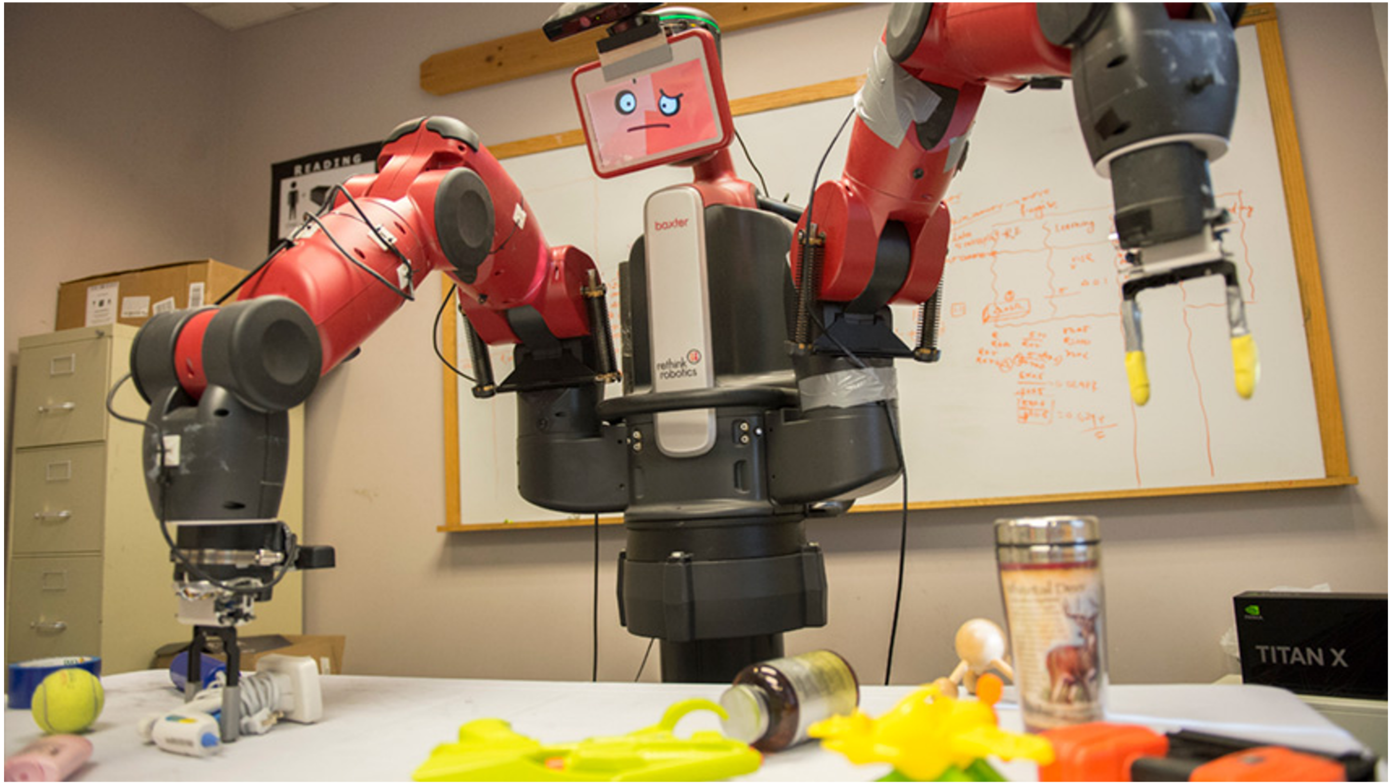
Synergies Between Affordance and Geometry: 6-DoF Grasp Detection via Implicit Representations

Presenter: Kevin Black

9/9/2021

Robot Grasping



Robot Grasping: Considerations

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

Robot Grasping: Considerations

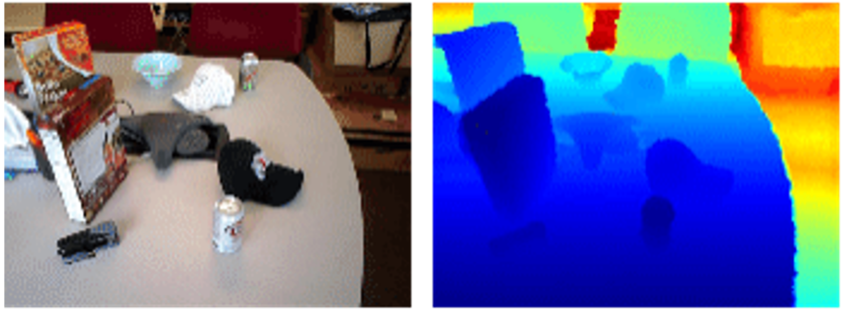
- Geometric vs. **data-driven**
- Object model: known vs. **unknown**
- Sensor data:
  - RGB vs. **RGB-D**
  - **Single-view** vs. multi-view
- **Open-loop** vs. closed-loop
- Human-supervised vs. **self-supervised**

Prior Work

- Efficient grasping from RGBD images: Learning using a new rectangle representation (ICRA 2011)
  - Introduces the hand-labeled Cornell grasping dataset, which became widely used
- Dex-Net 2.0: Deep Learning to Plan Robust Grasps with Synthetic Point Clouds and Analytic Grasp Metrics (RSS 2017)
  - A continuing line of work that has explored multiple gipper types (e.g. suction)
- Deep Object Pose Estimation for Semantic Robotic Grasping of Household Objects (CoRL 2018)
  - Performs pose estimation on known objects in order to geometrically compute grasps
- Grasping in the Wild: Learning 6DoF Closed-Loop Grasping from Low-Cost Demonstrations (ICRA 2020)
  - Learns from real human demonstrations, uses reinforcement learning for closed-loop feedback

Problem Formulation

Input: single-view RGB-D image



[Fu et al. 2015]

Output: grasp affordance

$\boldsymbol{t} \in \mathbb{R}^3$	Grasp center
$w \in [0, w_{max}]$	Gripper width
$\boldsymbol{r} \in SO(3)$	Gripper rotation
$q \in [0, 1]$	Grasp quality (success probability)

Performance measured using grasp success rate

Target Scenarios

Packed objects (more occlusion)

Piled objects (less occlusion)

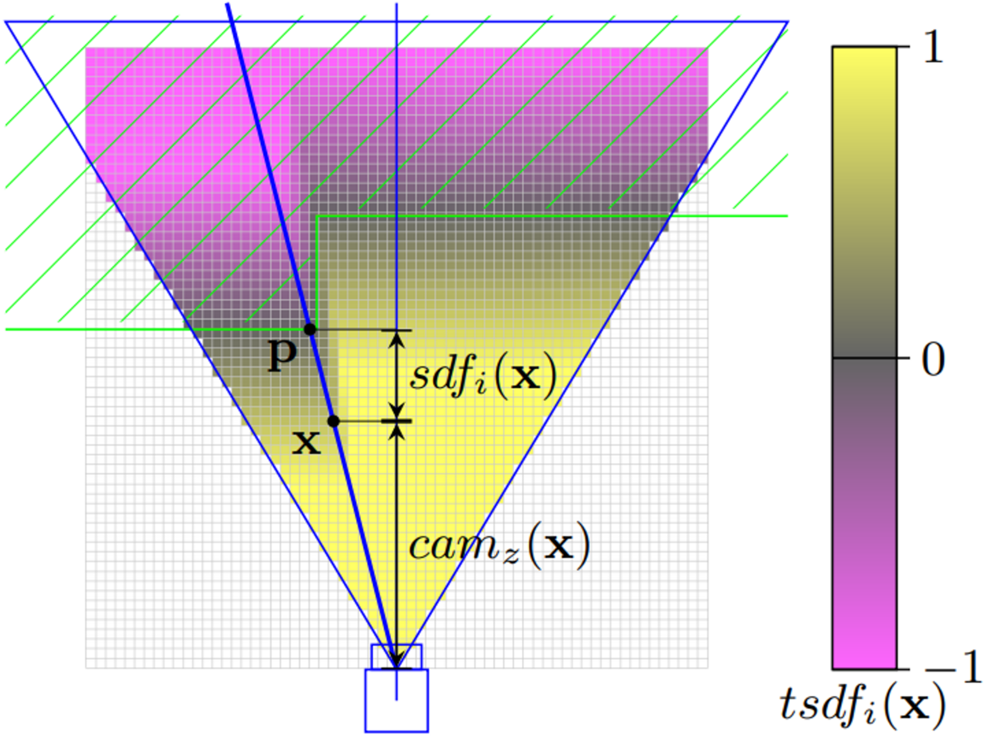
Primary Contribution

Grasp Detection via Implicit Geometry and Affordance (GIGA)

- Takes a single side view RGB-D image as input
- Network architecture which utilizes implicit neural representations
- Also learns 3D reconstruction as an auxiliary task

- Key insight: learning grasp detection and 3D reconstruction in tandem will synergize to improve overall performance in both tasks
- 3D reconstruction will especially help to reason about occluded regions

Truncated Signed Distance Function (TSDF)



[Werner et al. 2014]

Implicit Neural Representations

Represent data as a function from spatial coordinates to values, i.e.

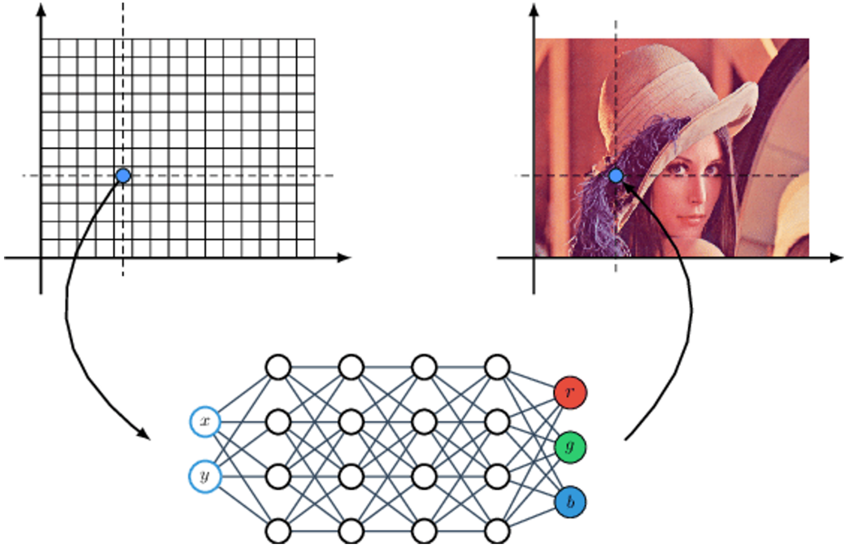
$$f : \mathbb{R}^n \to \mathcal{Y}$$

Usually conditioned on additional input:

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

Advantages:

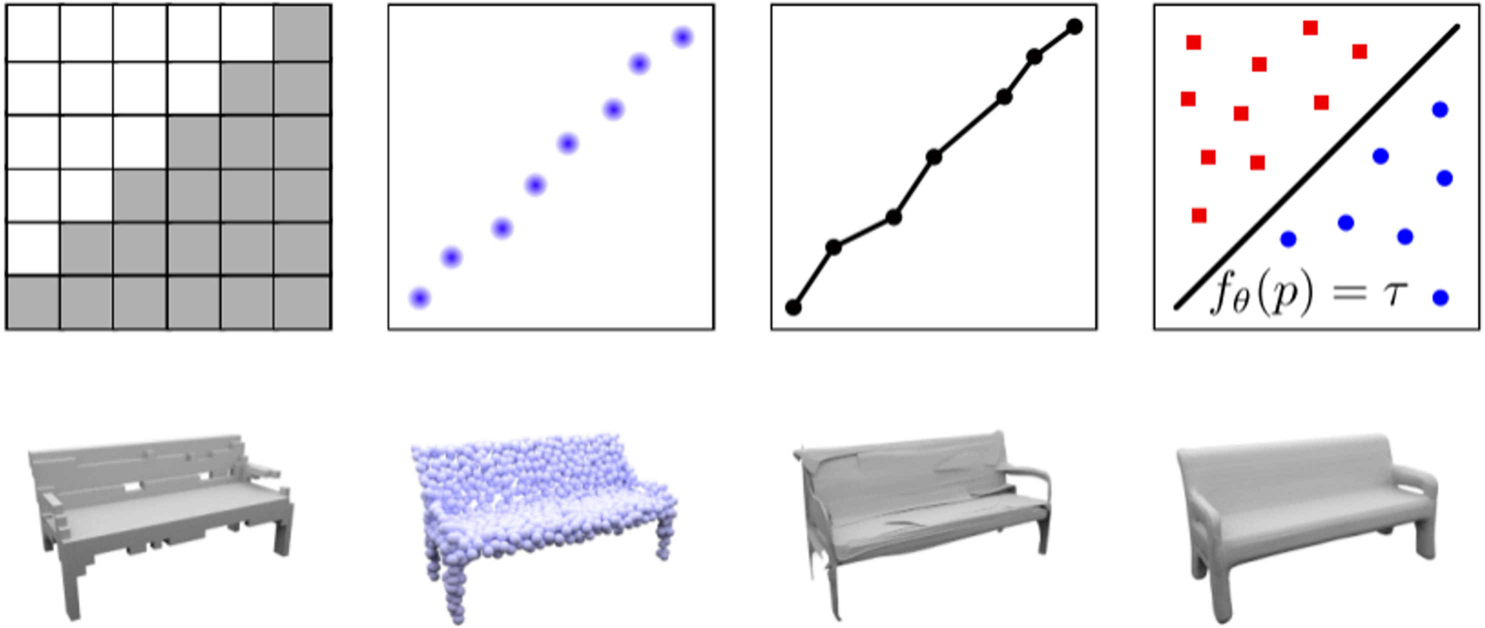
- "Infinite resolution" (no discretization)
- Memory-efficient



[Skorokhodov et al. 2020]

Implicit Representations for 3D Reconstruction

[Mescheder et al. 2019]



Voxels

Point cloud

Mesh

Implicit occupancy function

GIGA: Grasp Detection + 3D Reconstruction

Candidate grasp center
(3D point)

Grasp quality, rotation, width

Grasp detection: $f_a : \boldsymbol{t}, \boldsymbol{x} \rightarrow \overbrace{q, \boldsymbol{r}, w}$

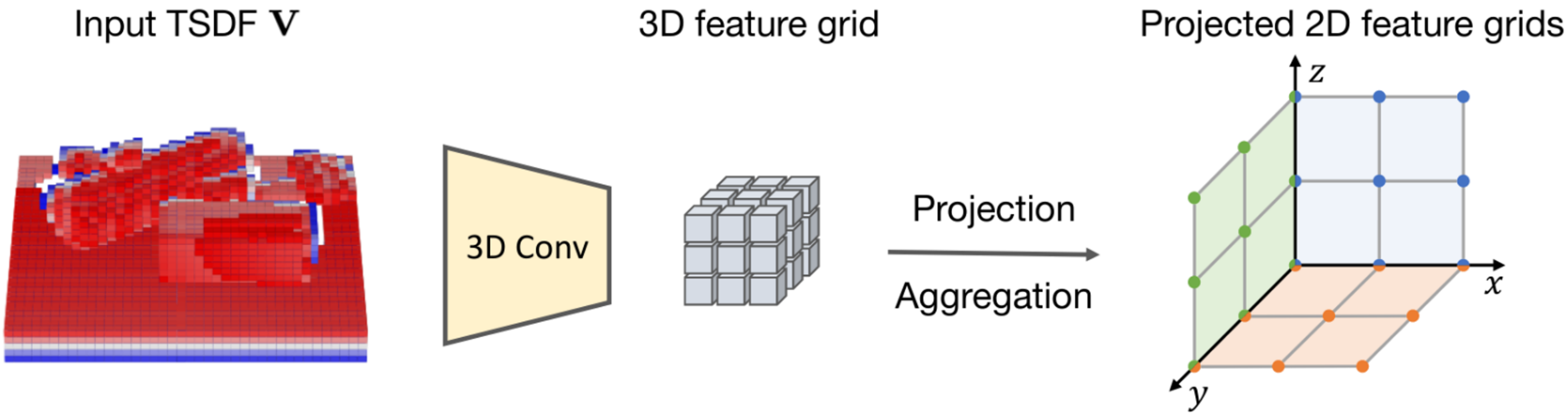
Input scene

3D reconstruction: $f_g : \boldsymbol{p}, \boldsymbol{x} \rightarrow b$

Any 3D point

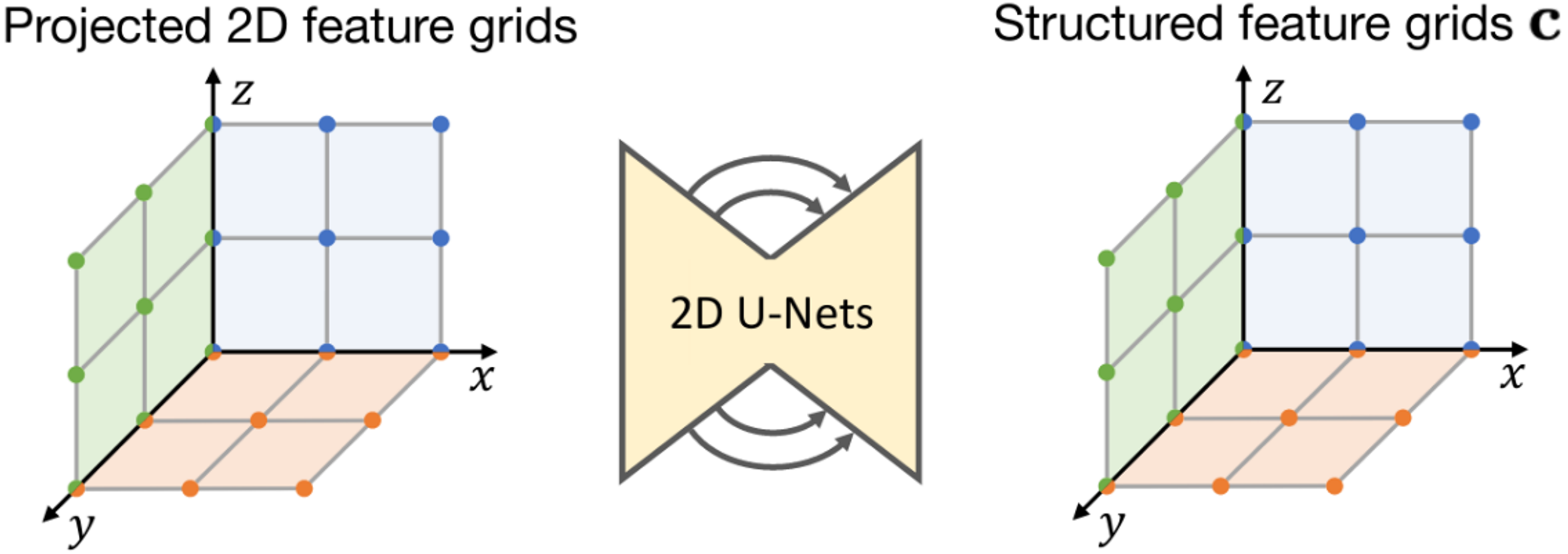
Occupancy value

GIGA Architecture



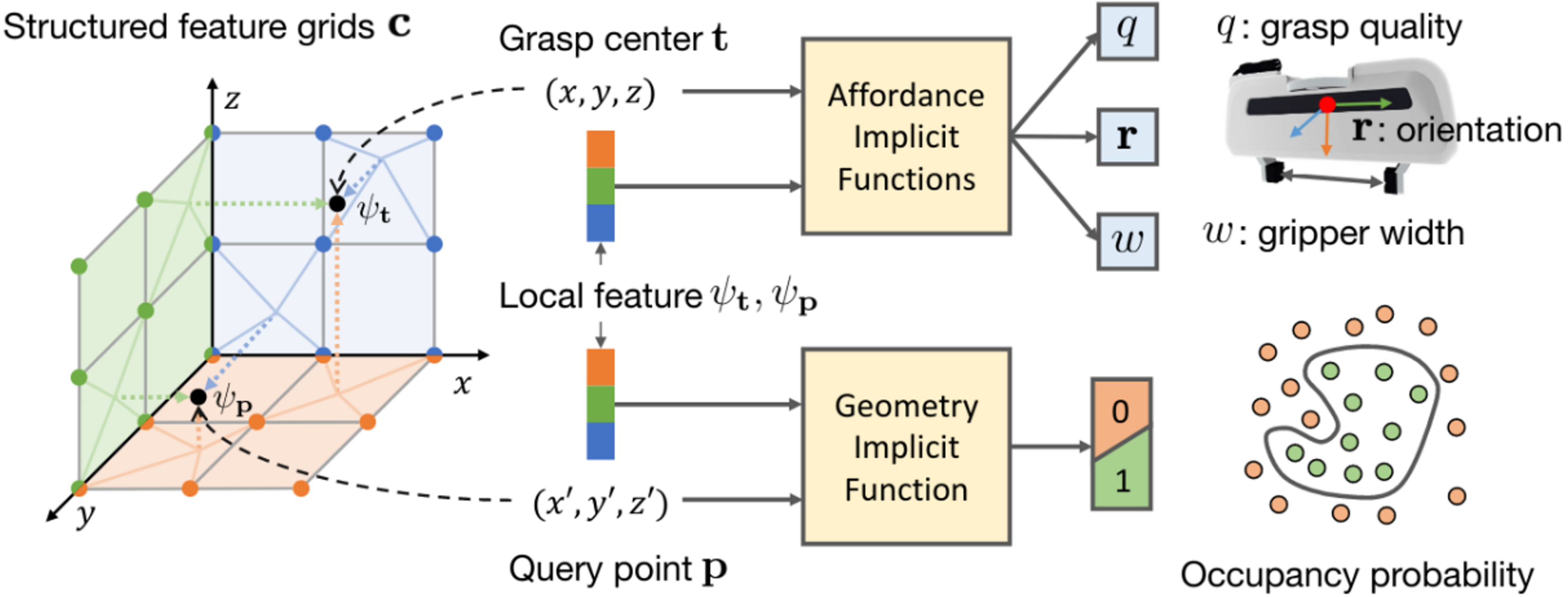
*Backbone architecture largely based on prior work: Convolutional Occupancy Networks (2020) by Peng et al.

GIGA Architecture



*Backbone architecture largely based on prior work: Convolutional Occupancy Networks (2020) by Peng et al.

GIGA Architecture



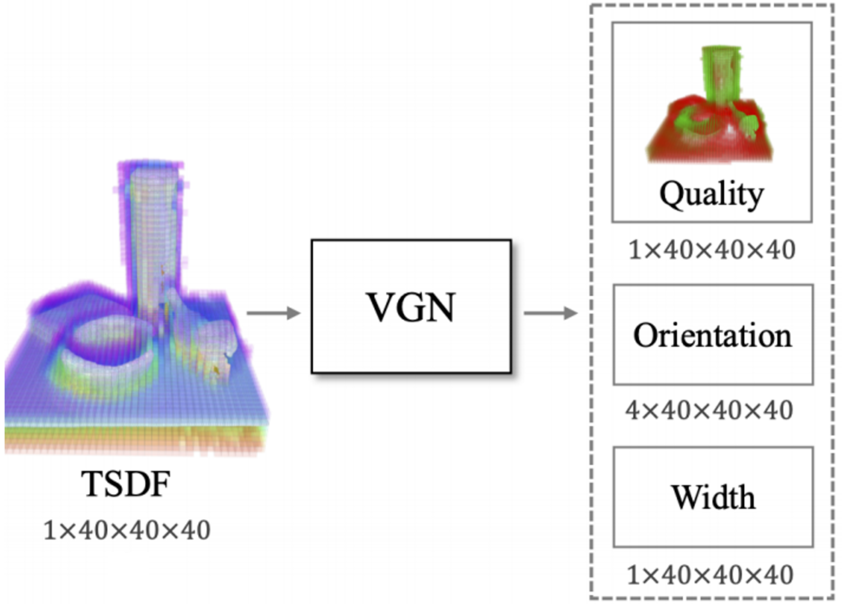
*Backbone architecture largely based on prior work: Convolutional Occupancy Networks (2020) by Peng et al.

Training Details

- Trained in simulation
- Grasps randomly sampled around surface of objects
- Grasp width and rotation only trained on successful grasps
- Data must be balanced by eliminating extra unsuccessful grasps
- 3D reconstruction trained using uniformly sampled points
- Noise added to images to aid with sim2real transfer

Closest Existing Approach: Volumetric Grasping Network (VGN)

- Predicts grasp affordance for each voxel rather than using an implicit neural representation
- Does not learn 3D reconstruction



[Breyer et al. 2021]

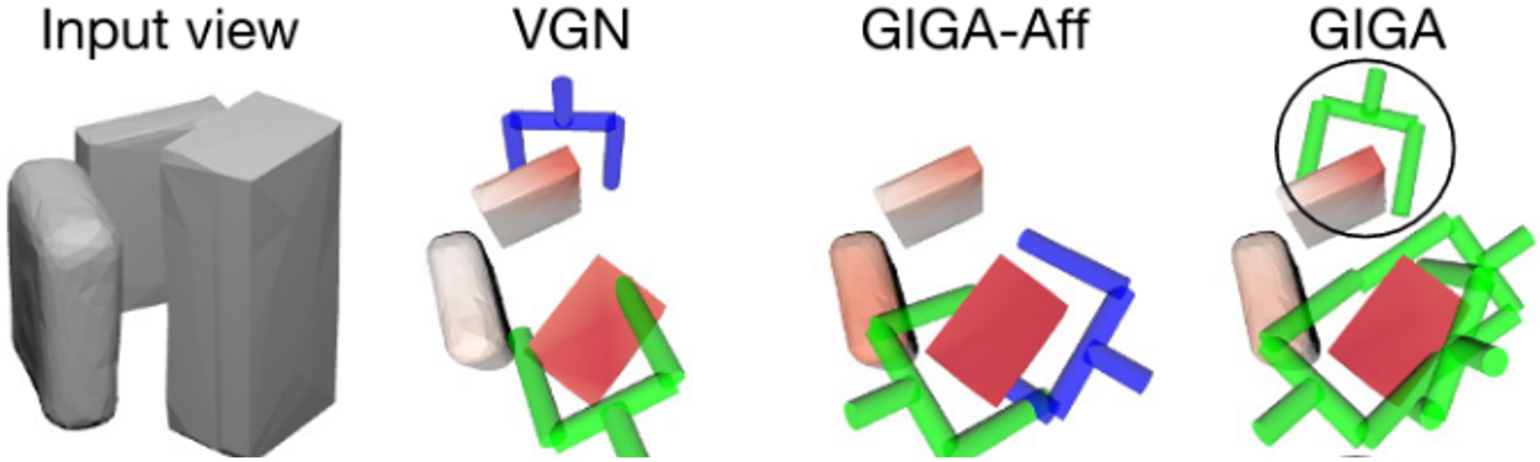
GIGA Grasping Results

Method	Packed		Pile	
	GSR (%)	DR (%)	GSR (%)	DR (%)
SHAF [13]	56.6 ± 2.0	58.0 ± 3.0	50.7 ± 1.7	42.6 ± 2.8
GPD [16]	35.4 ± 1.9	30.7 ± 2.0	17.7 ± 2.3	9.2 ± 1.3
VGN [4]	74.5 ± 1.3	79.2 ± 2.3	60.7 ± 4.2	44.0 ± 4.9
GIGA-Aff	77.2 ± 2.3	78.9 ± 1.7	67.8 ± 3.0	49.7 ± 1.9
GIGA	83.5 ± 2.4	84.3 ± 2.2	69.3 ± 3.3	49.8 ± 3.9
GIGA (HR)	$\mathbf{87.9 \pm 3.0}$	$\mathbf{86.0 \pm 3.2}$	$\mathbf{69.8 \pm 3.2}$	$\mathbf{51.1 \pm 2.8}$

GIGA-Aff: without the reconstruction component
GIGA (HR): high resolution (60x60x60 sampled grasp candidates rather than 40x40x40)

GSR: grasp success rate
DR: declutter rate (proportion of items removed after running until failure)

GIGA Grasping Results



GIGA 3D Reconstruction Results

Method	IoU (%)	IoU-Grasp (%)	$\Delta\%$ (IoU-Grasp$-$IoU)
GIGA-Detach	53.2	68.8	**+15.6**
GIGA	70.0	78.1	+8.1
GIGA-Geo	**80.0**	**84.0**	+4.0

GIGA-Detach: features trained for grasping, weights frozen, final layers trained for reconstruction

GIGA-Geo: end-to-end trained for reconstruction only

IoU: intersection-over-union of reconstructed object

IoU-Grasp: the IoU around graspable regions only

GIGA 3D Reconstruction Results

Real Robot Experiments

Method	Packed	
	GSR (%)	DR (%)
VGN [4]	77.2 (61 / 79)	81.3
GIGA	**83.3** (65 / 78)	**86.6**

Method	Pile	
	GSR (%)	DR (%)
VGN [4]	79.0 (64 / 81)	85.3
GIGA	**86.9** (73 / 84)	**97.3**

Limitations/Future Work

- Currently throws away reconstructed 3D information at test time
  - Could be used for closed-loop control
- Assumes a single static viewpoint
  - Not generalizable to a mobile robot or camera system
- sim2real transfer not very thoroughly evaluated
  - A real out-of-laboratory setting would be much more noisy and likely hurt performance significantly
- Could possibly be extended to other manipulation tasks?

Summary

- Grasping arbitrary objects is a very hard but fundamentally useful task
- GIGA does not rely on known object models, multiple views, or uncluttered/unoccluded scenes
- Key insights:
    - Implicit neural representations work well for efficiently representing grasp affordance
    - Learning 3D reconstruction synergizes with grasping, especially for occluded objects
- GIGA demonstrates state-of-the-art results on cluttered grasping from a single view

Thank you!

Extended Reading

- A Survey on Learning-Based Robotic Grasping
- Volumetric Grasping Network
- Occupancy Networks
- Convolutional Occupancy Networks