Robust Grasp Planning Under Partial Point-Cloud Observation

A THESIS

submitted by

Ritika Chawla
2021CS60123

under the guidance of

Prof. Subhasis Banerjee

for the award of the degree

of

Doctor of Philosophy

Department Of Computer Science and Engineering
INDIAN INSTITUTE OF TECHNOLOGY DELHI

THESIS CERTIFICATE

This is to certify that the thesis titled **Robust Grasp Planning Under Partial Point-Cloud Observation**, submitted by **Ritika Chawla**, to the Indian Institute of Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by her under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Subhasis Banerjee
Professor
Dept. of Computer Science
IIT-Delhi, 110016

ACKNOWLEDGEMENTS

I would like to thank my advisor for his guidance and patience throughout this thesis, my doctoral committee for their feedback at each stage of this work, and the robotics lab's technical staff for their help maintaining the manipulator and depth-camera rig used in every experiment in this thesis.

ABSTRACT

Robotic grasp planners trained on complete, noise-free 3D models perform substantially worse when deployed against real depth-camera point clouds, which are partial (occluded by the object itself and by neighbouring clutter) and noisy. This thesis studies grasp planning under exactly this partial-observation regime, rather than assuming the complete object model available in simulation.

We propose a grasp-quality estimator that is trained directly on partially-observed point clouds, using a data-augmentation scheme that simulates camera occlusion during training rather than relying on complete-model training data alone. The estimator is evaluated on a physical manipulator against 22 household objects with varying geometry and surface material, under three camera-viewpoint conditions of increasing occlusion severity.

Trained on occlusion-augmented data, the estimator achieves a grasp success rate of 84% under the most severe occlusion condition tested, compared to 61% for an equivalent estimator trained only on complete point clouds — a 23-percentage-point gap that grows, rather than shrinks, as occlusion severity increases. The thesis closes by analysing which object geometries remain hard for the occlusion-augmented estimator, and shows that thin, elongated objects account for a disproportionate share of the remaining failures.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS **i**

ABSTRACT **ii**

LIST OF TABLES **v**

LIST OF FIGURES **vi**

1 INTRODUCTION **1**

1.1 Contributions 1

2 BACKGROUND **2**

2.1 Learned Grasp-Quality Estimation 2

2.2 Sim-to-Real Transfer for Grasping 2

2.3 Gap Addressed by This Thesis 2

3 APPROACH **3**

3.1 Occlusion-Augmented Training 3

3.2 Clutter Sampling 3

3.3 Network Architecture 3

4 RESULTS **5**

4.1 Experimental Setup 5

4.2 Grasp Success Rate 5

4.3 Failure Analysis 5

4.4 Discussion 6

A SUPPLEMENTARY **7**

REFERENCES **8**

LIST OF TABLES

4.1 Grasp success rate by training condition and occlusion severity, 22 objects, 10 trials each. 5

LIST OF FIGURES

Chapter 1

INTRODUCTION

Robotic grasp planning has advanced substantially on the strength of learned grasp-quality estimators trained on large simulated datasets of complete 3D object models [1]. In deployment, however, a robot never observes a complete model: a single depth camera sees only the visible surface of an object, occluded further by neighbouring clutter, and the resulting point cloud is both partial and noisy.

This thesis studies grasp planning under exactly this partial-observation regime. Rather than treating the sim-to-real gap as purely a rendering or noise-modelling problem, we ask a more specific question: does a grasp-quality estimator trained on complete point clouds generalise to partial ones, or does the partiality itself require a different training signal?

1.1 Contributions

- An occlusion-augmentation training scheme that simulates camera-realistic partial observation during training, described in Chapter 3.
- A physical-robot evaluation across 22 household objects and three occlusion severities, in Chapter 4, quantifying the gap between complete-model and occlusion-augmented training.

Chapter 2

BACKGROUND

2.1 Learned Grasp-Quality Estimation

Grasp-quality estimators trained on large simulated datasets, such as Dex-Net [1], have driven most recent progress in robotic grasping. These datasets are built from complete object meshes, rendered from many viewpoints, but the grasp-quality label attached to each candidate grasp is computed from the complete mesh regardless of which viewpoint generated the training image — so the estimator is never explicitly trained to cope with an incomplete point cloud.

2.2 Sim-to-Real Transfer for Grasping

Prior work addressing the sim-to-real gap for grasping has focused primarily on visual domain randomisation — varying rendering parameters, textures, and camera noise models — to make simulated training images look more like real ones [3]. This thesis instead focuses on *geometric* domain gap: even a perfectly rendered, noise-free depth image of a partially-occluded scene is geometrically different from a complete-model training example, independent of any visual rendering imperfection.

2.3 Gap Addressed by This Thesis

No prior work isolates the effect of point-cloud partiality itself, as distinct from sensor noise or visual domain gap, on grasp-quality estimator performance. Chapter 3 develops a training scheme that targets this gap specifically, and Chapter 4 quantifies it.

Chapter 3

APPROACH

3.1 Occlusion-Augmented Training

Standard grasp-quality datasets pair a candidate grasp with a label computed from the object's complete mesh. To train an estimator that copes with partial observation, we instead render each training example from a randomly sampled camera viewpoint against the complete scene (object plus randomly placed clutter), producing a genuinely partial point cloud, and compute the grasp-quality label against that same partial observation — not the complete mesh.

$$\mathcal{L}(\theta) = \mathbb{E}_{(o,c,g)\sim\mathcal{D}}\left[\ell\big(f_\theta(\text{PartialPC}(o,c)),\, q(g,o)\big)\right] \tag{3.1}$$

where o is the object, c a sampled camera viewpoint, g a candidate grasp, $\text{PartialPC}(o,c)$ the partial point cloud rendered from that viewpoint, and $q(g,o)$ the grasp-quality label computed from the complete object mesh (the label itself still uses the complete mesh, since ground-truth grasp quality is a property of the object, not of what the camera happened to see — only the *input* to f_θ is partial).

3.2 Clutter Sampling

Clutter objects are drawn from the same object pool as target objects and placed at randomly sampled positions and orientations around the target, subject to a minimum clearance constraint so that at least one collision-free grasp remains feasible. This is what produces the varying occlusion severities evaluated in Chapter 4: light clutter yields near-complete point clouds, and heavy clutter yields point clouds covering less than a third of the target object's visible surface.

3.3 Network Architecture

The estimator itself is a PointNet++-style point-cloud encoder [2] feeding a grasp-quality regression head, chosen so that the architecture itself is not novel — the contribution is

the training data distribution, not the network design, and Chapter 4 holds the architecture fixed across both the occlusion-augmented and complete-model training conditions to isolate that effect.

© 2026, *Indian Institute of Technology Delhi*

Chapter 4

RESULTS

4.1 Experimental Setup

The estimator was evaluated on a physical manipulator (a 7-DOF arm with a parallel-jaw gripper) fitted with a wrist-mounted depth camera, against 22 household objects spanning a range of geometries (boxes, cylinders, irregular and thin/elongated shapes) and surface materials (matte plastic, glossy ceramic, fabric). Three occlusion conditions were tested per object: light clutter, moderate clutter, and heavy clutter, following the clutter sampling procedure of Chapter 3.

Two estimators were compared, identical in architecture and differing only in training data: one trained on complete point clouds (the conventional approach), and one trained with the occlusion-augmentation scheme of Chapter 3.

4.2 Grasp Success Rate

Table 4.1: Grasp success rate by training condition and occlusion severity, 22 objects, 10 trials each.

Training condition	Light clutter	Moderate clutter	Heavy clutter
Complete point clouds	91%	78%	61%
Occlusion-augmented (this work)	93%	89%	84%

Table 4.1 shows the two training conditions perform similarly under light clutter (near-complete point clouds), but diverge sharply as occlusion severity increases: the gap widens from 2 percentage points under light clutter to 23 percentage points under heavy clutter, confirming that the occlusion-augmentation scheme's benefit is specifically in the partial-observation regime it targets, not a general improvement that would show up even without occlusion.

4.3 Failure Analysis

Of the occlusion-augmented estimator's remaining failures under heavy clutter, 61% occurred on the four thin, elongated objects in the test set (a ruler, a spatula, a pair of

chopsticks, and a letter opener) despite these objects making up only 18% of the object pool. These objects present very little visible surface area under heavy occlusion, which is consistent with the estimator's failures being driven by genuinely insufficient observed geometry, rather than a training-data artefact.

4.4 Discussion

The results support the thesis's central claim: point-cloud partiality is a distinct challenge from sensor noise or visual domain gap, and training data that matches the partiality the estimator will see at deployment time closes most of the resulting performance gap. The residual failures on thin, elongated objects suggest a geometric limit to what occlusion augmentation alone can fix — these objects may need active viewpoint selection rather than a passive training-data fix.

© 2026, *Indian Institute of Technology Delhi*

Appendix A

SUPPLEMENTARY

This appendix would normally list the full 22-object test set with their geometric category and surface material, and the per-object grasp success rates underlying Table 4.1 in Chapter 4. It is omitted from this starter template; a real thesis appendix would include the full object list and per-object breakdown as a table.

REFERENCES

[1] J. Mahler, J. Liang, S. Niyaz, M. Laskey, R. Doan, X. Liu, J. A. Ojea, and K. Goldberg. Dex-net 2.0: Deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. In *Robotics: Science and Systems (RSS)*, 2017.

[2] C. R. Qi, L. Yi, H. Su, and L. J. Guibas. PointNet++: Deep hierarchical feature learning on point sets in a metric space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[3] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, 2017.

CITATIONS

Parts of the work in this thesis are based on the following publication:

- R. Chawla and S. Banerjee, "Occlusion-Augmented Training for Grasp-Quality Estimation from Partial Point Clouds," *under submission*, 2026.