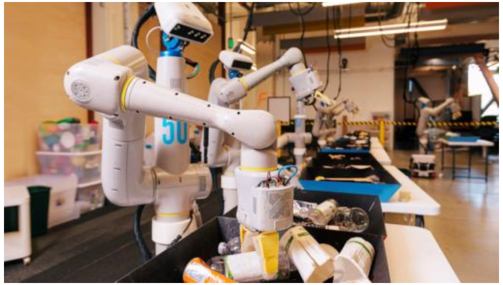
Synergies Between Affordance and Geometry: 6-DoF Grasp Detection via Implicit Representations

Presenter: Cheng-Chun Hsu

9/8/2022

Robotic Grasping

- Modules in robot manipulation
  - Bin picking
  - Part assembly
  - Logistics



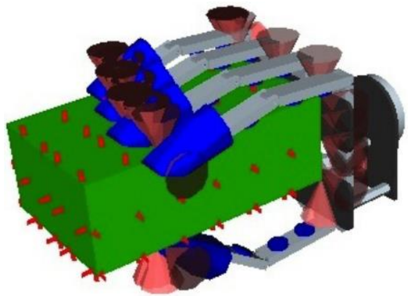
Robotic Grasping

- Geometric vs. data-driven
- Object model: known vs. unknown
- Sensor data:
  - Single-view vs. multi-view
- Open-loop vs. closed-loop
- Human-supervised vs. self-supervised

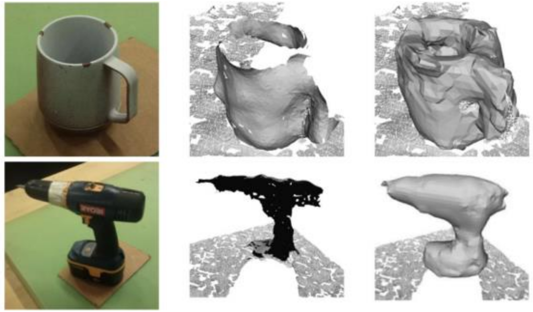
Robotic Grasping

- Geometric vs. data-driven
- Object model: known vs. unknown
- Sensor data:
    - Single-view vs. multi-view
- Open-loop vs. closed-loop
- Human-supervised vs. self-supervised

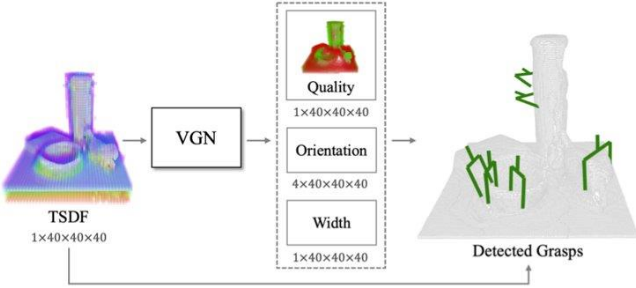
Prior work



[Miller et al. 2003, Goldfeder et al. 2007, Hübner et al. 2008, Diankov et al. 2008]



[Bohg et al. 2011, Varley et al. 2017, Lundell et al. 2019]



[Mahler et al. 2017, Morrison et al. 2018, Liang et al. 2019, Breyer et al. 2020]

Geometry Analysis

- Analytical solution
- Require full 3D model

Reconstruction → Grasp Synthesis

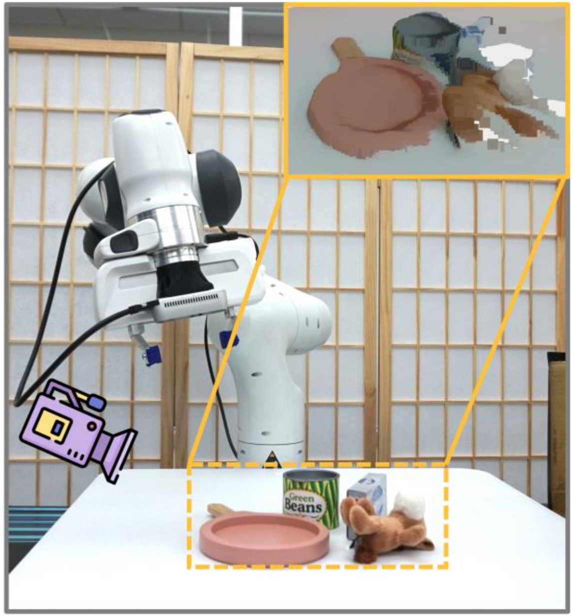
- Operate on raw visual observation
- Subject to 3D reconstruction quality

End-to-end Deep Learning

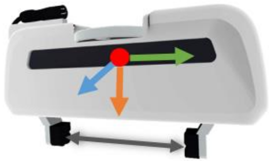
- High grasp performance
- No explicit geometry reasoning

Problem Formulation

Input: partial point cloud



Output: 6-DoF grasp pose



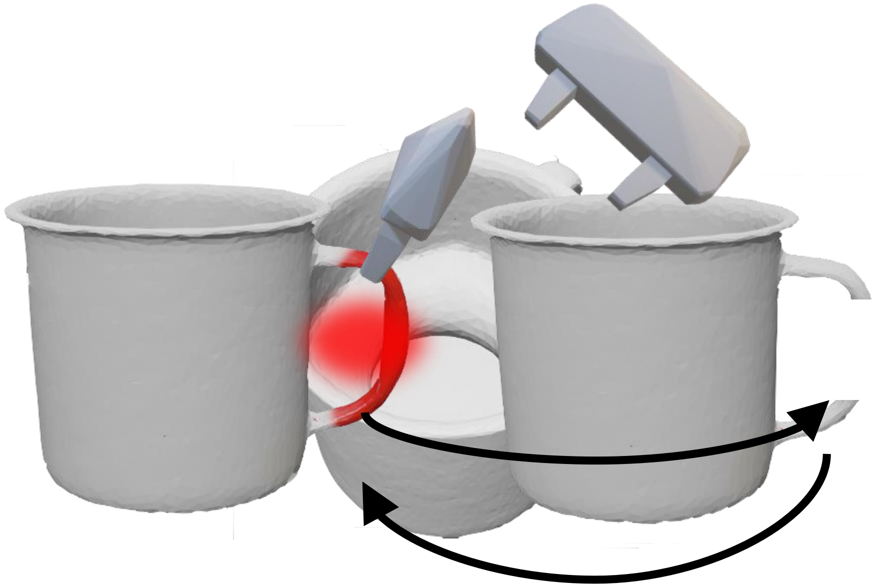
$t \in \mathbb{R}^3$	Grasp center
$\omega \in [0, \omega_{max}]$	Grasp width
$r \in SO(3)$	Gripper rotation
$q \in [0, 1]$	Grasp quality

Key Idea

Affordance and geometry reasoning are not isolated

Affordance

Predict affordance of reconstructed part



Geometry

Reconstruct graspable region

Implicit Neural Representations

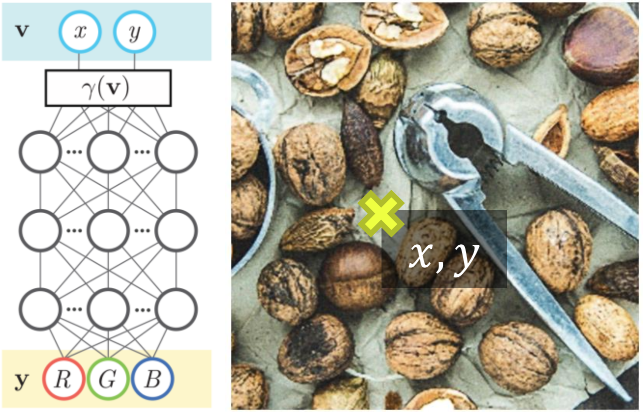
A mapping function
from spatial coordinates to values

$$f : \mathbb{R}^n \rightarrow \mathcal{Y}$$

Sometimes conditioned on additional input

$$f : \mathbb{R}^n \times \mathcal{X} \rightarrow \mathcal{Y}$$

E.g., x-y coordinate → RGB value



[Tancik et al. 2020]

Implicit Neural Representations

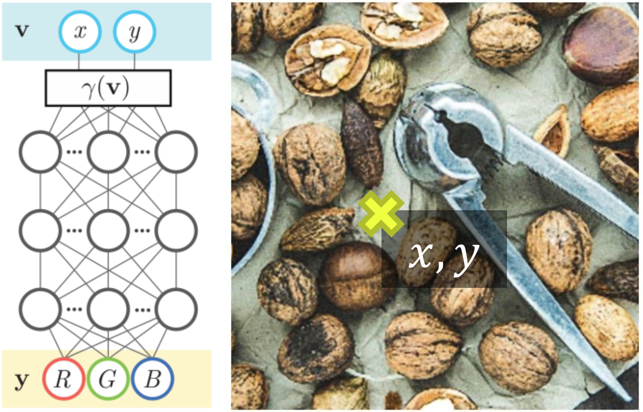
A mapping function
from spatial coordinates to values

$$f : \mathbb{R}^n \to \mathcal{Y}$$

It can also be conditioned on additional input

$$f : \mathbb{R}^n \times \mathcal{X} \to \mathcal{Y}$$

E.g., x-y coordinate → RGB value



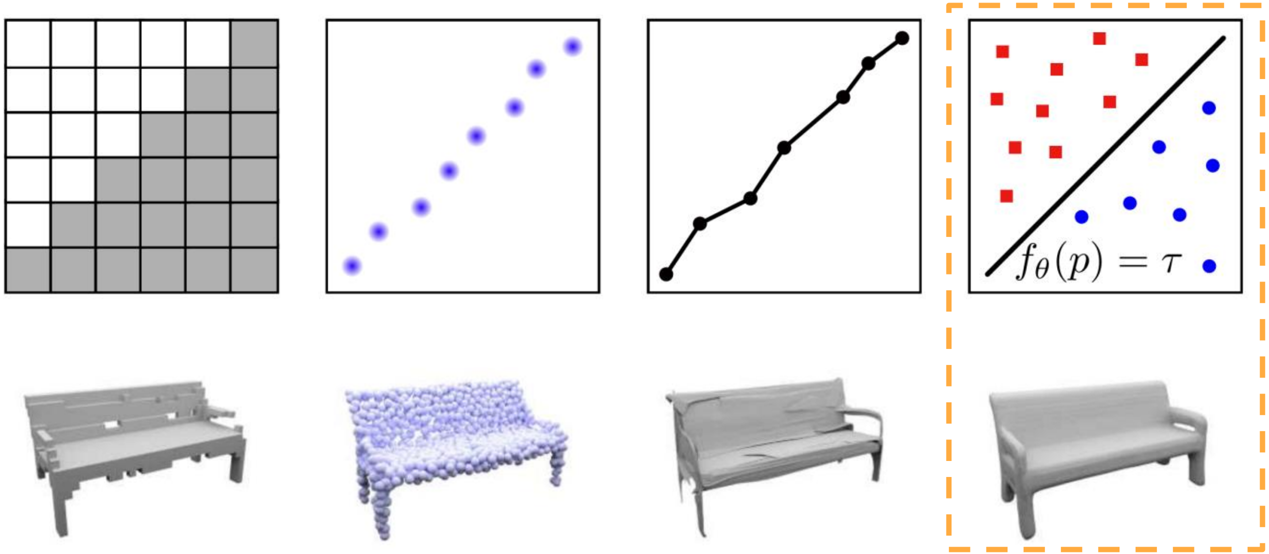
[Tancik et al. 2020]

Advantages:

- Continuous and memory-efficient
- End-to-end differentiable
- Adaptively allocate representation resources

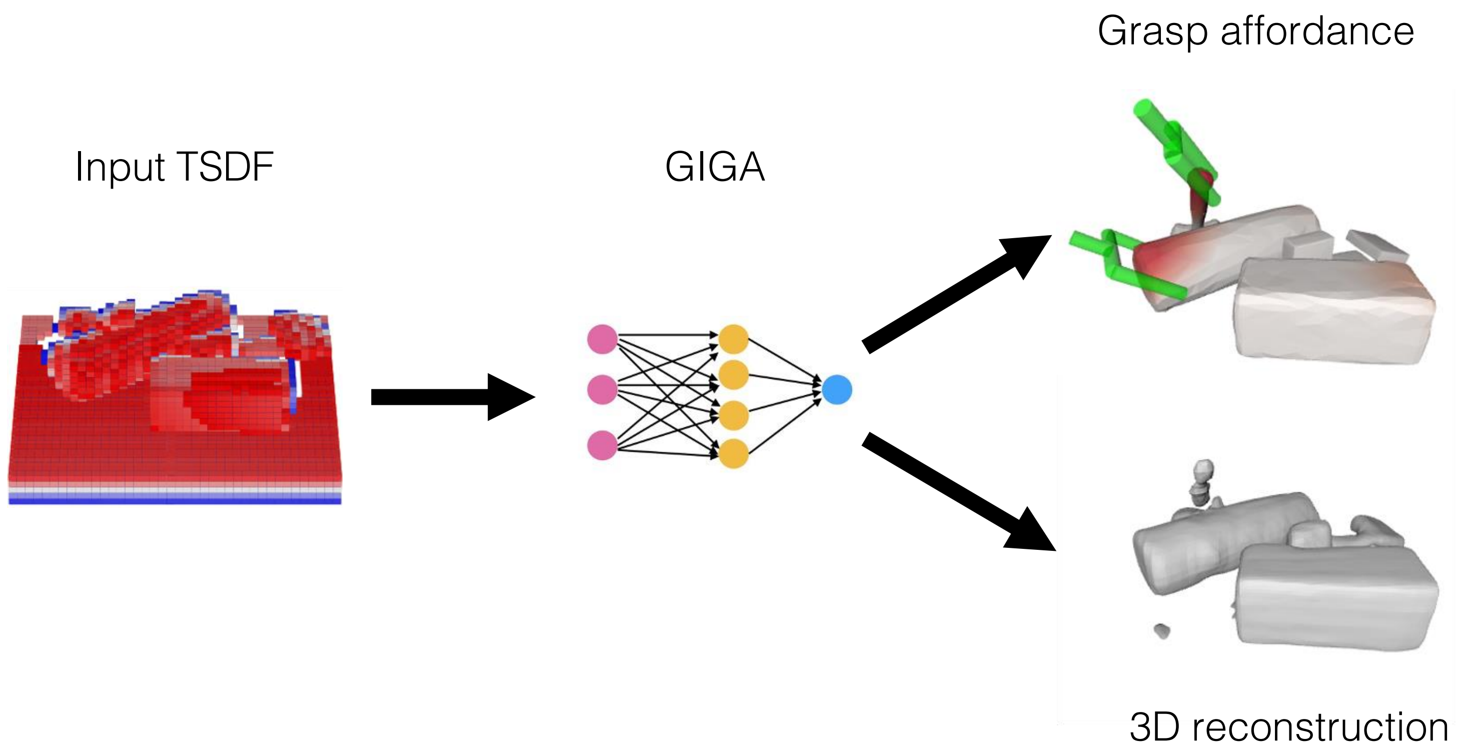
Implicit Neural Representations

Occupancy Network [Mescheder et al. 2019] maps 3D coordinates to occupancy in 3D reconstruction



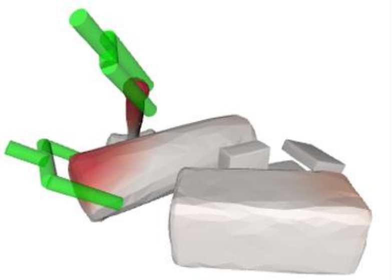
Voxel Point cloud Mesh Implicit occupancy function

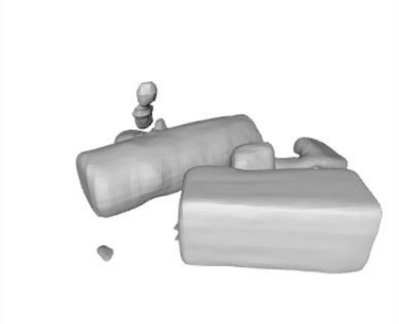
Approach



Approach

Grasp affordance





3D reconstruction

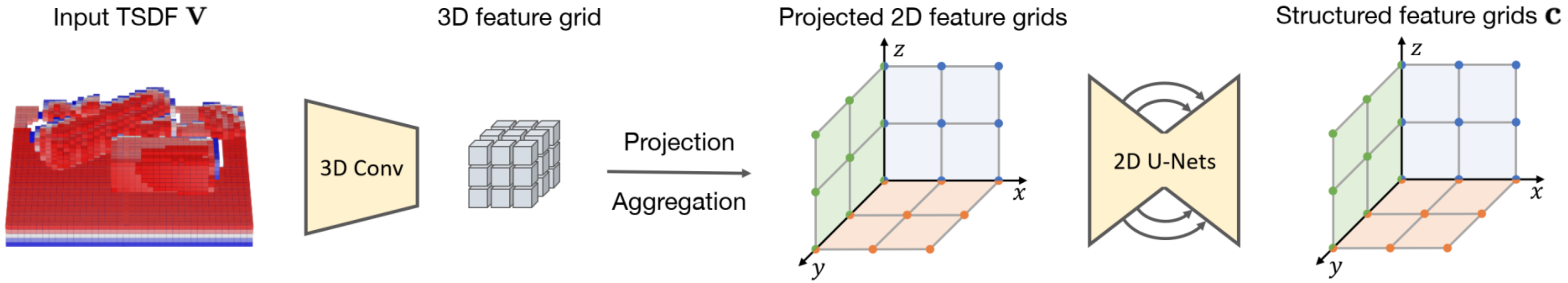
$$f_a : \mathbf{t} \to q, \mathbf{r}, w,$$

3D location (Grasp center) — Grasp quality, rotation, width

$$f_g : \mathbf{p} \to b.$$

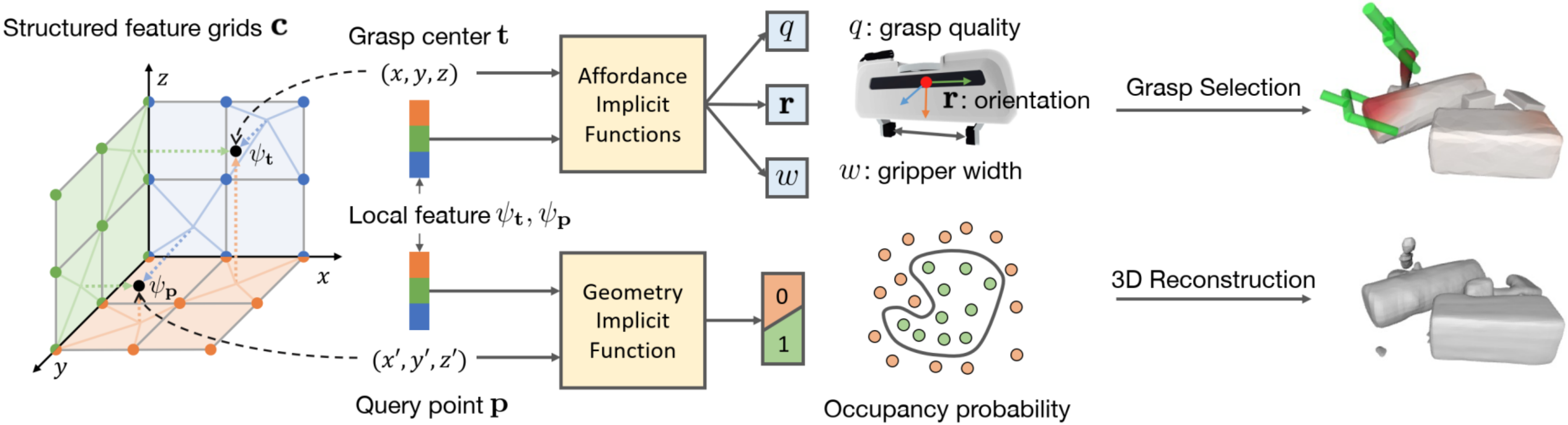
3D location (Any) — Occupancy

Approach



Input TSDF $\mathbf{V}$
3D Conv
3D feature grid
Projection
Aggregation
Projected 2D feature grids
z
x
y
2D U-Nets
Structured feature grids $\mathbf{c}$
z
x
y

Approach



Structured feature grids $\mathbf{c}$
z
x
y
$\psi_{\mathbf{t}}$
$\psi_{\mathbf{p}}$
Grasp center $\mathbf{t}$
(x, y, z)
Local feature $\psi_{\mathbf{t}}, \psi_{\mathbf{p}}$
(x', y', z')
Query point $\mathbf{p}$
Affordance Implicit Functions
Geometry Implicit Function
q
$\mathbf{r}$
w
0
1
q: grasp quality
$\mathbf{r}$: orientation
w: gripper width
Occupancy probability
Grasp Selection
3D Reconstruction

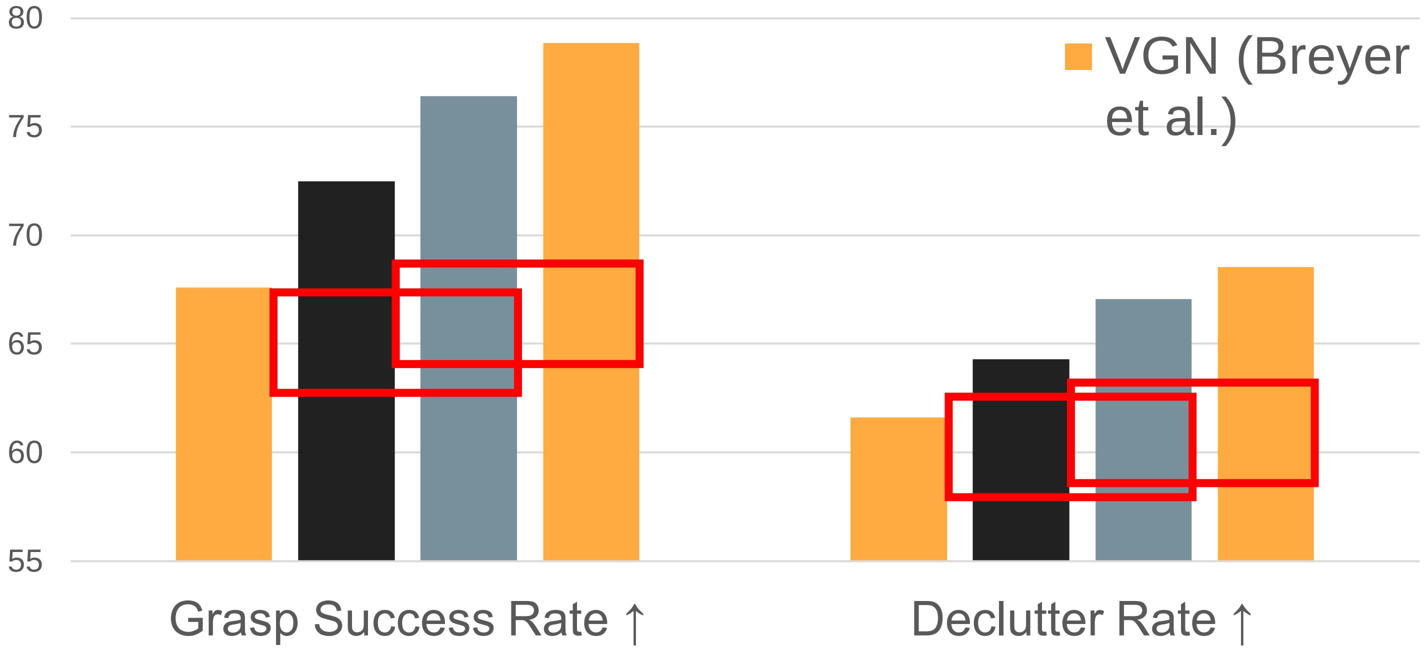
Experimental Setup - Scenarios

Packed objects (more occlusion)

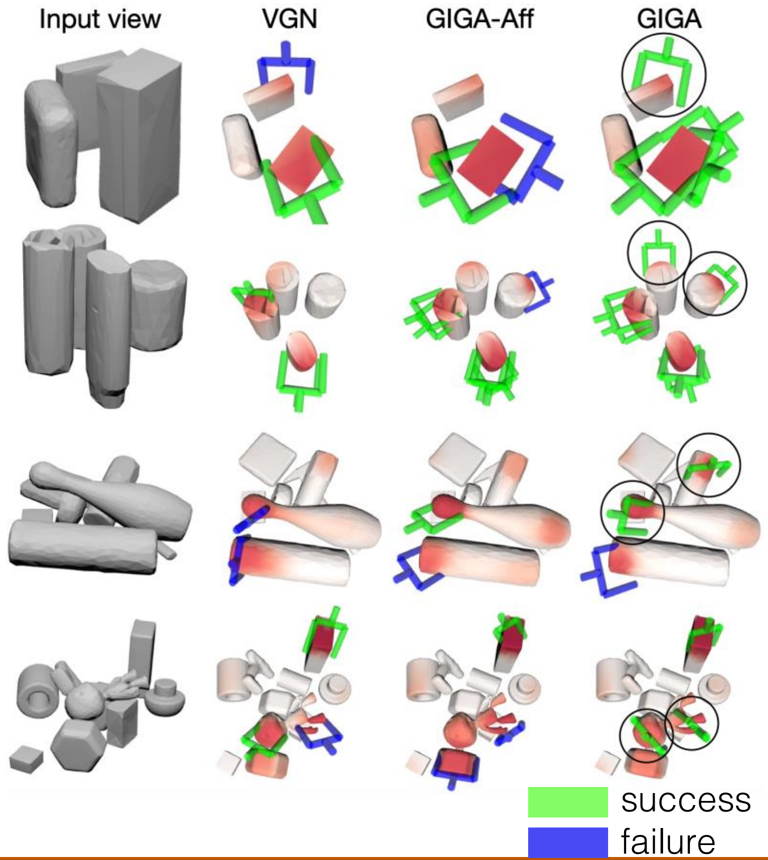
Piled objects (less occlusion)

Quantitative Comparison

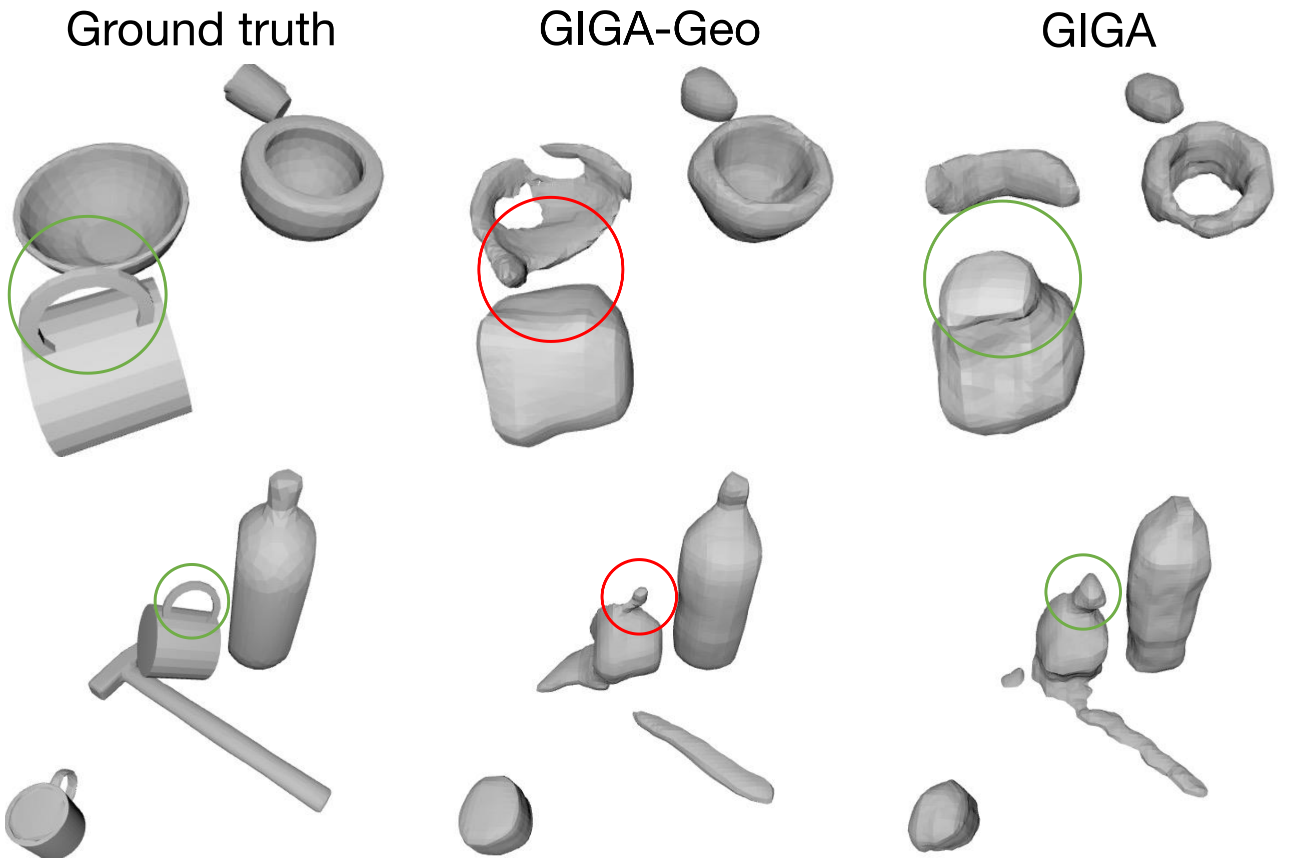
- Geometry learning facilitates affordance learning
- Continuity of implicit function enables higher precision



Geometry Learning Facilitates Occluded Grasps



Reconstruction Focuses on Graspable Parts



Strengths

- Not require known object models or multiple views
- Deal with cluttered and occluded scenes
- Continuous and compact representation

Weaknesses

- 3D reconstruction is only used as an auxiliary task during training
  - Reconstructed 3D information can be used for test-time optimization or closed-loop control
- GIGA relies on several assumptions
  - Single fixed viewpoint – what about a mobile robot?
  - Static scene and object
- Unrealistic real-world scenario
  - Evaluated only on tabletop scenarios

Future Directions

- Explore the potential of reconstructed 3D information
- Extended to a mobile robot
- Tested in more varied environments

Summary

GitHub Page

- Synergies between affordance and geometry
  - Better grasp prediction, especially in occluded regions
  - 3D reconstruction focuses on action-relevant parts
- Structured implicit neural representation
  - Continuous and compact representation for both affordance and geometry
  - Combine voxel grids with neural implicit functions