UC San Diego

S^{4}G: Amodal Single-view Single-Shot $\mathbb{SE}$(3) Grasp Detection in Cluttered Scenes

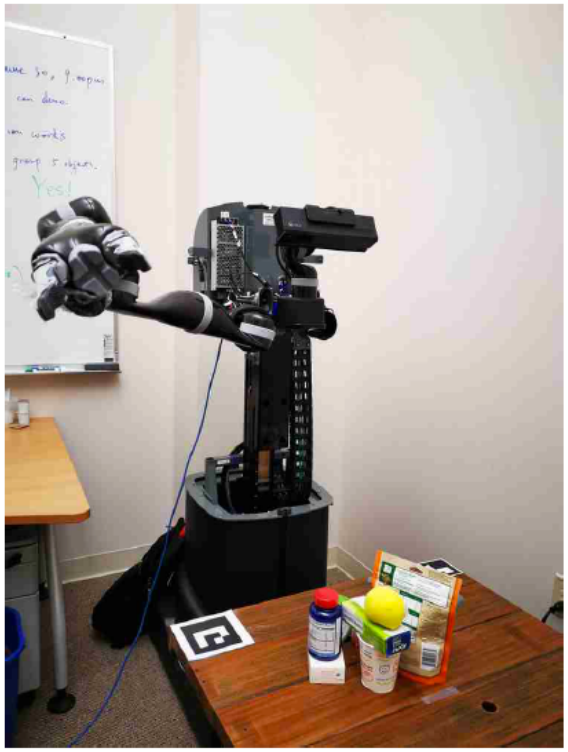
Yuzhe Qin, Rui Chen, Hao Zhu, Meng Song, Jing Xu, Hao Su

Presenter: Yiran Xu
May 7th 2020

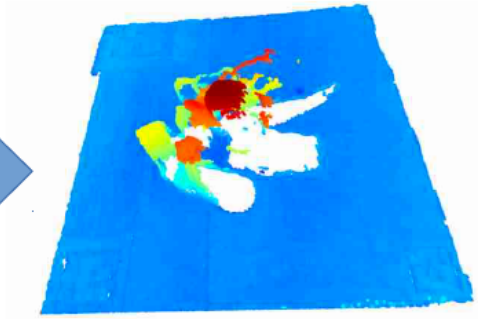
Outline

- Big Idea
- Related work
- Method
  - Flat Surface Contact Gripper model
  - Training Data Generation
  - Single-shot Grasps Generation
- Take-home message

Problem Setting



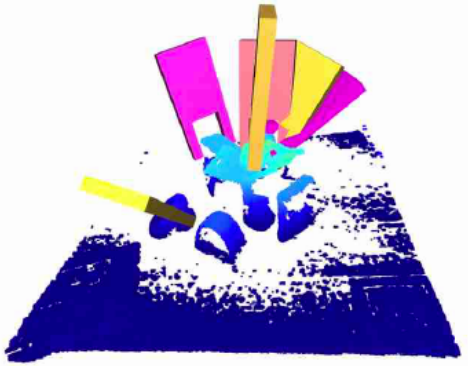
Robot initial state



Single-view point cloud

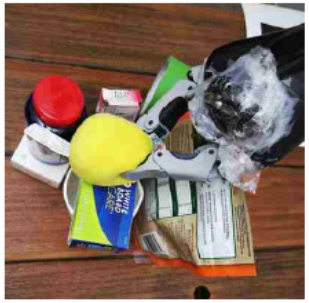
Single-shot Grasp Proposal Network

NMS



High quality grasp set

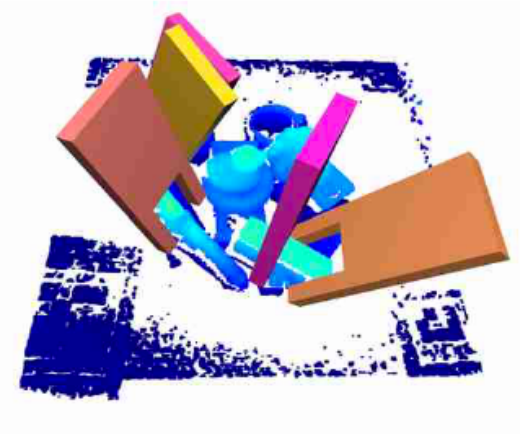
Weighted Sampling



Grasp execution

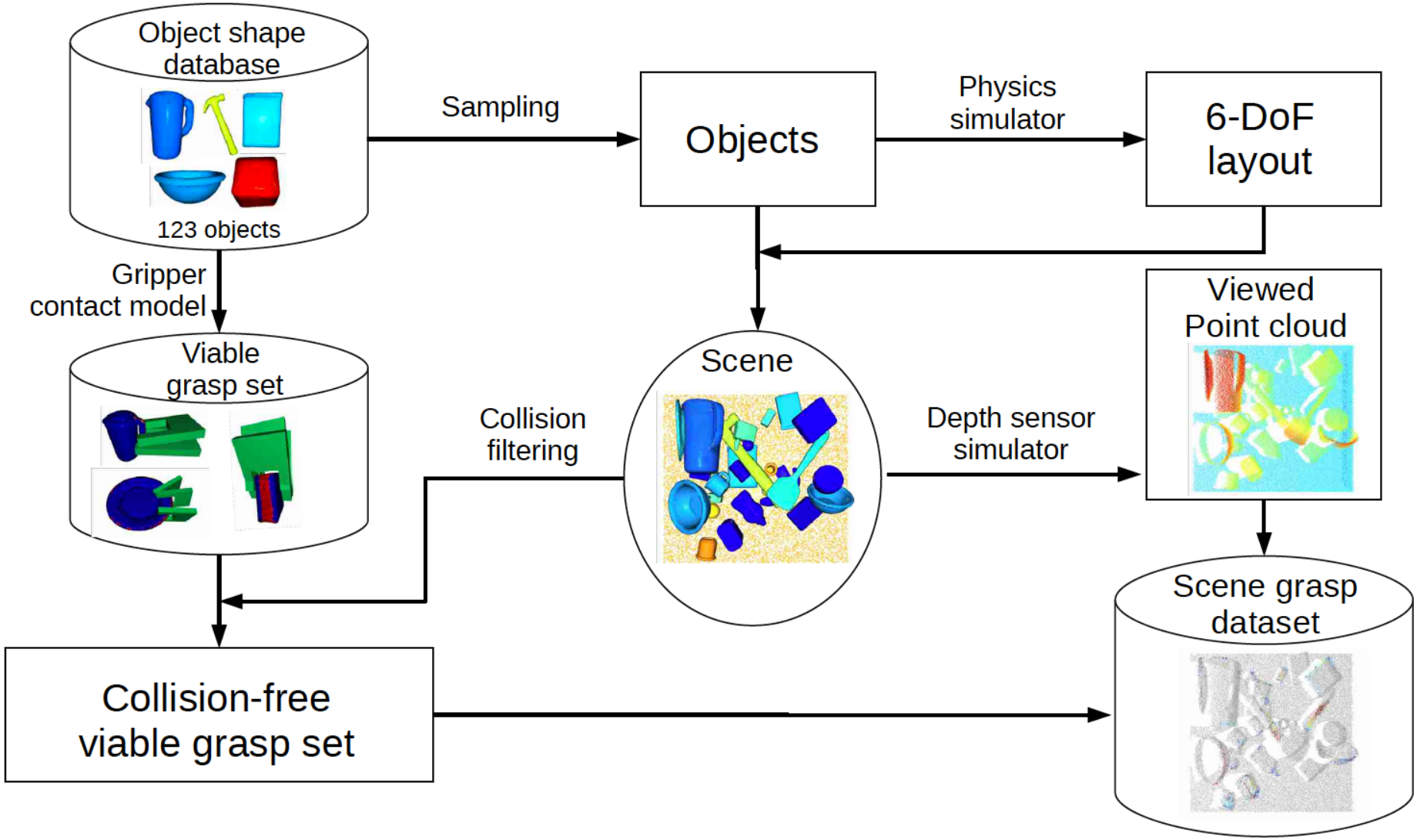
- Model-based ⟶ Data-Driven
- 3 / 4 Degree of Freedom(DoF) ⟶ 6 DoF

Motivation

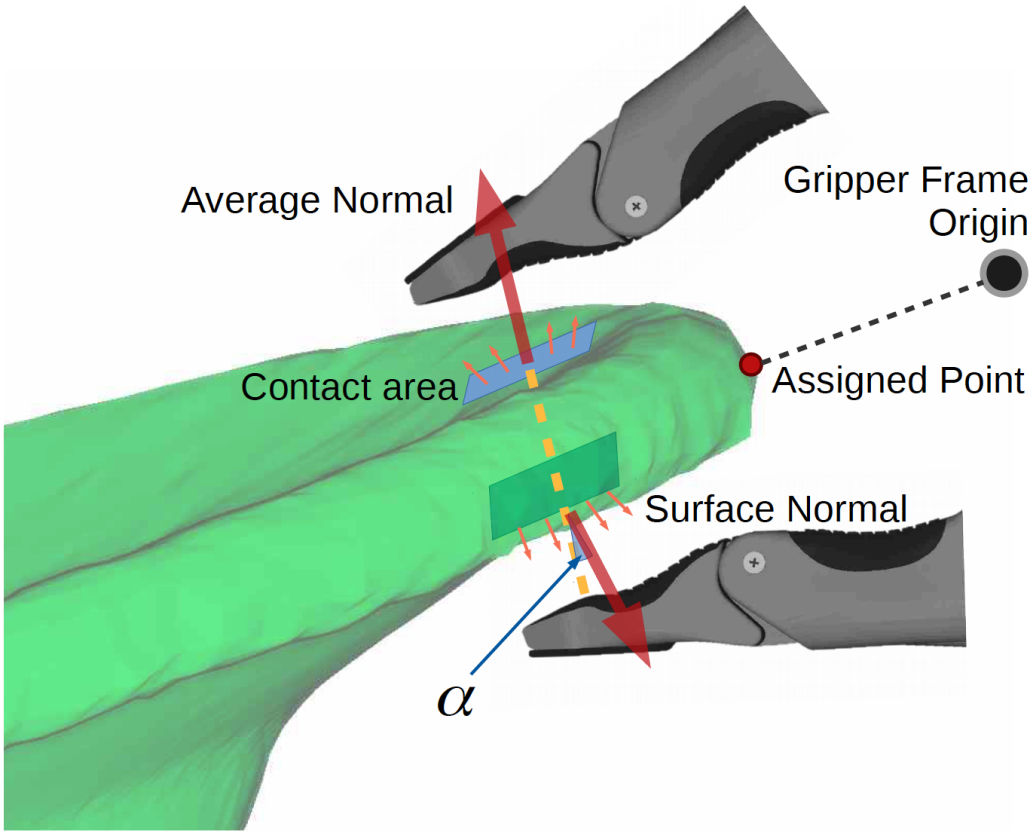


- Sampling ⟶ Regression
- Single object ⟶ Many Objects

Generating Training Data

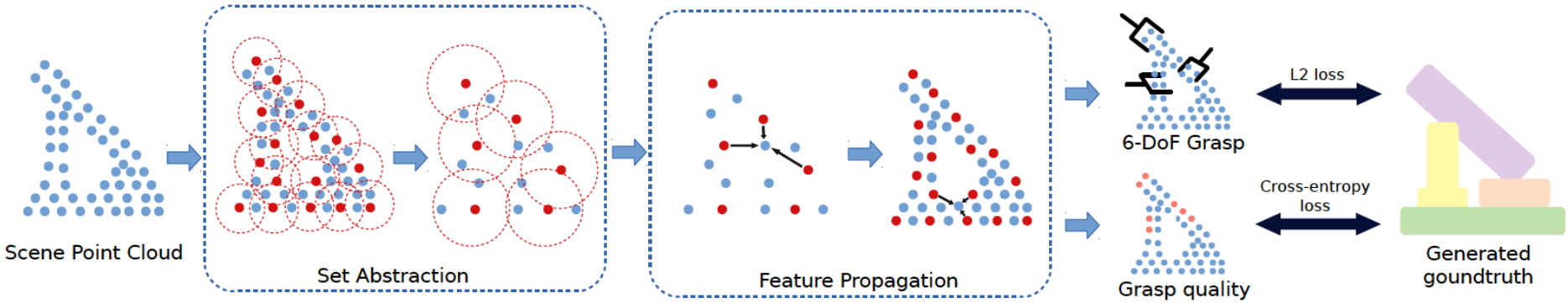


Store the Grasping Pose on the Surface



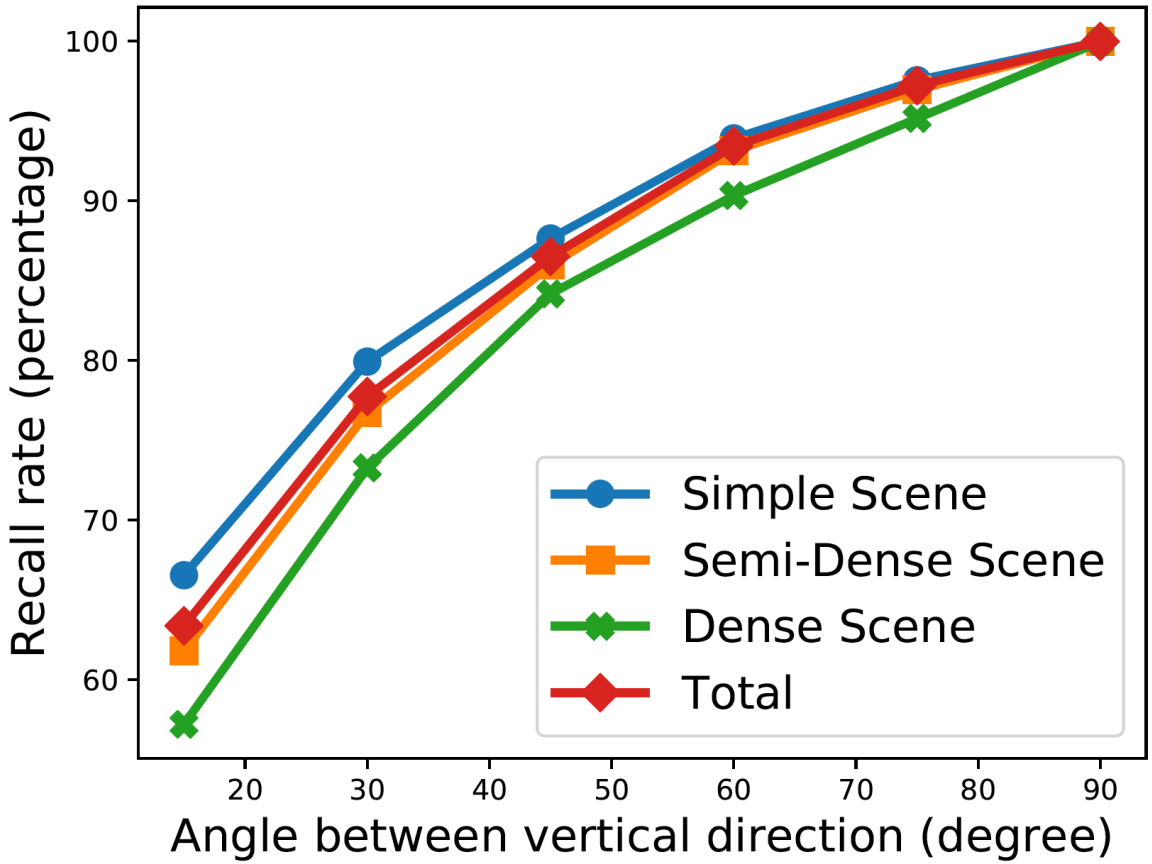
- Grasp proposal ⟶ Per-points labeling

Grasp Proposal as Per-point Labeling



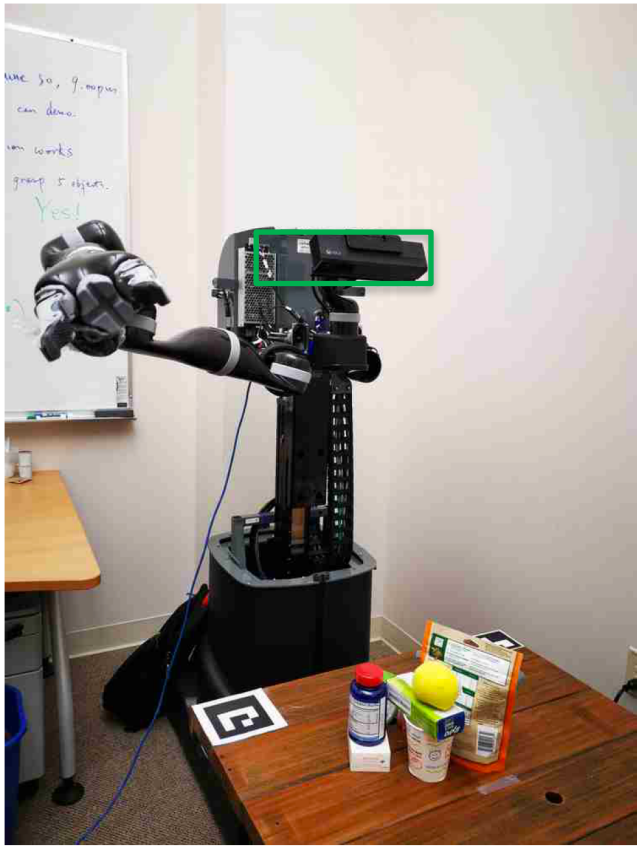
- Single-view
- Single-shot
- $\mathbb{SE}(3)$

Superiority of $\mathbb{SE}(3)$ Grasp



- Only 63.38% objects can be grasped by nearly vertical grasps (0, 15)

Experiments



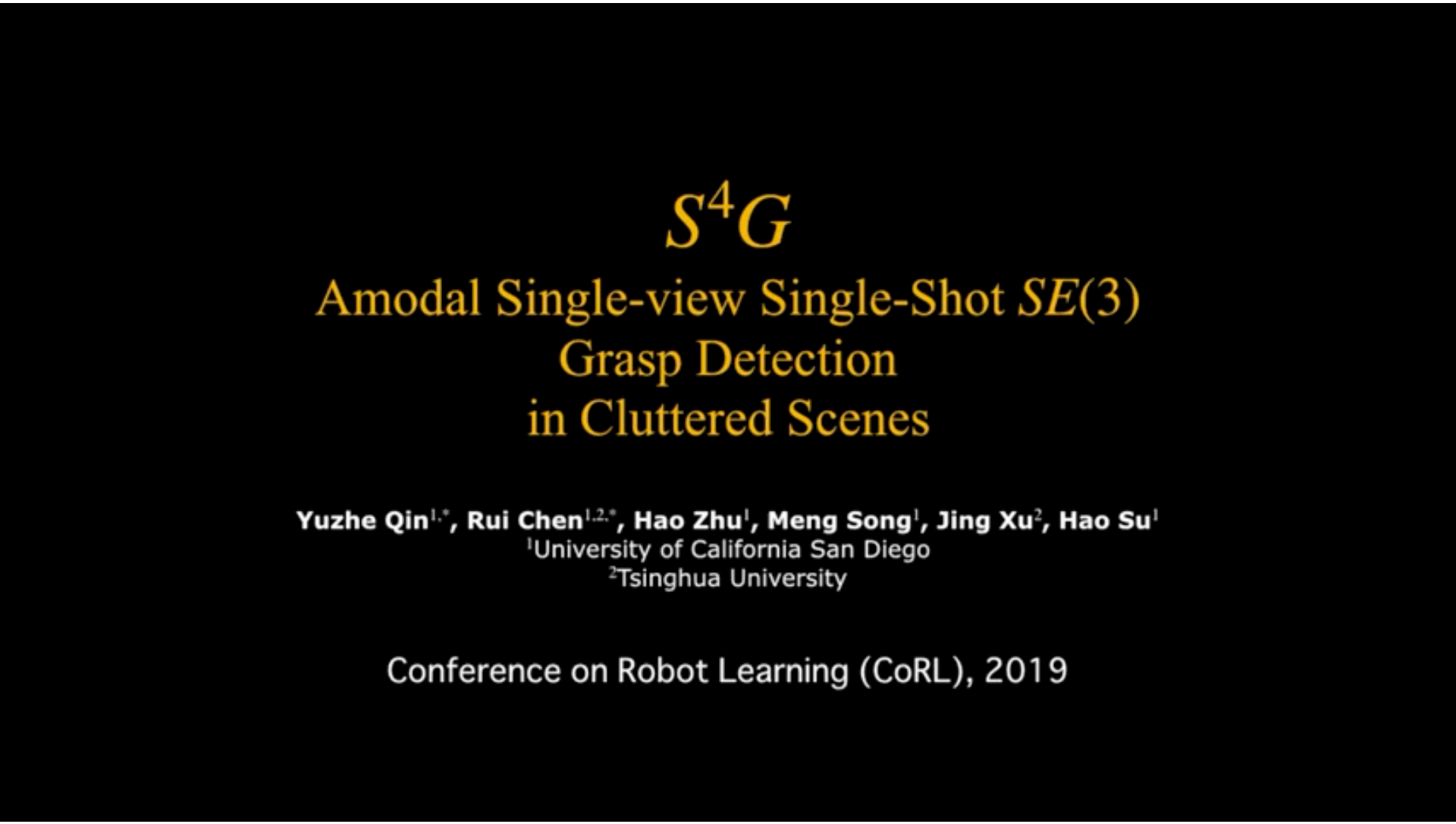
- Single-view depth Kinect
- Novel test set

Experimental Results

	Grasp quality		Time-efficiency		
	Success rate	Completion rate	Processing	Inference	**Total**
GPD (3 channels)	40.0%	60.0%	24106 ms	**1.50 ms**	24108 ms
GPD (12 channels)	33.3%	50.0%	27195ms	1.70ms	27197ms
PointNetGPD	40.0%	60.0%	17694ms	2.86ms	17697ms
Ours	**77.1%**	**92.5%**	**5804ms**	12.60 ms	**5817 ms**

- Outperform the baseline
- Much more time-efficient

Result Demonstration



Related Work

- Deep Learning based Grasping Methods
- Training Data Synthesis for Grasping
- Deep Learning on 3D Data

Deep Learning based Grasping Methods

- Deep learning is effective for robotic grasping
- To retrieve 6-DoF pose:

Fit object model to the scan point cloud

↓ to achieve better generalizability of novel objects

Generate grasp hypotheses based on local geometry prior

↓ further extended

Replacing multi-view projection features with direct point cloud representation

Training Data Synthesis for Grasping

- Analytic grasp synthesis
  - Complete and precise geometric models
- S^4G:
  - Use this to generate viable grasps
  - Reject unfeasible grasps

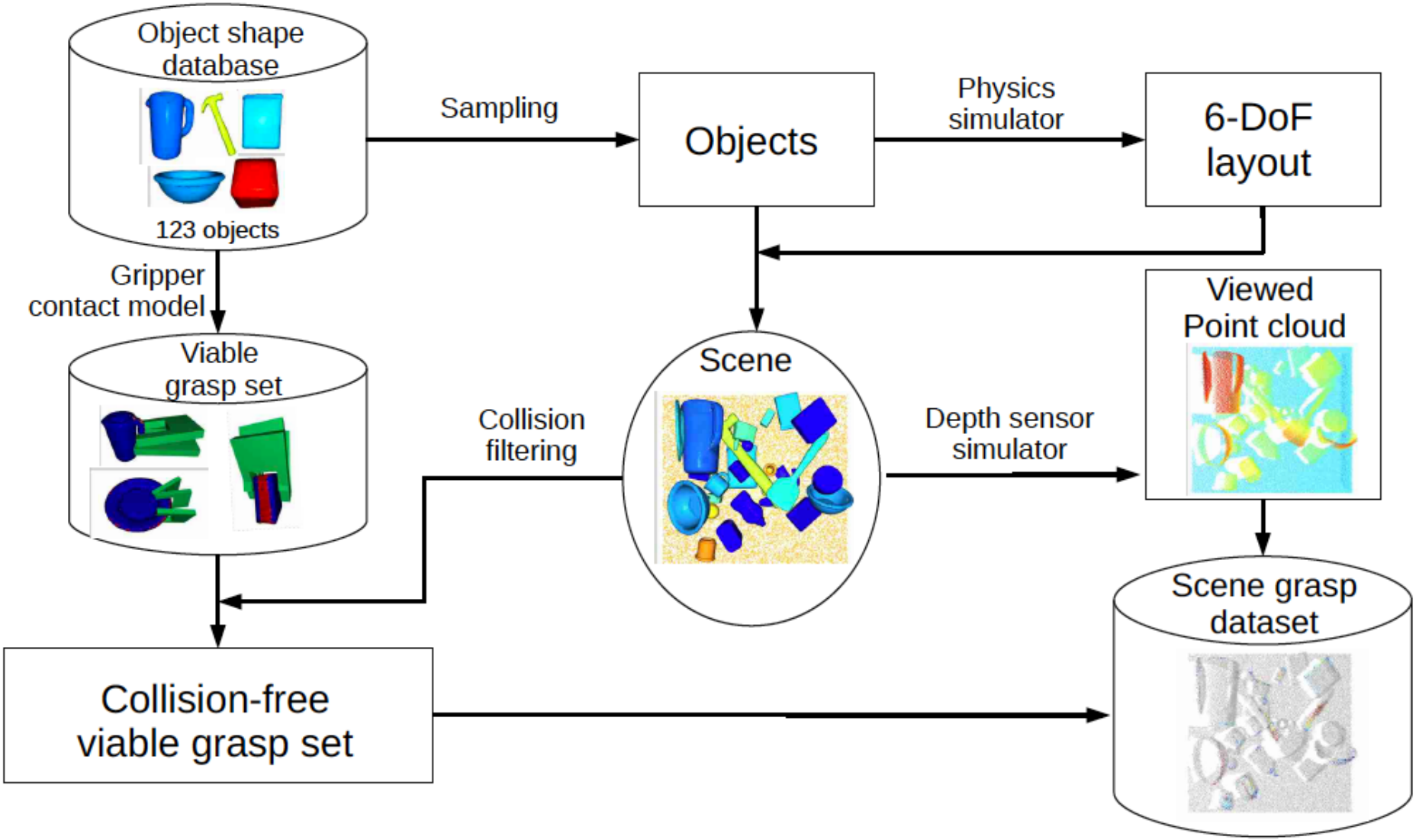
Deep Learning on 3D Data

- PointNet and PointNet++:
  - Extract useful representations from 3D point clouds
- S^4G:
  - Use PointNet++ as the backbone of single-shot grasp detection

Problem Setting

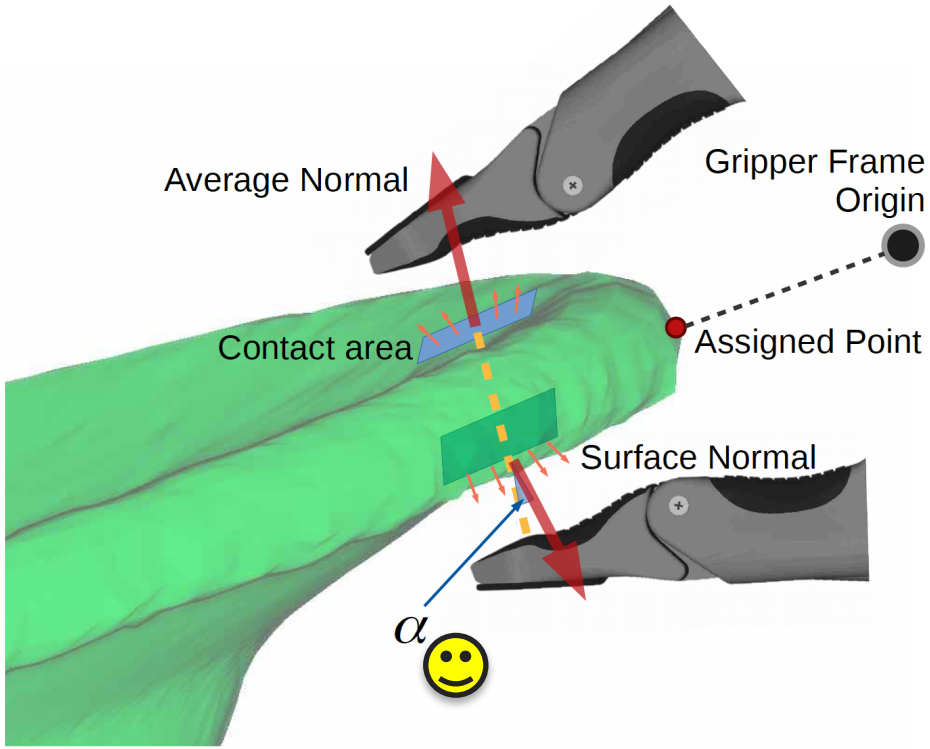
- Single-view point cloud: P
- Gripper description: G
- Grasp configuration:
  - $c = (\mathbf{h}, s_{\mathbf{h}})$
  - where $\mathbf{h} \in \mathbb{SE}(3)$ and $s_{\mathbf{h}} \in \mathbb{R}$ is a score measuring the quality of $\mathbf{h}$.

Training Data Generation



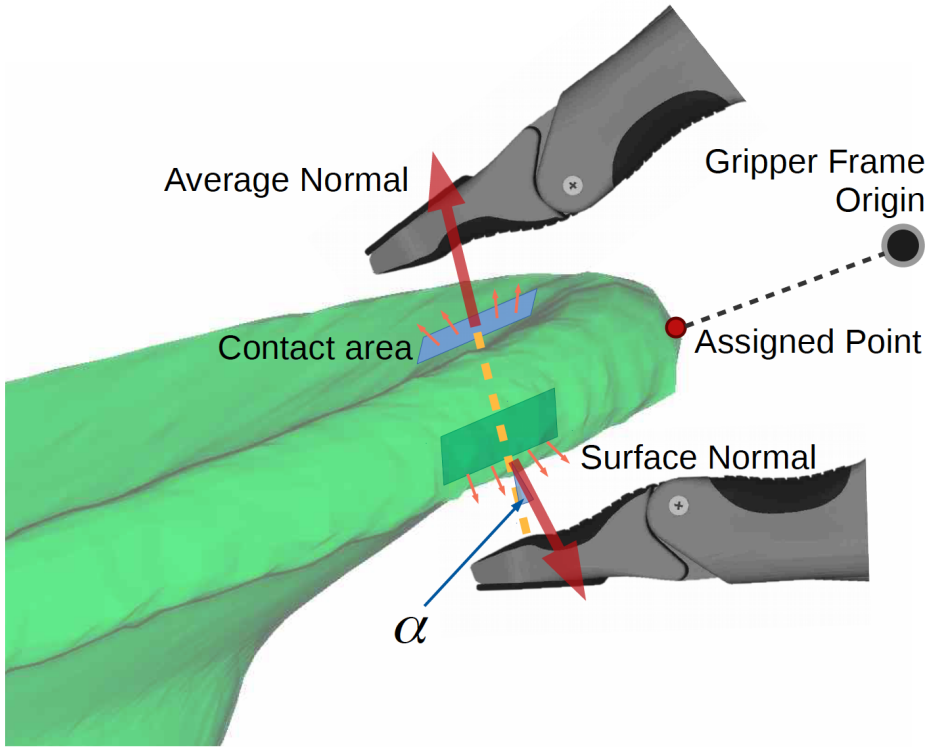
Gripper Contact Model

- Antipodal score $s_h^a = \cos(\alpha_1)\cos(\alpha_2)$
- α_i : the **angle** between the outward **normal** and the **line** connecting two **contact points**



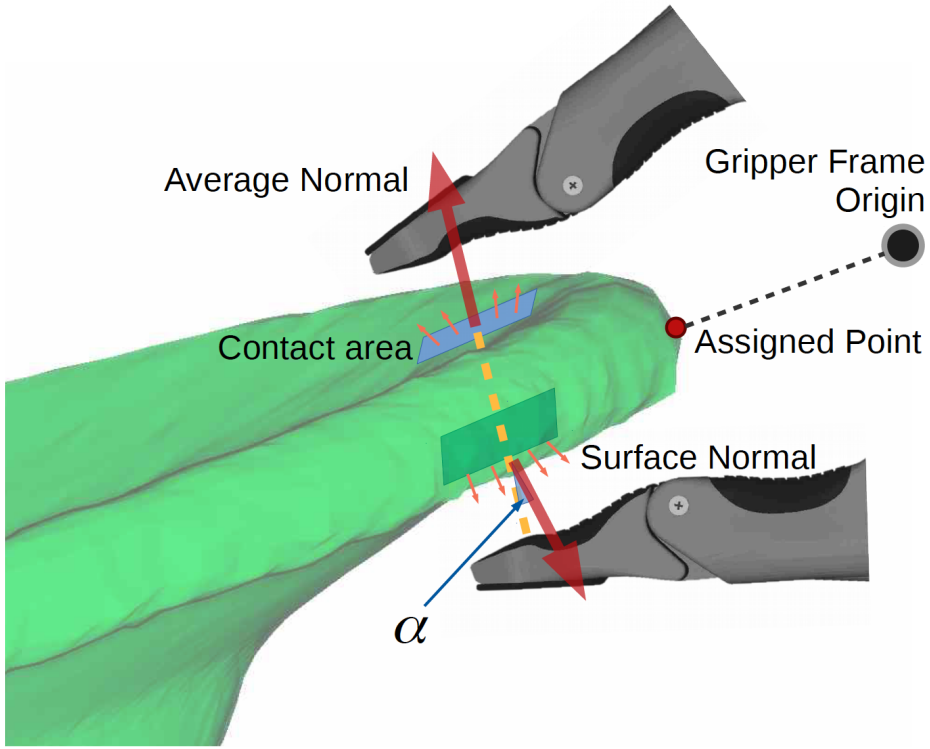
Details of Calculation

- For each **contact pair** (p_i, p_j), the **normal n_i at point p_i** is **smoothed** with radius $r$mm.



Details of Calculation

- Remove the neighbors which **has a distance along the normal direction** larger than a threshold
- p_i^k is the k-th neighbor of point i
- $r_i^k = \left|\left(p_i^k - p_i\right) \cdot \frac{n_i}{|n_i|}\right|$



Stability of the Grasp

- Occupancy score s_h^o to judge stability
- $s_h^o = \min\{\ln(|P_{close}|), 6\}, P_{close} = R(c) \cap P$
  - $R(c)$: gripper closing region
  - P_{close}: the number of points within the closing region
  - s_h^o: represents the volume of objects with in $R(c)$

Robustness Grasp Generation by Scene Analysis

- Collision score s_h^c is a scene-specific **boolean mask** indicating the **occurrence of collision**

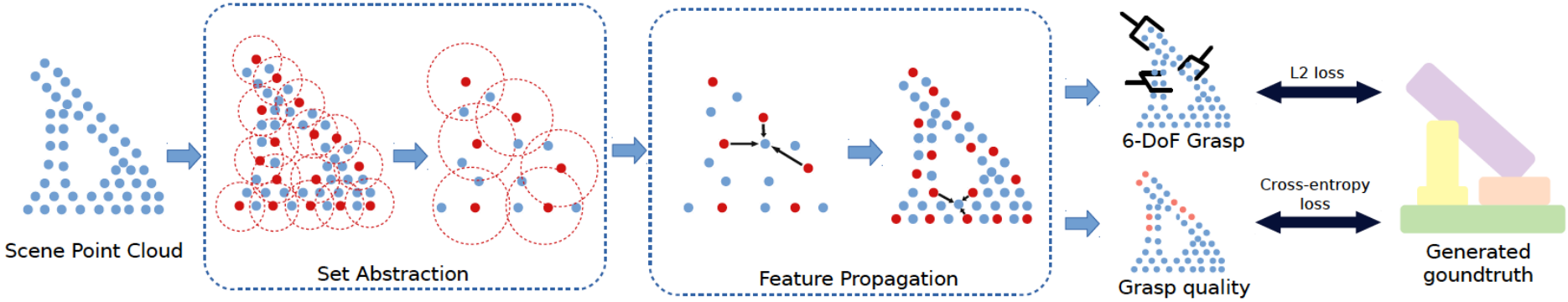
Robustness Grasp Generation by Scene Analysis

- The final score of each grasp
- $s_h = \min_j [s^a_{h_j}, s^o_{h_j}, s^c_{h_j}]$, $h_j = \exp(\hat{\xi})\, h$
  - $\hat{\xi} \in \mathfrak{se}(3)$: the pose perturbation
  - exp is the exponential mapping
- Training data for S^{4}G:
  - Viewed point cloud with ground truth grasps and scores

Single-Shot Grasp Generation

- PointNet++ based Grasp Proposal
- Non-maximum Suppression and Grasp Sampling

PointNet++ based Grasp Proposal



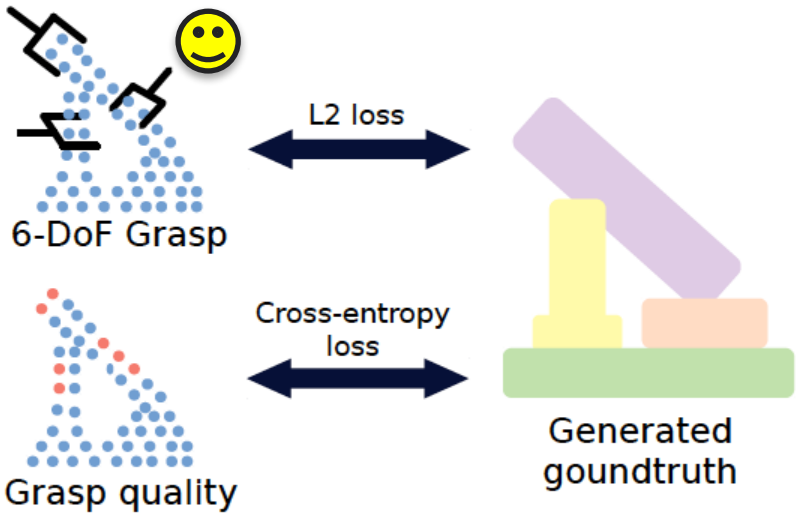
- Extracts hierarchical point set features
- Propagates the point set features to all the original points
- Predicts h_i and s_{h_i} of every point

Single-Shot 6-DoF Grasp Direct Regression Task

- $R \in \mathbb{SO}(3)$
- $a = [a_1, a_2], a_1, a_2 \in \mathbb{R}^3$
- The mapping $f: a \rightarrow R$ is:

$$\mathbf{R} = [\mathbf{b}_1, \mathbf{b}_2, \mathbf{b}_3]$$
$$\mathbf{b}_1 = N(\mathbf{a}_1)$$
$$\mathbf{b}_2 = N(\mathbf{a} - \langle \mathbf{a}_2, \mathbf{b}_1 \rangle \mathbf{b}_1)$$
$$\mathbf{b}_3 = \mathbf{b}_1 \times \mathbf{b}_2,$$

$N()$: normalization function

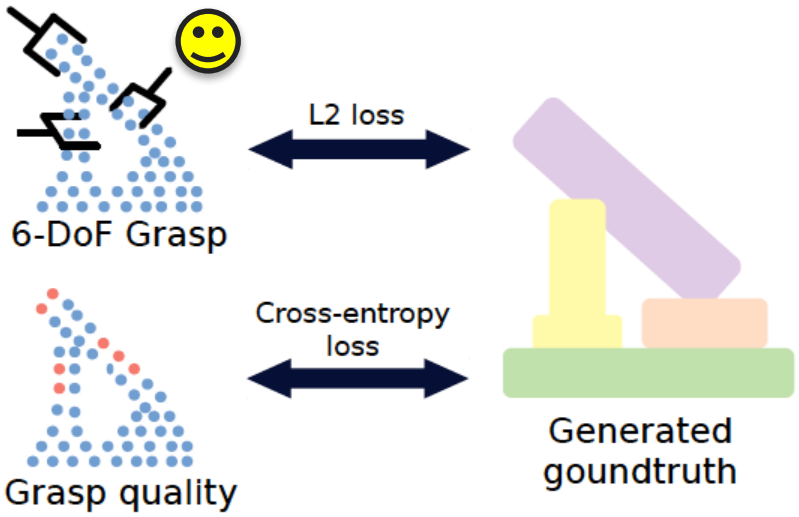


Single-Shot 6-DoF Grasp Direct Regression Task

- Given the groundtruth rotation matrix R_{GT}, the rotation loss L_{rot} is defined as:

$$L_{rot} = \min_{i \in \{0,1\}} \| f(\mathbf{a}_{pred}) - \mathbf{R}_{GT}^{(i)} \|^2$$

$$\mathbf{R}_{GT}^{(i)} = \mathbf{R}_{GT} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\pi i) & 0 \\ 0 & 0 & \cos(\pi i) \end{bmatrix}$$

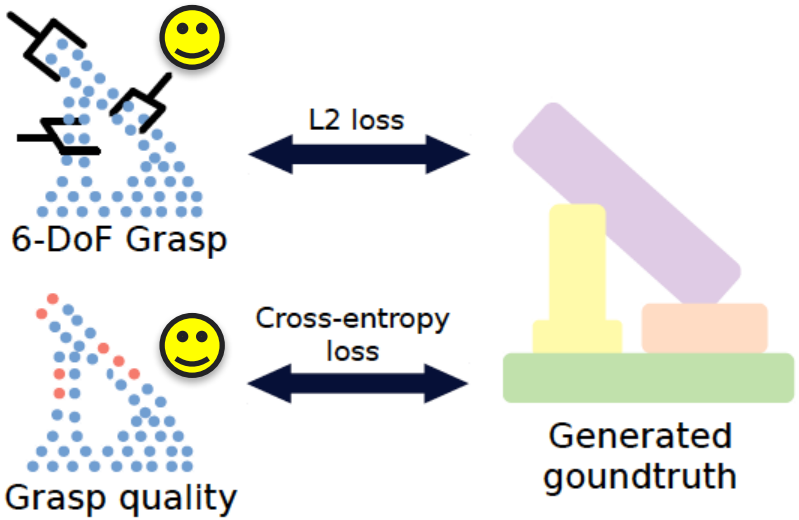


- This is treated as a regression task, thus we use L2 loss

Grasp Quality Multi-class Classification Task

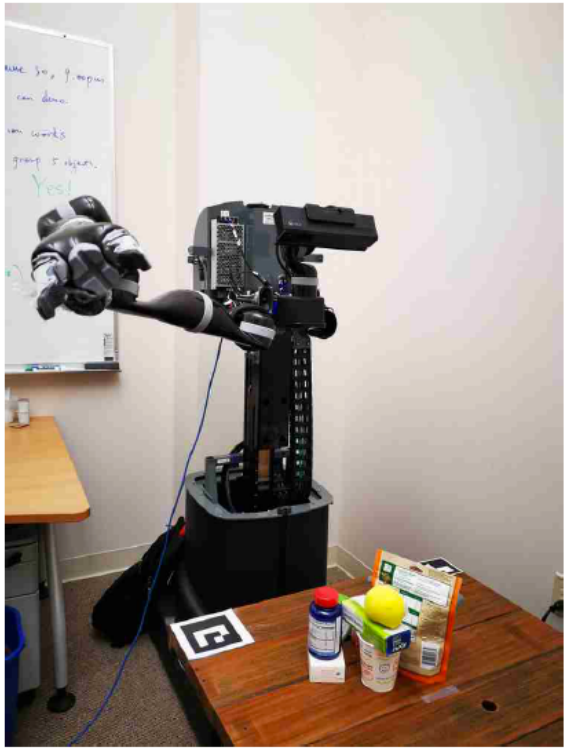
- The total is:

$$L = \sum_{\mathbf{P}_v} (\lambda_{rot} \cdot L_{rot} + \lambda_t \cdot L_t) + \sum_{\mathbf{P}_s} (\lambda_s \cdot L_s)$$

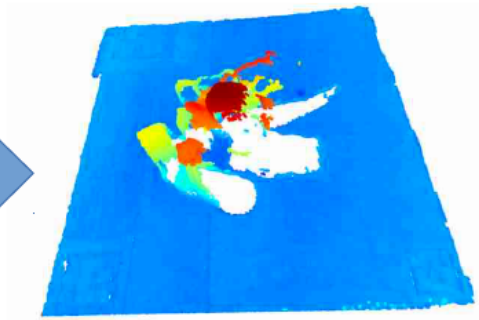


- P_v: point set with feasible grasps
- P_s: whole scene point cloud
- $\lambda_{rot}, \lambda_t, \lambda_s$ are set to 5.0, 20.0, 1.0 in experiments

Non-maximum Suppression and Grasp Sampling



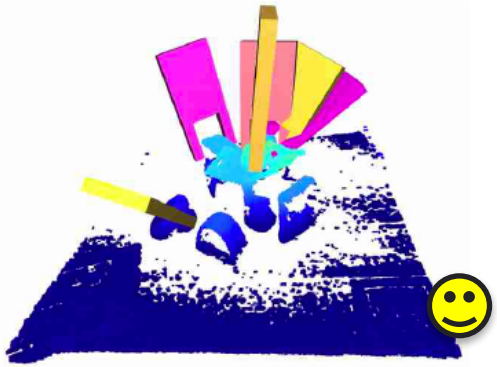
Robot initial state



Single-view point cloud

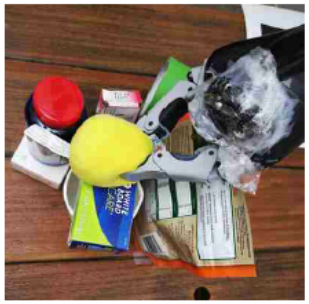
Single-shot Grasp Proposal Network

NMS



High quality grasp set

Weighted Sampling



Grasp execution

Non-maximum Suppression and Grasp Sampling

- Input is:
- Network prediction $\mathcal{C}$

Algorithm 1: NMS and Grasp sampling

Input: Prediction $\mathcal{C}$: $\{(\mathbf{h}_i, s_{\mathbf{h}_i})\}$

Export: Grasp Execution: h

Executable Grasps $\mathcal{H} = \{\}$

$Sort\ \{(\mathbf{h}_i, s_{\mathbf{h}_i}\}\ by\ s_{\mathbf{h}_i}$

$i = 0$

while $\mathcal{L}ength(\mathcal{H}) < N$ **do**

  if $(Collision == False)$ and $\mathbf{h}_k \in \mathcal{H} \underset{min}{dist}(\mathbf{h}_i, \mathbf{h}_k) > \epsilon$ **then**

    $Add\ (\mathbf{h}_i, s_{\mathbf{h}_i})$ to $\mathcal{H}$

  end if

  $i = i + 1$

end while

$$p_k = \frac{g(s_{\mathbf{h}_k})}{\sum_l g(s_{\mathbf{h}_l})} \text{ for } \mathbf{h}_k \in \mathcal{H}$$

while Motion planning fails **do**

  Sample h according to $\{p_k\}$

end while

Non-maximum Suppression and Grasp Sampling

- Input is:
- Network prediction $\mathcal{C}$
- Output is
- One grasp execution h

Algorithm 1: NMS and Grasp sampling

Input: Prediction $\mathcal{C}$: $\{(\mathbf{h}_i, s_{\mathbf{h}_i})\}$
Export: Grasp Execution: h 🙂
Executable Grasps $\mathcal{H} = \{\}$
$Sort\ \{(\mathbf{h}_i, s_{\mathbf{h}_i}\}\ by\ s_{\mathbf{h}_i}$
$i = 0$
while $\mathcal{L}ength(\mathcal{H}) < N$ **do**
  if $(Collision == False)$ and
  $\mathbf{h}_k \in \mathcal{H} \underset{min}{dist}(\mathbf{h}_i, \mathbf{h}_k) > \epsilon$ **then**
    $Add\ (\mathbf{h}_i, s_{\mathbf{h}_i})$ to $\mathcal{H}$
  end if
  $i = i + 1$
end while
$p_k = \dfrac{g(s_{\mathbf{h}_k})}{\sum_l g(s_{\mathbf{h}_l})}$ for $\mathbf{h}_k \in \mathcal{H}$
while Motion planning fails **do**
  Sample h according to $\{p_k\}$
end while

Non-maximum Suppression and Grasp Sampling

- …
- Use non-maximum suppression (NMS) to **select grasps** h with **local maximum** s_{h_i} to generate executable **grasp set** $\mathcal{H}$

Algorithm 1: NMS and Grasp sampling

Input: Prediction $\mathcal{C}$: $\{(\mathbf{h}_i, s_{\mathbf{h}_i})\}$
Export: Grasp Execution: h
Executable Grasps $\mathcal{H} = \{\}$
$Sort\ \{(\mathbf{h}_i, s_{\mathbf{h}_i}\}\ by\ s_{\mathbf{h}_i}$
$i = 0$
while $\mathcal{L}ength(\mathcal{H}) < N$ **do**
  if $(Collision == False)$ and $\mathbf{h}_k \in \mathcal{H} \underset{min}{dist}(\mathbf{h}_i, \mathbf{h}_k) > \epsilon$ **then**
    $Add\ (\mathbf{h}_i, s_{\mathbf{h}_i})$ to $\mathcal{H}$
  end if
  $i = i + 1$
end while
$p_k = \dfrac{g(s_{\mathbf{h}_k})}{\sum_l g(s_{\mathbf{h}_l})}$ for $\mathbf{h}_k \in \mathcal{H}$
while Motion planning fails **do**
  Sample h according to $\{p_k\}$
end while

Non-maximum Suppression and Grasp Sampling

- Weighted random sampling to generate executed grasp

Algorithm 1: NMS and Grasp sampling

Input: Prediction $\mathcal{C}$: $\{(\mathbf{h}_i, s_{\mathbf{h}_i})\}$
Export: Grasp Execution: h
Executable Grasps $\mathcal{H} = \{\}$
$Sort\ \{(\mathbf{h}_i, s_{\mathbf{h}_i}\}\ by\ s_{\mathbf{h}_i}$
$i = 0$
while $\mathcal{L}ength(\mathcal{H}) < N$ **do**
  if $(Collision == False)$ and
  $\mathbf{h}_k \in \mathcal{H} \underset{min}{dist}(\mathbf{h}_i, \mathbf{h}_k) > \epsilon$ **then**
    $Add\ (\mathbf{h}_i, s_{\mathbf{h}_i})$ to $\mathcal{H}$
  end if
  $i = i + 1$
end while
$p_k = \dfrac{g(s_{\mathbf{h}_k})}{\sum_l g(s_{\mathbf{h}_l})}$ for $\mathbf{h}_k \in \mathcal{H}$
while Motion planning fails **do**
  Sample h according to $\{p_k\}$
end while

Take-home Message

- The paper studied the problem of **6-DoF grasping** by a parallel gripper in a **cluttered scene** captured using a commodity depth sensor from a **single viewpoint**
- Deep learning + Robotics ⟶ more effective way to tackle Robotics issues

Reference

- Yuzhe Qin, Rui Chen, et al. S4G: A modal Single-view Single-Shot SE(3) Grasp Detection in Cluttered Scenes. In *Conference on Robot Learning (CoRL),* 2019.

- CoRL 2019 Osaka Day2, 2019. [Online]. Available: https://www.youtube.com/watch?v=b7StSnt85S4. [Accessed: 26- Apr- 2020].

Thank you!

Questions?