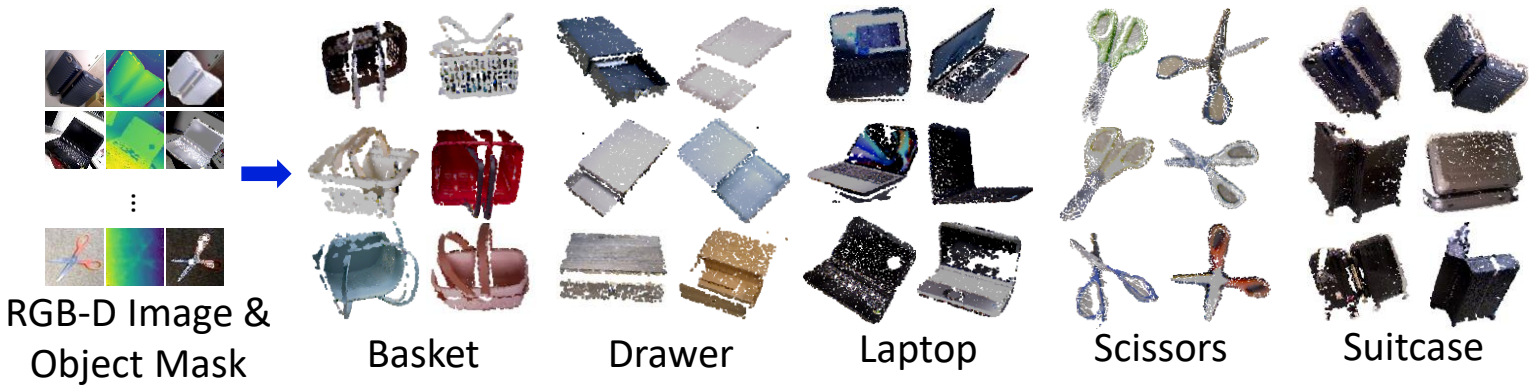
- Predicted joint direction, pivot, and states
- Each part of the input can be aligned with the **corresponding part** of the reconstruction

- Using **part segmentation probability** as point weight
- Combine part segmentation and joint movement

Real-world Dataset

- Articulated Objects from 5 categories, 4 training objects & 2 testing objects
- At least 8 different joint states for each object
- Object mask generated by Segment Anything Model or Mask-RCNN

Category	Training		Testing		Object			Detection
	Image	Instance	Image	Instance	Part	Joint(prismatic)	Joint(revolute)	
basket	974	4	449	2	3	0	2	SAM [17]
drawer	884	4	452	2	2	1	0	SAM [17]
laptop	740	4	412	2	2	0	1	Mask-RCNN [9]
scissors	922	4	421	2	2	0	1	Mask-RCNN [9]
suitcase	813	4	381	2	2	0	1	Mask-RCNN [9]



Inputs

- RGB-D
- Object Mask

Annotations

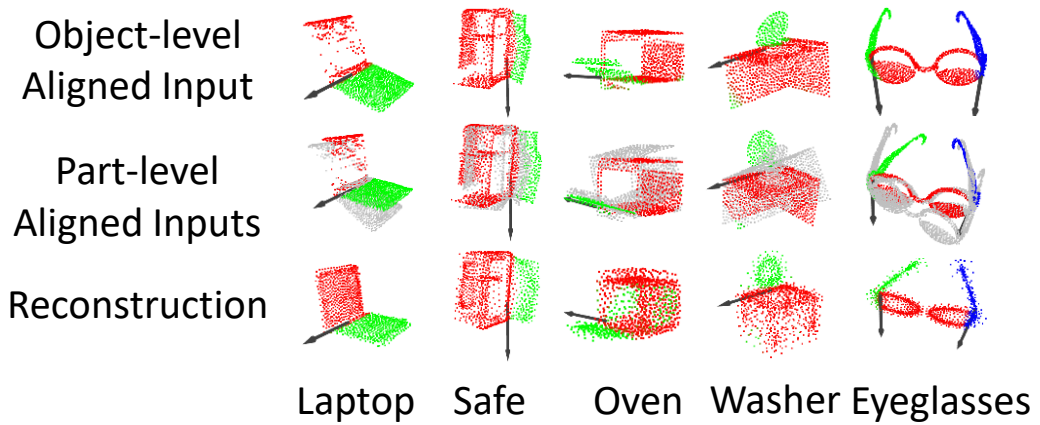
- Part Segmentation
- Part Rotation
- Part Translation
- Part Scale
- Joint Pivot
- Joint Direction

Experiments on the Synthetic Dataset

- **Partially observed point cloud**
  generated with random camera position

- State-of-the-art performance
- Real-time inference speed

Visualization



Mean Error Results

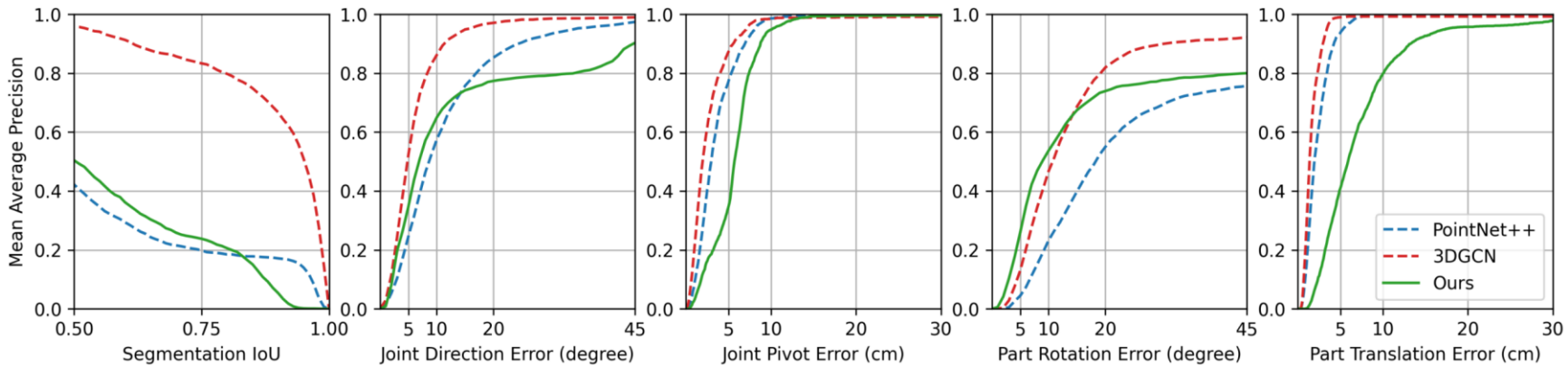
Method	Supervision			Segmentation	Rotation	Translation	Pivot	Direction	Memory	Speed
	Pose	Segmentation	Joint	IoU↑	(degree)↓	↓	↓	(degree)↓	(GB)↓	(FPS)↑
3DGCN [22]	✓	✓	✓	**94.05**	11.61	0.093	**0.084**	9.78	-	-
NPCS-EPN [20]	✓	✓	✓	-	11.05	**0.080**	0.147	15.20	-	-
ICP		✓		66.45	44.12	0.242	-	-	-	-
EAP [24]				68.46	10.44	0.121	0.162	23.09	9.23	<1
Ours				80.70	**8.10**	0.129	0.110	**6.63**	**2.31**	**41**

Experiments on the Real-world Dataset

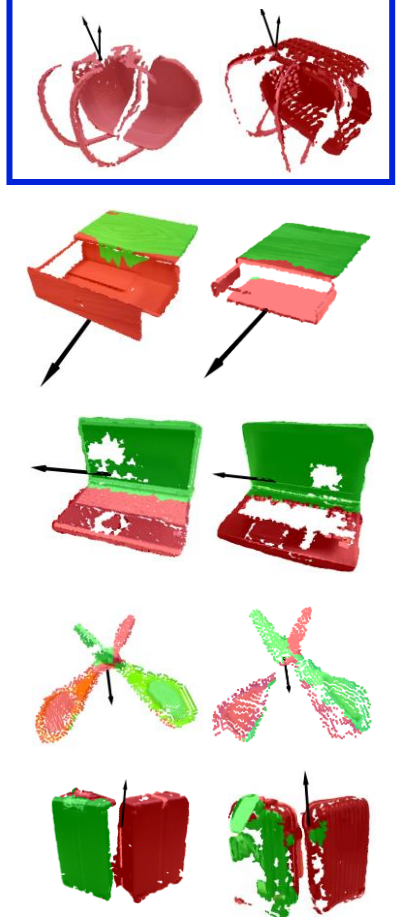
- Compared with other **supervised** methods
- Still have rooms for improvement

mAP(%) Results

Method	Supervision			Segmentation↑		Joint↑			Part↑		
	Pose	Segmentation	Joint	IoU75%	IoU50%	5°5cm	10°10cm	15°15cm	5°5cm	10°10cm	15°15cm
3DGCN [22]	✓	✓	✓	**83.31**	**95.83**	**47.51**	**85.79**	**94.59**	13.07	**46.77**	**68.66**
PointNet++ [33]	✓	✓	✓	19.83	42.20	21.06	57.38	75.56	4.47	23.25	39.82
Ours				23.79	50.42	12.57	63.59	74.04	**14.79**	46.09	59.76



Failure cases



Conclusion

A New Self-supervised Model

Focusing on articulated object pose estimation.

- Expand SE(3)-equivariant backbone usage range
- Simulation of joint movement
- Segmentation probability weighted chamfer distance

A New Real-world Dataset

Focusing on point cloud of articulated object with high-accuracy annotation.

- Multiple category / joint type
- Pose / joint states variance
- High accuracy annotations

Repository

Contact

cheyuchen.titech@gmai.com
rfurukaward@gmail.com
kanezaki@c.titech.ac.jp

Thank you for listening!

東京工業大学
Tokyo Institute of Technology
Automation & Knowledge Lab.