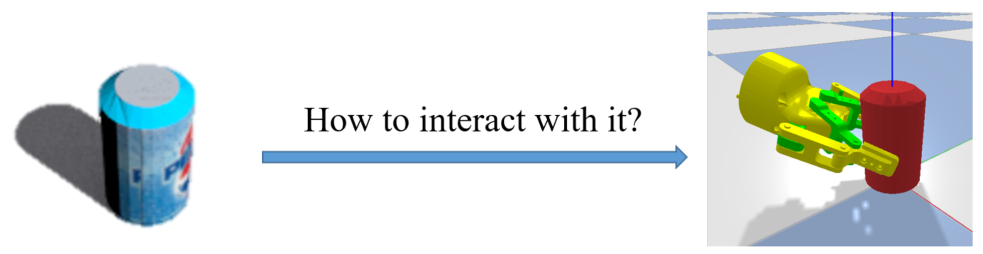
Grasp Proposal Networks: An End-to-End Solution for Visual Learning of Robotic Grasps

Chaozheng Wu*, Jian Chen*, Qiaoyu Cao, Jianchi Zhang, Yunxin Tai, Lin Sun, Kui Jia†

Introduction

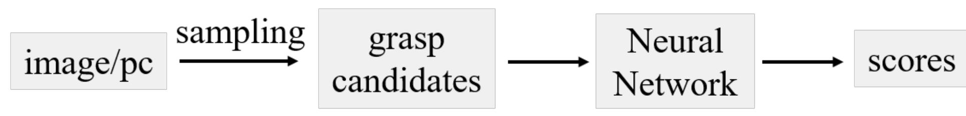
Visual Grasp Learning (VGL)



Learn to grasp from observation (image or point cloud).

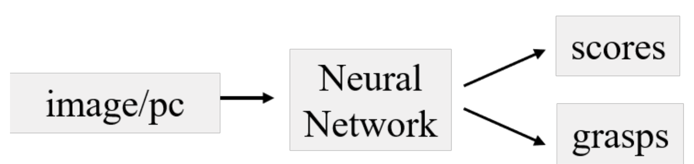
Existing Methods

Sampling based:



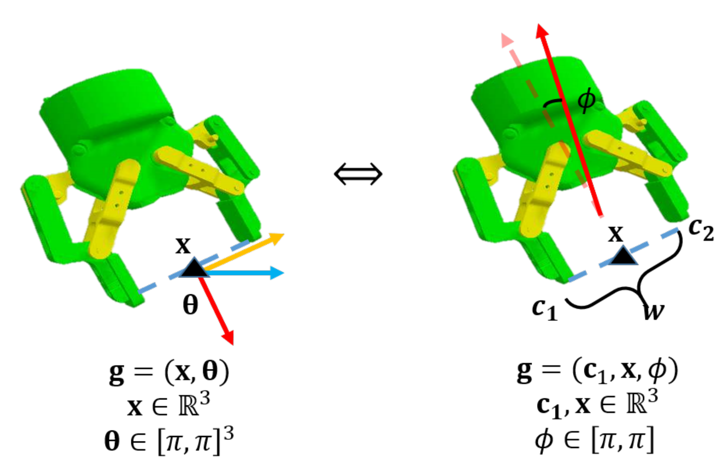
- Pros:
  easy to learn.
- Cons:
  a) finite number;
  b) may miss optimal grasps;
  c) time consuming.

Generation based:



- Pros:
  a) can generate a large number of grasps;
  b) can learn the optimal grasps;
  c) fast.
- Cons:
  a little hard to learn.

6-DOF Grasp Representation

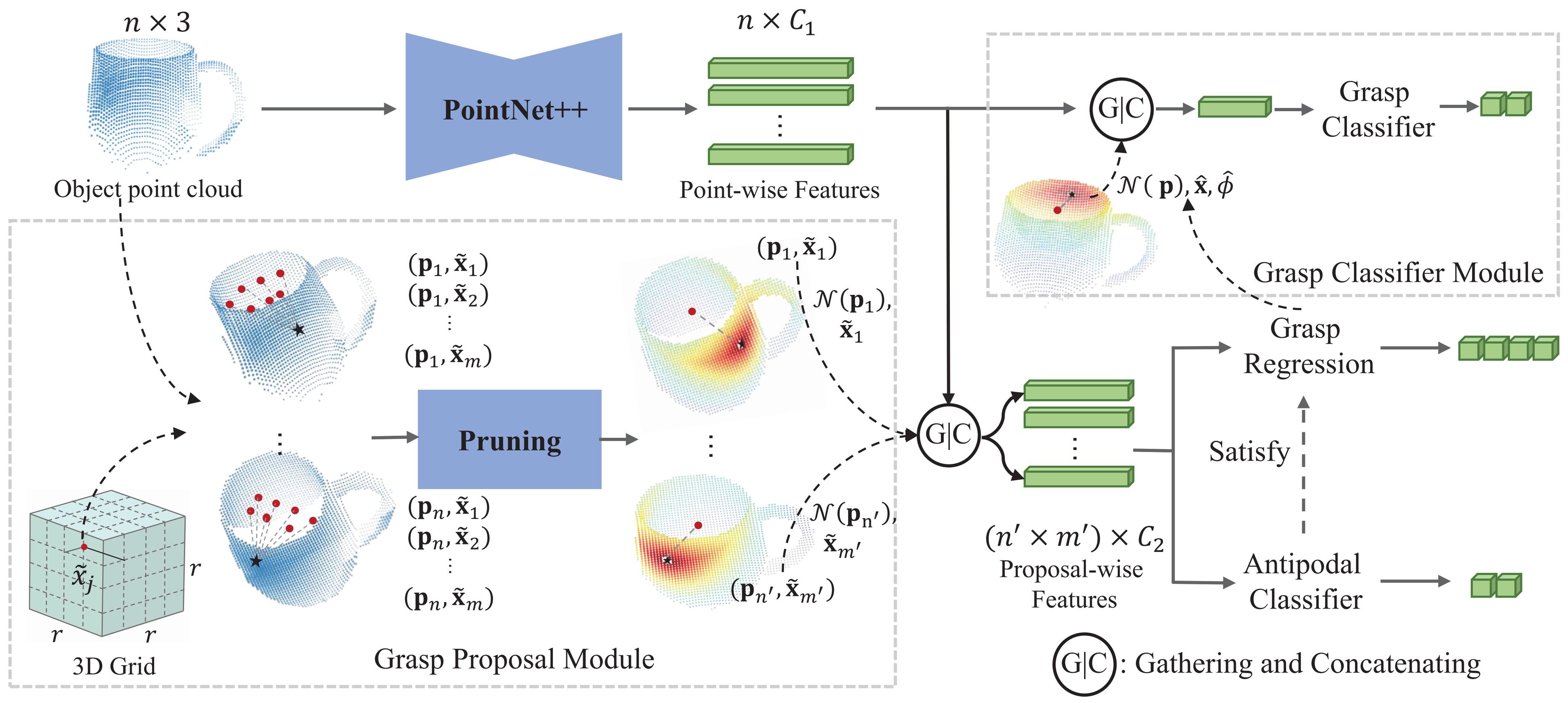


paper

code

Method

Grasp Proposal Network (GPNet)

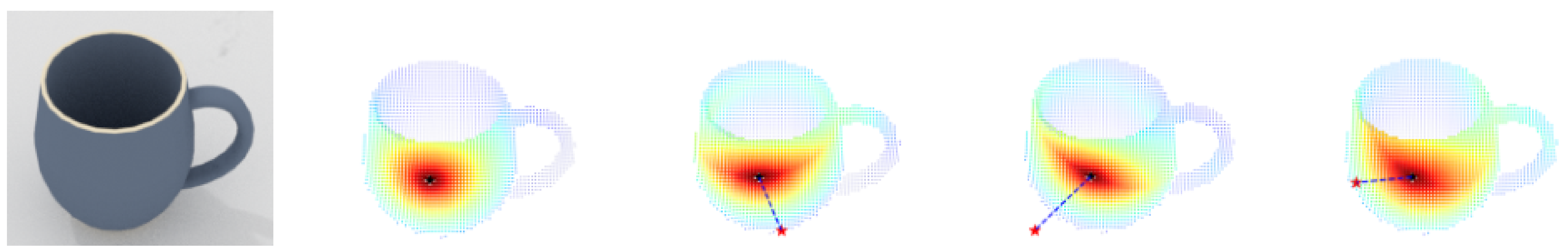


Overview

Features Extraction - We use PointNet++ to extract features for each point.

Grasp Proposal

- Each point $\mathbf{p}_i$ on point cloud can be a contact point; the vertices $\tilde{\mathbf{x}}_j$ of 3D grid can serve as anchors of grasp centers.
- Pair up $\mathbf{p}_i$ and $\tilde{\mathbf{x}}_j$, $(\mathbf{p}_i, \tilde{\mathbf{x}}_j)$ is our called grasp proposal.
- Pruning: a) remove the vertices out of the bounding box of object; b) remove some proposals far away from GT annotations.
- Anchor-dependent grasp features extraction:
  We use an anchor-dependent manner to determine $\mathbf{p}_i$ neighborhood $\mathcal{N}(\mathbf{p}_i)$.



(a) Mug (b) Distance-only (c) $(0, 0, 3)$cm (d) $(0, 3, 3)$cm (e) $(3, 3, 3)$cm

$$\mathcal{N}(\mathbf{p}_i, \widetilde{\mathbf{x}}_j) = \{\mathbf{p}_{i'} \mid d(\mathbf{p}_{i'}, \mathbf{p}_i) \cdot (|\cos(\overrightarrow{\mathbf{p}_i\mathbf{p}_{i'}}, \overrightarrow{\widetilde{\mathbf{x}}_j\mathbf{p}_i})| + 1) \leq \varepsilon\}$$

Antipodal Classifier

To check whether the grasp proposal satisfy the antipodal constraint.

$$\mathcal{L}_{AP}(\mathbf{p}_i, \widetilde{\mathbf{x}}_j) = -l^*_{AP}(\mathbf{p}_i, \widetilde{\mathbf{x}}_j) \log \hat{l}_{AP}(\mathbf{p}_i, \widetilde{\mathbf{x}}_j) - (1 - l^*_{AP}(\mathbf{p}_i, \widetilde{\mathbf{x}}_j)) \log \left(1 - \hat{l}_{AP}(\mathbf{p}_i, \widetilde{\mathbf{x}}_j)\right).$$

Grasp Regression

For the antipodal grasp proposal, we regress its offset to GT grasp center and the 'pitch' angle to get a precise grasp.

$$\mathcal{L}_{REG}(\mathbf{p}_i, \widetilde{\mathbf{x}}_j) = \|\Delta^{*+}_{\widetilde{\mathbf{x}}_j} - \Delta_{\widetilde{\mathbf{x}}_j}\| + \frac{1}{K}\sum_{k=1}^{K} \omega_k |\cos\hat{\phi} - \cos\phi^{*+}_k|$$

Grasp Classifier

We train a grasp classifier to score the predicted grasps, the regressed grasp center $\hat{\mathbf{x}}$, pitch angle $\hat{\phi}$, and $\mathcal{N}(\mathbf{p})$ are used as input features.

$$\mathcal{L}_{CLS}(\hat{\mathbf{g}}) = -l^*_{CLS}(\hat{\mathbf{g}}) \log \hat{l}_{CLS}(\hat{\mathbf{g}}) - (1 - l^*_{CLS}(\hat{\mathbf{g}})) \log \left(1 - \hat{l}_{CLS}(\hat{\mathbf{g}})\right).$$

Experiments

Dataset

- 226 CAD models from ShapeNetSem covering 8 categories (*bowl, bottle, mug, cylinder, cuboid, tissue box, sodacan,* and *toy car*), 196 objects for training and 30 for test.
- 22.6M grasp annotations ($\sim$ 100,000 per object, $\sim$ 23.6% positives and $\sim$ 76.4% negatives).
- 1000 RGB-D images per object rendered under arbitrary views.

Rule-based Evaluation

Criterion: (1) $\|\hat{\mathbf{x}} - \mathbf{x}^{*+}\|_2 \leq 25mm$ (2) $\|\hat{\theta} - \theta^{*+}\|_\infty \leq 30°$

Methods		success rate@k%				coverage rate@k%			
		10	30	50	100	10	30	50	100
6-DOF GraspNet	w/o refinement	0.867	0.850	0.711	0.534	0.039	0.039	0.094	0.132
	w/ refinement	0.867	0.833	0.733	0.534	0.063	0.063	0.122	0.168
GPNet-Naive	$r = 10, b = 22cm$	0.372	0.313	0.278	0.215	0.022	0.058	0.100	0.142
GPNet	$r = 3, b = 22cm$	0.844	0.833	0.800	0.649	0.051	0.107	0.191	0.273
	$r = 7, b = 22cm$	0.898	0.833	0.822	0.713	0.061	0.113	0.201	0.300
	$r = 10, b = 22cm$	**0.933**	**0.932**	0.820	**0.729**	0.068	0.144	0.199	0.307
	$r = 10, b = 10cm$	0.856	0.776	0.695	0.570	0.055	0.112	0.169	0.274
	$r = 10, b = 30cm$	0.900	0.869	**0.846**	0.712	**0.073**	**0.157**	**0.231**	**0.308**

Simulation-based Evaluation

Methods		$k = 10$	$k = 30$	$k = 50$	$k = 100$
GQCNN of planar grasp in DexNet		0.783	0.742	0.663	0.464
6-DOF GraspNet	w/o refinement	0.433	0.367	0.311	0.207
	w/ refinement	0.800	0.594	0.508	0.354
GPNet-Naive	$r = 10, b = 22cm$	0.100	0.095	0.083	0.054
GPNet	$r = 3, b = 22cm$	0.644	0.637	0.561	0.371
	$r = 7, b = 22cm$	0.767	0.711	0.656	0.557
	$r = 10, b = 22cm$	**0.900**	**0.761**	**0.723**	**0.588**
	$r = 10, b = 10cm$	0.494	0.433	0.393	0.253
	$r = 10, b = 30cm$	0.833	0.702	0.679	0.574

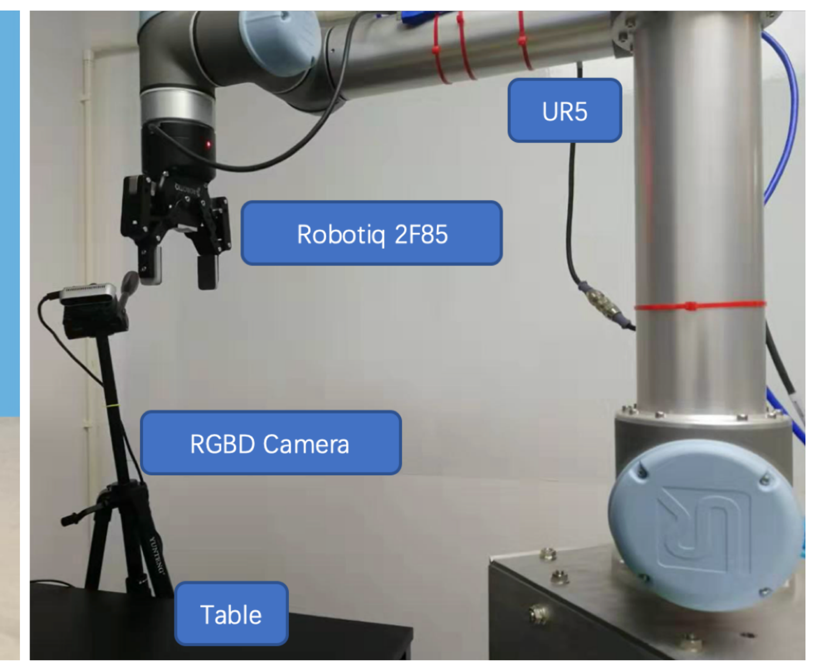
# Avg. annotations per object	Accuracy
10K	0.650
50K	0.730
100K	**0.900**

Ratio of training set	Accuracy
1/4	0.522
1/2	0.700
1	**0.900**

Robot Experiment

Real test setting:

Objects



Robot disposition

Real test results:

Object index	#1	#2	#3	#4	#5	#6	#7	#8	#9	#10	
GPNet	2/3	3/3	3/3	3/3	3/3	3/3	3/3	2/3	3/3	2/3	
6-DOF GraspNet	2/3	2/3	3/3	2/3	1/3	0/3	2/3	3/3	1/3	3/3	
Object index	**#11**	**#12**	**#13**	**#14**	**#15**	**#16**	**#17**	**#18**	**#19**	**#20**	**Overall**
GPNet	3/3	2/3	2/3	3/3	3/3	2/3	3/3	0/3	3/3	3/3	85%
6-DOF GraspNet	3/3	3/3	3/3	3/3	2/3	3/3	3/3	0/3	3/3	2/3	73%

video