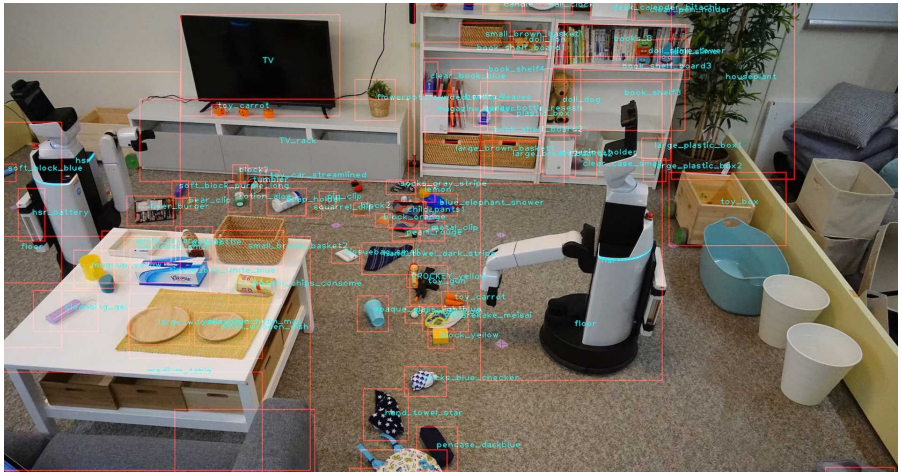
Synergies Between Affordance and Geometry: 6-DoF Grasp Detection via Implicit Representations

Presenter: Marlan McInnes-Taylor

Sept. 7, 2023

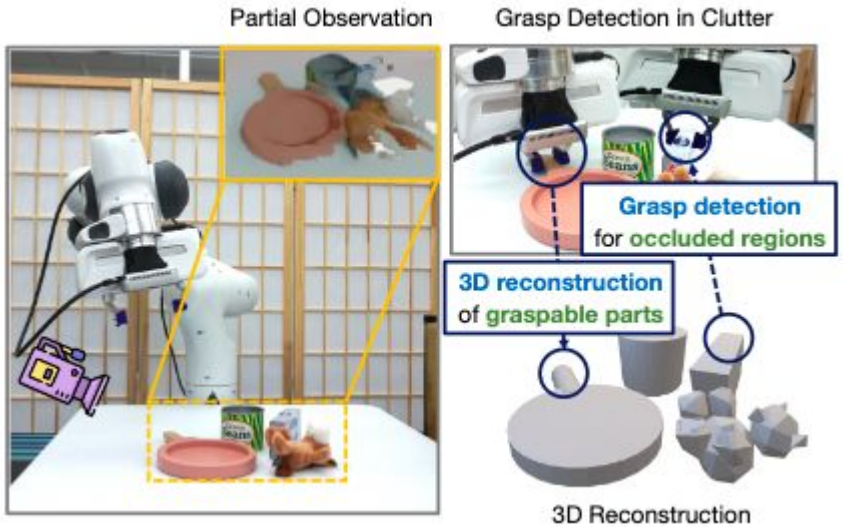
Motivation - Robot Grasping

- Robots must physically interact with (often) unstructured environments
- Grasp detection determines arm & gripper behavior
- Progressing robot autonomy requires robust grasping
- Modern deep nets used to grasp prediction



Proposed Work

"...a learned representation capable of reconstructing the 3D scene encodes relevant geometry information for predicting grasp points and vice versa."



Problem Formulation

Problem: *6-DoF grasp detection for unknown rigid objects in clutter from a single-view depth image*

- Robot arm equipped with a parallel-jaw gripper in a cubic workspace with a planar tabletop
- Single-view depth image taken with a fixed side view depth camera is fused into a Truncated Signed Distance Function (TSDF) and then passed into the model.
- The model outputs 6-DoF grasp pose predictions and associated grasp quality.

Prior Work - Grasp Detection

- Analytical & 3D reconstruction methods
- **Dex-Net** [2017] 4-DoF
  - two-stage pipeline for top-down antipodal grasping
- **GraspNet** [2019] 6-DoF
  - variational autoencoder for grasp proposals on a singulated object
- **GPD** [2016] & **PointGPD** [2019] 6-DoF
  - two-stage pipeline for grasp detection in clutter
- **VGN** [2020] 6-DoF
  - one-stage pipeline from input depth images for grasp detection in clutter

Prior Work - Geometry-Aware Grasping

- **DGGN** [2018]
  - predicts a voxel occupancy grid from partial observations and evaluates grasp quality from feature of the reconstructed grid
- **PointSDF** [2020]
  - learns 3D reconstruction via implicit functions and shares the learned geometry features with the grasp evaluation network

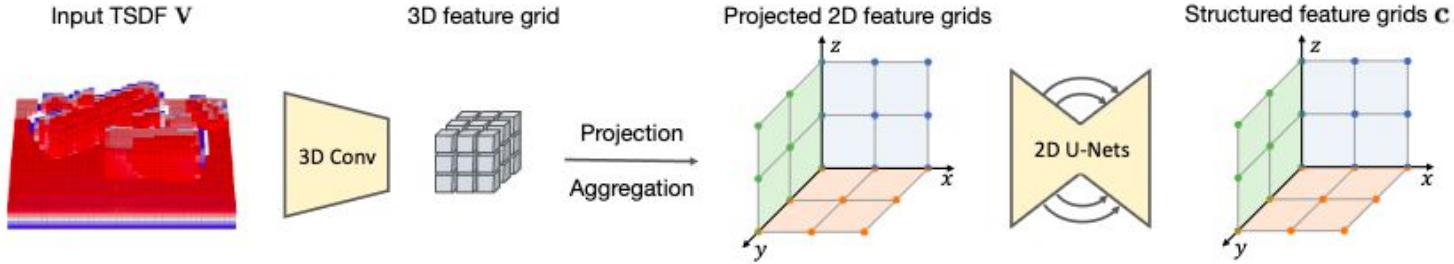
Problem Setting - Notation

- **Observations**
  - given a depth image captured by a depth camera and fused it into a TSDF, we have an N^3 voxel grid $\mathbf{V}$ where each cell V_i contains the truncated signed distance to the nearest surface
- **Grasps**
  - a 6-DoF grasp g as the grasp center position $\mathbf{t} \in R^3$, the orientation $\mathbf{r} \in SO(3)$ of the gripper, and the opening width $w \in R$ between the fingers
- **Grasp Quality**
  - a scalar grasp quality $q \in [0, 1]$ estimates the probability of grasp success
- **Occupancy**
  - for an arbitrary point $\mathbf{p} \in R^3$, the occupancy $b \in \{0, 1\}$ is a binary value indicating whether this point is occupied by any of the objects in the scene

Problem Setting - Objectives

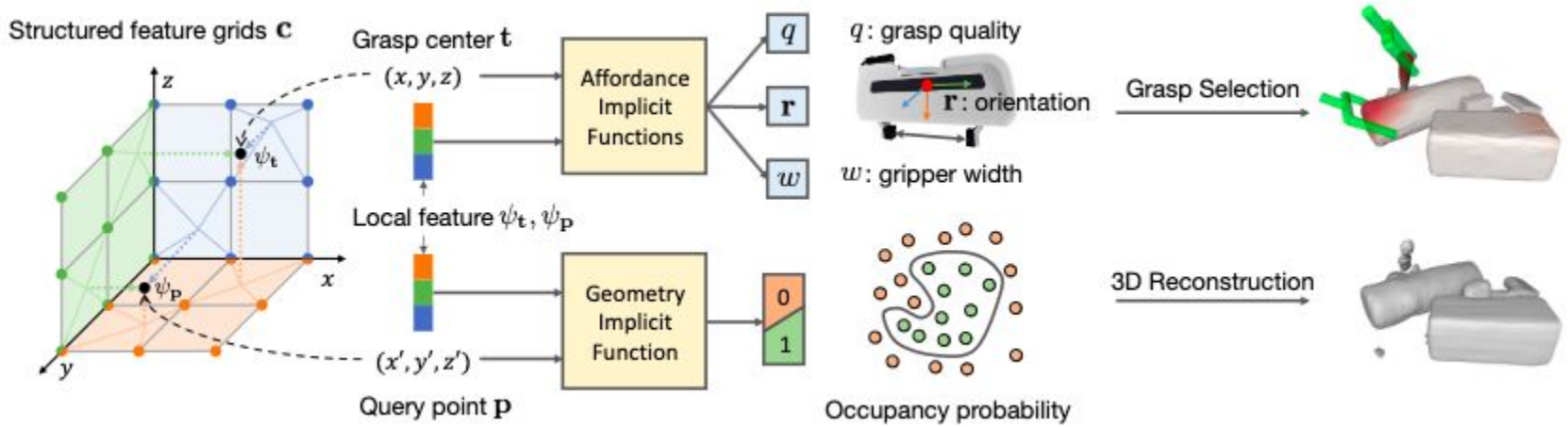
- **Primary goal**: detect 6-DoF grasp configurations that allow the robot arm to successfully grasp and remove the objects from the workspace
- Learn two functions:
  - $f_a : \mathbf{t} \rightarrow q, \mathbf{r}, w$
    - maps from a grasp center to the rotation, gripper width, and grasp quality of the best grasp at that location
  - $f_g : \mathbf{p} \rightarrow b$
    - maps any point in the workspace to the estimated occupancy value at that point

GIGA - Grasp detection via Implicit Geometry and Affordance



1. An encoder adopted from ConvONets accepts a TSDF voxel field and creates a feature embedding for each voxel
2. 3D feature grids are projected and aggregated 2D feature grids
3. Feature grids are processed by 2D U-Nets, creating a structured feature planes

GIGA



1. Given a grasp center and occupancy query point **p**, query the local features on the structured planes using bilinear interpolation
2. Affordance implicit functions predict grasp parameters from the local feature at the grasp center
3. Geometry implicit function predicts occupancy probability from the local feature at the query point

GIGA - Training

- Loss is comprised of affordance loss and geometry loss
- Affordance loss (L_A)
  - $L_A(\hat{g}, g) = L_q(\hat{q}, q) + q(L_r(\hat{r}, r) + L_w(\hat{w}, w))$
- Geometry loss (L_G)
  - Binary cross-entropy loss between the predicted occupancy and the ground-truth occupancy label
- Final loss function is the direct sum of affordance and geometry loss
  - $L = L_A + L_G$

Experimental Setup

- Evaluation tasks - clutter removal in packed and piled scenarios
- Simulated environment utilizes PyBullet
  - free gripper samples grasps in 30x30x30 cm^3 tabletop workspace
- Model trained in a self-supervised manner with ground-truth grasp labels collected from physical trials in simulation and occupancy data obtained from the object meshes
  - dataset balanced by discarding redundant negative samples

Experimental Setup

1. Can the structured implicit neural representations encode action-related information for grasping?
2. Does joint learning of geometry and affordance improve grasp detection?
3. How does grasp affordance learning impact the performance of 3D reconstruction?

Experimental Setup (Grasp Detection) - Baselines

SHAF

- Uses the *highest point heuristic* where from all grasps of quality >= 0.5 predicted by GIGA, the highest one is selected

GPD (Grasp Pose Detection)

- two-stage 6-DoF grasp detection algorithm which generates a large set of grasp candidates and classifies each of them

VGN (Volumetric Grasping Network)

- single-stage 6-DoF grasp detection algorithm which generates a large number of grasp parameters in parallel given input TSDF volume

GIGA-Aff (abalation)

- GIGA model with only affordance implicit function branch

Experimental Setup - Metrics

Grasp Detection (simulated & real)

- Grasp success rate (GSR) - the ratio of success grasp execution
- Declutter rate (DR) - the average ratio of objects removed

3D Reconstruction

- Volumetric IoU - interaction over the union between predicted mesh and ground-truth
- IoU-Grasp - IoU near graspable parts given by grasp trials

Results - Grasp Detection

TABLE I: Quantitative results of clutter removal. We report mean and standard deviation of grasp success rates (GSR) and declutter rates (DR). HR denotes high resolution.

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)	$\mathbf{87.9 \pm 3.0}$	$\mathbf{86.0 \pm 3.2}$	$\mathbf{69.8 \pm 3.2}$	$\mathbf{51.1 \pm 2.8}$

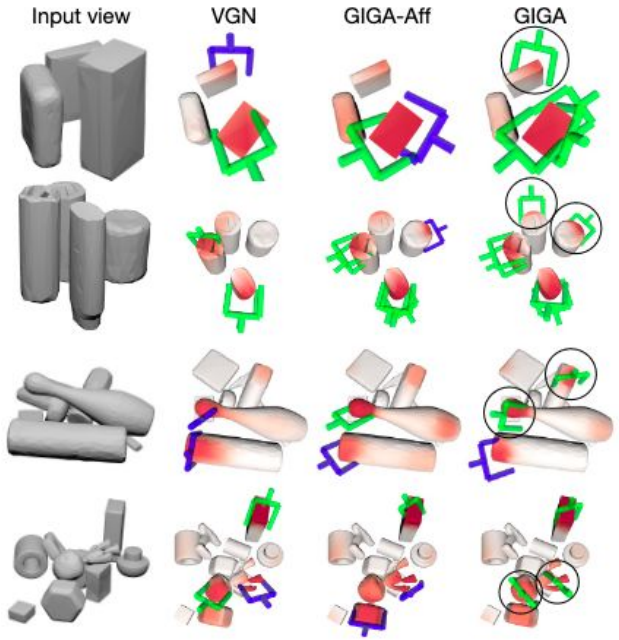


Fig. 4: Visualization of the grasp affordance landscape and predicted grasps. **Red** indicates that the method predicts high grasp affordance near the corresponding area. **Green** indicates successful grasps and **Blue** failures. The circles highlight interesting examples, such as asymmetric affordance heatmaps and highly occluded objects.

Results - 3D Reconstruction

TABLE II: Quantitative results of 3D reconstruction. We see that models trained with grasp supervision tend to have better reconstruction performance near graspable regions than average by a larger margin.

Method	IoU (%)	IoU-Grasp (%)	Δ% (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

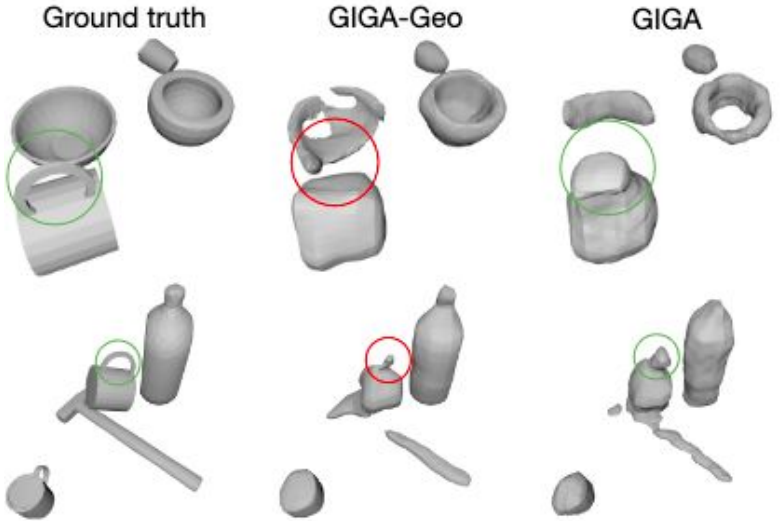


Fig. 5: Qualitative 3D reconstruction results of a scene rendered from the top view. The circles highlight the contrast.

Results - Real Robot

TABLE III: Quantitative results of clutter removal in real world. We report GSR, DR, the number of successful grasps, and the number of total grasp trials (in bracket).

Method	Packed		Pile	
	GSR (%)	DR (%)	GSR (%)	DR (%)
VGN [4]	77.2 (61/79)	81.3	79.0 (64/81)	85.3
GIGA	**83.3** (65/78)	**86.6**	**86.9** (73/84)	**97.3**

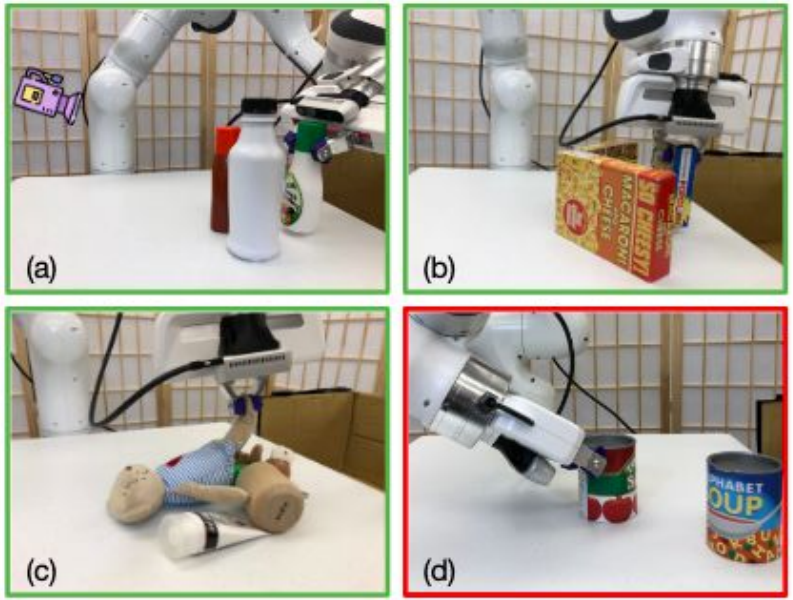


Fig. 6: Examples of real-world grasps by GIGA. (a) and (b) show two examples of grasps of partially occluded objects. The tilted camera looks at the workspace from the left. (c) shows an example where GIGA picks up the bear doll from a localized graspable part. (d) illustrates a typical failure where the gripper slips off the object due to the small contact surface.

Discussion

- Grasp Detection
  - GIGA takes into account the context of the scene and predicts collision-free grasps
  - GIGA produces more diverse and accurate grasp detections compared with the baselines
  - GIGA improves grasp efficiency compared to VGN
  - GIGA performs similarly in single-view and multi-view contexts
    - supervision from 3D reconstruction facilitates reasoning about the occluded parts of the scene
- 3D Reconstruction
  - GIGA's reconstructions are more robust due to better understanding of local geometry

Future Work for Paper / Reading

- Extend GIGA to learn to predict the full distribution of viable grasp parameters with generative modeling
- Utilize the reconstructed 3D scene to constrain the grasp prediction to be collision-free
- Adapt GIGA to close-loop grasp planning by integrating real-time feedback

Extended Readings

- **T. Weng, D. Held, F. Meier and M. Mukadam**, "*Neural Grasp Distance Fields for Robot Manipulation*," 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 1814-1821, doi: 10.1109/ICRA48891.2023.10160217.
- **Haoran Geng, Helin Xu, Chengyang Zhao, Chao Xu, Li Yi, Siyuan Huang, He Wang**; "*GAPartNet: Cross-Category Domain-Generalizable Object Perception and Manipulation via Generalizable and Actionable Parts*", Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 7081-7091
- **Z. Jiang, C. -C. Hsu and Y. Zhu**, "*Ditto: Building Digital Twins of Articulated Objects from Interaction*," 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 2022, pp. 5606-5616, doi: 10.1109/CVPR52688.2022.00553.

Summary

- **Problem**: 6-DoF grasp detection for unknown rigid objects in clutter from a single-view depth image
- Challenging due to unstructured environments, partial observations, and not enough local geometric information
- **Prior work** primarily focused only on learning grasp prediction
- **Insight**: grasp-prediction and 3D reconstruction share underlying similarities
- **Results**: jointly learning grasp-prediction and 3D reconstruction improves both tasks