DexGraspNet 2.0: Learning Generative Dexterous Grasping in Large-Scale Synthetic Cluttered Scenes

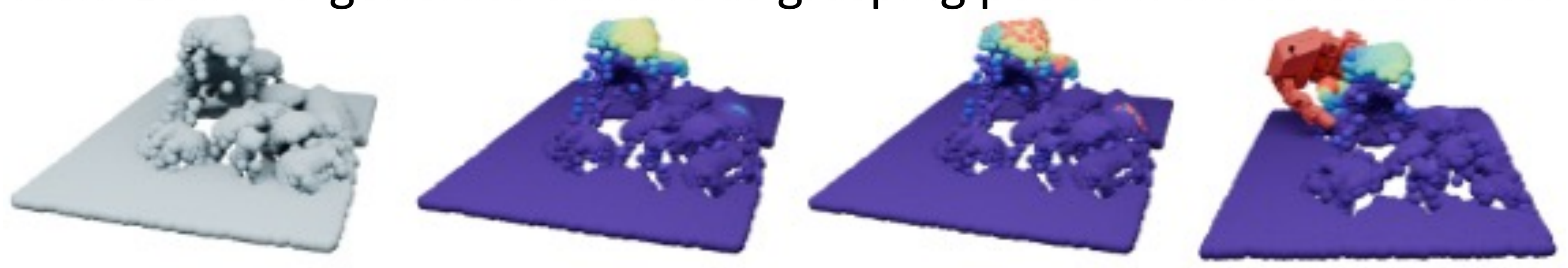
Jialiang Zhang[1,2]*, Haoran Liu[1,2]*, Danshi Li[2]*, Xinqiang Yu[2]*, Haoran Geng[1,2], Yufei Ding[1,2], Jiayi Chen[1,2], He Wang[1,2,3]†

[1]CFCS, School of CS, Peking University [2]Galbot [3]Beijing Academy of Artificial Intelligence

Videos available

Introduction

➢ **Task**: Learn to generate dexterous grasping poses in cluttered scenes.



① ② ③ ④

①Input: single view scene point cloud ②Predict per-point graspness

③sampled seed points ④generate grasps and select the best

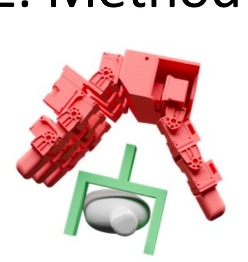
⑤Our method generates diverse grasp poses in large scale with diverse scenes

➢ **Motivation:**

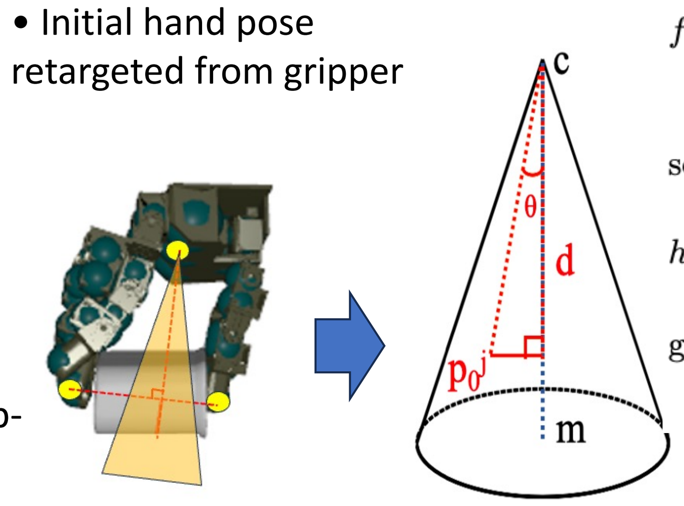
- Grasping in cluttered scenes remains highly challenging for dexterous hands due to the scarcity of **data**
- Existing grasp pose detection methods are data-inefficient without utilization of **local geometric features**
- It is unclear how does **scaling the dataset** impact grasping generation performance

Main contribution

1. Large-scale Benchmark for dexterous grasping in cluttered scenes
   - 7500 synthetic training scenes composed with 60 GrsapNet-1Billion objects annotated with 426M dexterous grasp poses
   - 670 test scenes encompassing 1319 objects and varied cluttery
2. Method



- Initial hand pose retargeted from gripper
- graspness cone retargeted from palm center and midpoint of thumb-middlefinger tip



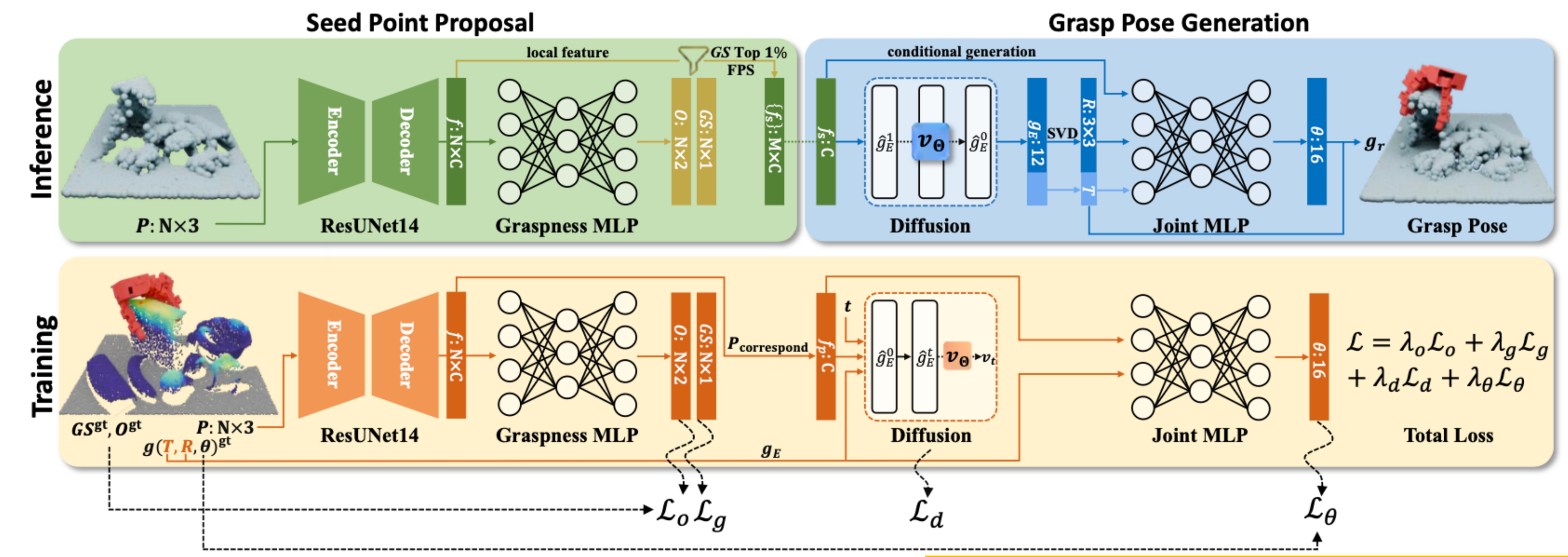
$$f(g_o^i, p_o^j) = \begin{cases} 0 & p_o^j \notin \text{this cone} \\ \exp\left(-\frac{\ln 2}{10}\frac{180}{\pi}\theta - \frac{\ln 2}{0.015}d\right) & p_o^j \in \text{this cone} \end{cases} \quad (2)$$

$$\text{seed\_point}(g_o^i) = \underset{p_o^j \in P_o}{\arg\max} f(g_o^i, p_o^j) \quad (3)$$

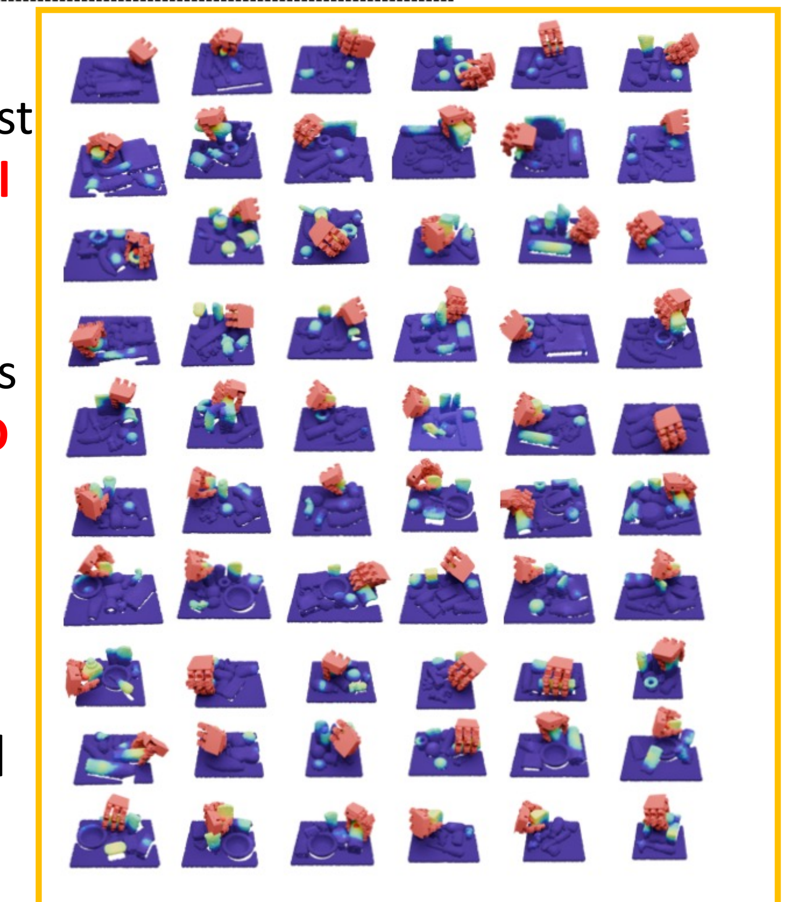
$$h(g_o^i, p_o^j) = 10^{-150||\text{seed\_point}(g_o^i) - p_o^j||_2} \quad (4)$$

$$\text{graspness\_score}(p_o^j) = \ln\left(0.001 + \sum_{g_o^i \in G_o} h(g_o^i, p_o^j)\right) \quad (5)$$

- Redefine **graspness** as likelihood of a valid grasp exists near the point



- Propose an end-to-end generative grasp prediction pipeline that first generates wrist **6D pose via diffusion conditioned on local geometric feature**, and **regress hand joint poses conditioned on wrist pose**
- We ablate among different representations of rotation and find the **9D Procrustes SVD** works best with diffusion model.



Predicted graspness and best grasp

3. We investigate the effect of **scaling number of grasps / number of scenes used in training set**, and find our model is surprisingly robust to low-data training, meanwhile scales better when the data scales up.

4. We test our model in the **real-world** and achieve **90.7%** success rate in decluttering very challenging scenes. With the help of point cloud restoration, our model can tackle scenarios with mixed diffuse-and-transparent objects.

Experiment

1. **Simulation**

Method	GraspNet-1Billion		ShapeNet	
	Dense	Loose	Dense	Loose
HGC-Net [10]	46.0	26.7	46.4	30.4
GraspTTA† [7]	62.5	42.8	56.6	46.4
ISAGrasp† [4]	63.4	51.4	64.0	52.7
Ours	**90.6**	**73.2**	**81.0**	**74.2**

Dense: 8-11 objects per scene
Loose: 1-2 objects per scene

Random: 1-10 objects per scene, obtained by randomly masking objects from Dense scenes

2. **Ablation**

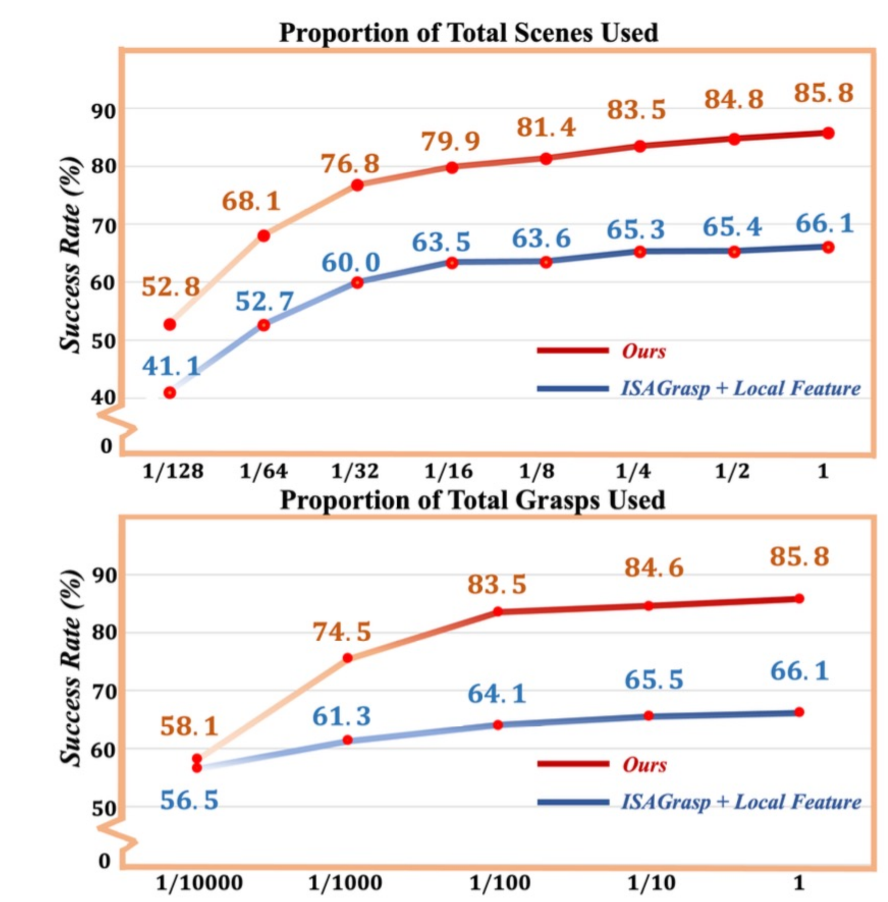
Method	GraspNet-1Billion			ShapeNet		
	Dense	Random	Loose	Dense	Random	Loose
local feature	16.8	10.9	4.8	21.3	17.6	10.9
decomposed model	84.2	80.7	71.3	74.9	72.5	66.4
random scene	90.0	**84.1**	68.2	78.9	78.8	71.3
Ours	**90.6**	83.7	**73.2**	**81.0**	**85.4**	**74.2**

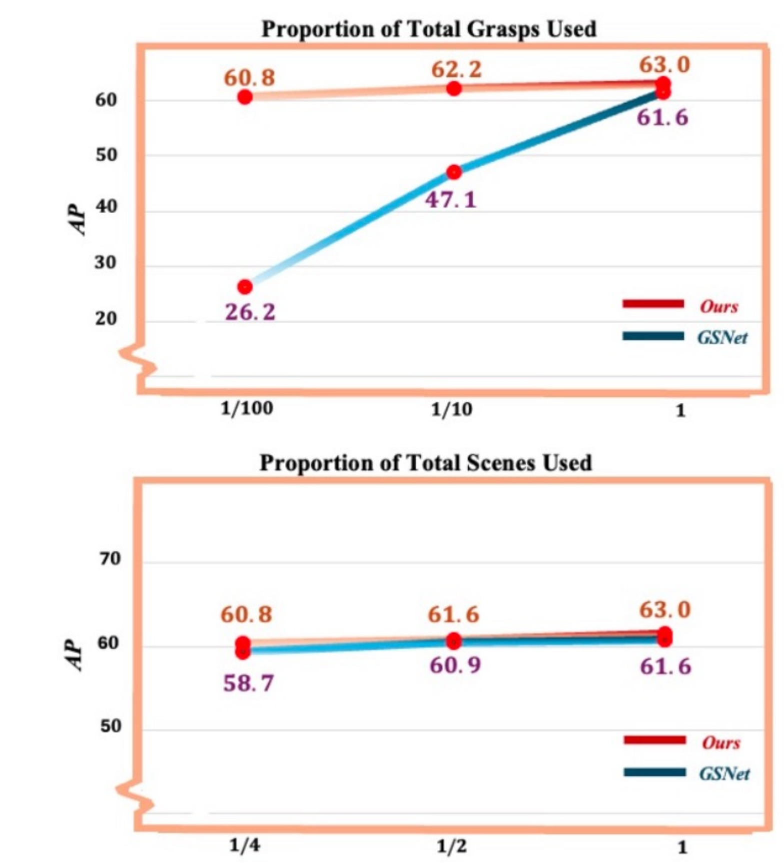
3. **Representation of rotation**

	Method	GraspNet-1Billion			ShapeNet		
		Dense	Random	Loose	Dense	Random	Loose
Ablation	Euler Angle	87.6	82.0	73.0	78.0	76.4	**75.2**
	Axis Angle	86.4	81.7	70.5	79.0	76.4	74.1
	Quaternion	87.9	81.5	72.0	78.6	77.0	72.9
	6D	88.2	81.5	71.9	80.2	79.0	73.0
	Ours	**90.6**	**83.7**	**73.2**	**81.0**	**85.4**	74.2

We empirically find **9D Procrustes SVD** works best with our diffusion pipeline

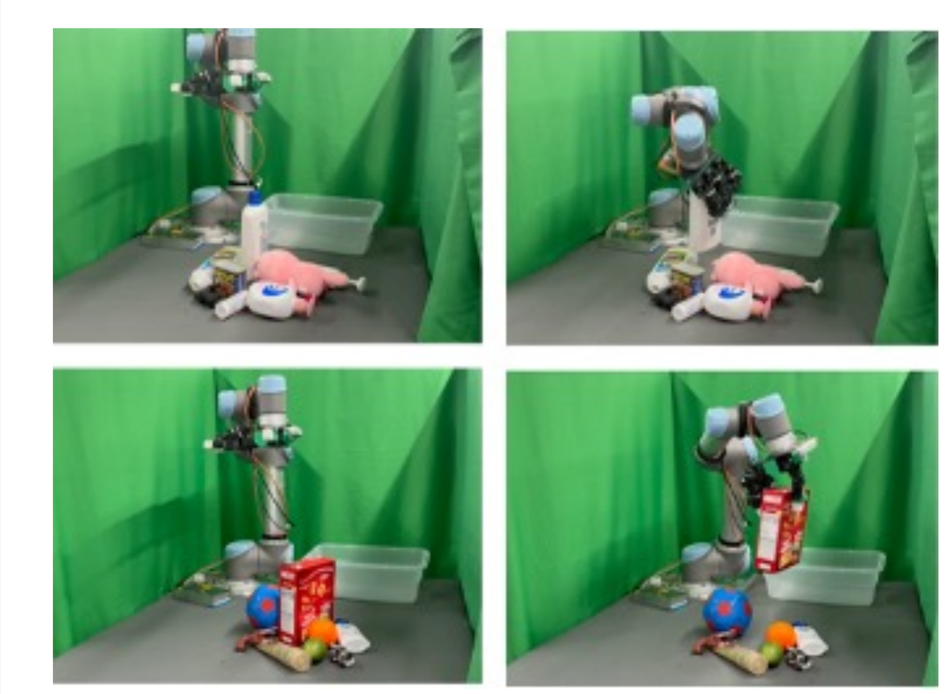
4. **Scaling the training data**

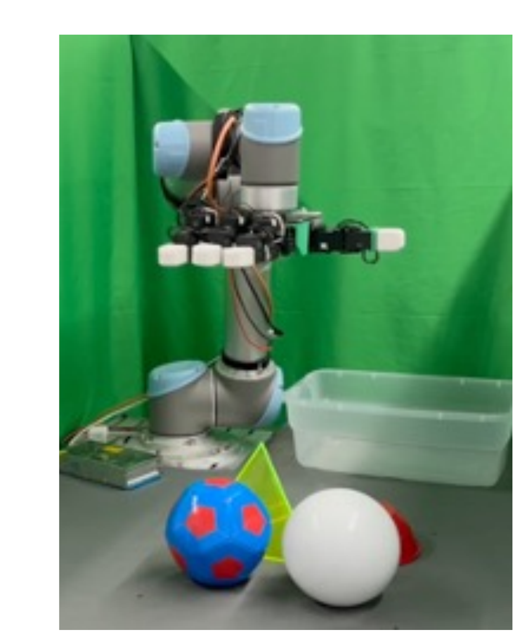




Fraction of Grasps	Success Rate
1/100(42k)	81.3
1(4.2M)	92.4

5. **Real-world experiments**





Method	HGC-Net	Ours
SR (%)	16.44	**90.70**

End Effector	Normal	Large
Parallel Gripper	92.4	0.0
Dexterous Hand	81.5	100.0

- 90.7% success rate in cluttered real scenes
- capable to work in scenarios where grippers fail