Supplementary Material for Dex-Net 2.0: Deep Learning to Plan Robust Grasps With Synthetic Point Clouds and Analytic Grasp Metrics

Jeffrey Mahler*, Jacky Liang*, Sherdil Niyaz*, Michael Laskey*, Richard Doan*, Xinyu Liu*,
Juan Aparicio Ojea†, and Ken Goldberg*
*Dept. of EECS, University of California, Berkeley
Email: {jmahler, jackyliang, sniyaz, laskeymd, rdoan, xinyuliu, goldberg}@berkeley.edu
† Siemens Corporation, Corporate Technology
Email: juan.aparicio@siemens.com

This document contains details on the parameters of our graphical model and the algorithms used to sample grasps in the robust grasping policy (fixed antipodal grasp sampling and the cross entropy method).

I. Parameters of Graphical Model

Our graphical model is illustrated in Fig. 1 and models $p(S, \mathbf{g}, \mathbf{x}, \mathbf{y})$ as the product of a state distribution $p(\mathbf{x})$, an observation model $p(\mathbf{y}|\mathbf{x})$, a grasp candidate model $p(\mathbf{g}|\mathbf{x})$, and a grasp success model $p(S|\mathbf{g}, \mathbf{x})$.

We model the state distribution as $p(\mathbf{x}) = p(\gamma)p(\mathcal{O})p(T_o|\mathcal{O})p(T_c)$. We model $p(\gamma)$ as a Gaussian distribution $\mathcal{N}(0.5, 0.1)$ truncated to $[0, 1]$. We model $p(\mathcal{O})$ as a discrete uniform distribution over 3D objects in a given dataset. We model $p(T_o|\mathcal{O}) = p(T_o|T_s)p(T_s|\mathcal{O})$, where is $p(T_s|\mathcal{O})$ is a discrete uniform distribution over object stable poses and $p(T_o|T_s)$ is uniform distribution over 2D poses: $\mathcal{U}([-0.1, 0.1] \times [-0.1, 0.1] \times [0, 2\pi))$. We compute stable poses using the quasi-static algorithm given by Goldberg et al. [1]. We model $p(T_c)$ as a uniform distribution on spherical coordinates $r, \theta, \varphi \sim \mathcal{U}([0.65, 0.75] \times [0, 2\pi) \times [0.05\pi, 0.1\pi])$, where the camera optical axis always intersects the center of the table.

Our distribution over grasps is a uniform distribution over pairs of antipodal points on the object surface that are parallel to the table plane. We sample from this distribution for a fixed coefficient of friction $\mu = 0.6$ and reject samples outside the friction cone or non-parallel to the surface.

We model images as $\mathbf{y} = \alpha * \hat{\mathbf{y}} + \epsilon$ where $\hat{\mathbf{y}}$ is a rendered depth image created using OSMesa offscreen rendering. We model α as a Gamma random variable with shape$= 1000.0$ and scale=0.001. We model ϵ as Gaussian Process noise drawn with measurement noise $\sigma = 0.005$ and kernel bandwidth $\ell = \sqrt{2}px$.

We compute grasp robustness metrics using the graphical model and noise parameters of [4].

II. Grasp Sampling Methods

Our goal is to learn policy parameters θ that maximize the success rate of planned grasps over a distribution on point clouds that can be generated from a set of possible objects $\mathcal{D}$:

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmax}}\ \mathbb{E}_{p(\mathbf{y}|\mathcal{D})}\left[Q(\pi_\theta, \mathbf{y})\right]. \qquad \text{(II.1)}$$

We propose to approach the objective of equation II.1 by using a greedy grasping policy $\pi_\theta(\mathbf{y}) = \operatorname{argmax}_{\mathbf{u} \in \mathcal{C}} Q_\theta(\mathbf{u}, \mathbf{y})$ with respect to a GQ-CNN robustness function $Q_\theta(\mathbf{u}, \mathbf{y})$, where $\mathcal{C}$ specifies constraints on candidate grasps such as kinematic feasibility. We explore two implementations of the robust grasping policy: (1) sampling a large, fixed set of antipodal grasps and choosing the most robust one and (2)

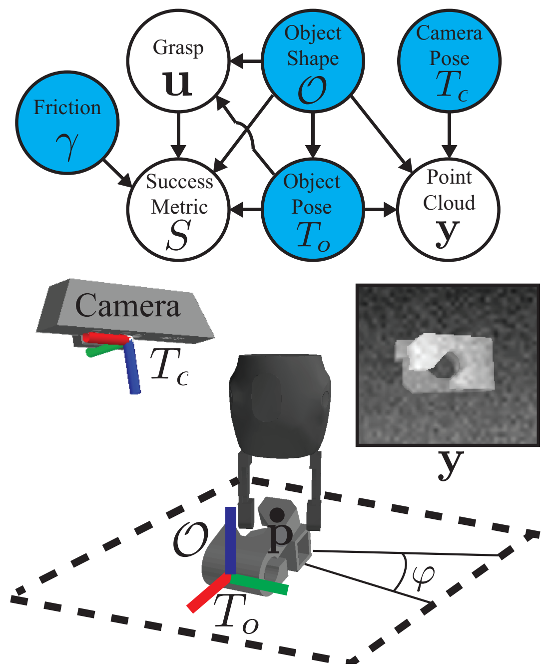

Fig. 1: Graphical model for robust parallel-jaw grasping of objects on a table surface based on point clouds. Blue nodes are variables included in the state representation. Object shapes $\mathcal{O}$ are uniformly distributed over a discrete set of object models and object poses T_o are distributed over the object's stable poses and a bounded region of a planar surface. Grasps $\mathbf{g} = (\mathbf{p}, \theta)$ are sampled uniformly from the object surface using antipodality constraints. Given the coefficient of friction γ we evaluate an analytic succes metric S for a grasp on an object. A synthetic point cloud $\mathbf{y}$ is generated from 3D meshes based on the camera pose T_c, object shape, and pose and corrupted with multiplicative and Gaussian Process noise.

optimizing for the most robust grasp using derivative free optimization.

A. Antipodal Grasp Sampling

The antipodal grasp sampling method used in the paper is designed to sample antipodal grasps specified as a planar pose, angle, and height with respect to a table. The algorithm is detailed in Algorithm 1. We first threshold the depth image to find areas of high gradient. Then, we use rejection sampling over pairs of pixels to generate a set of candidate antipodal grasps, incrementally increasing the friction coefficient until a desired number of grasps is reached in case the desired number cannot be achieved with a smaller friction coefficient. We convert antipodal grasps in image space to 3D by assigning discretizing the gripper height between the height of the grasp center pixel relative and the height of the table surface itself.

This grasp sampling method is used for all image based grasp planners in the paper. We used $M = 1000$, K set to the intrinsics of a Primesense Carmine 1.08, T_c determined by chessboard registration, $g = 0.0025m$, $\mu_\ell = 0.4$, $\delta_\mu = 0.2$, $N = 1000$, and $\delta_h = 0.01m$.

Input: Depth image $\mathbf{y}$, Number of grasps M, Camera Intrinsics Matrix K, Camera pose T_c, Depth gradient threshold g, Min friction coef μ_ℓ, Friction coef increment δ_μ, Max samples per friction coef N, Gripper height resolution δ_h
Result: $\mathcal{G}$, set of candidate grasps
`// Compute depth edges`
$G_x = \nabla_x \mathbf{y}, G_y = \nabla_y \mathbf{y}$;
$\mathcal{E} = \{\mathbf{u} \in \mathbb{R}^2 : G_x(\mathbf{u})^2 + G_y(\mathbf{u})^2 > g\}$;
`// Find antipodal pairs`
$\mathcal{G} = \{\}$, $i = 0$, $j = 0$;
while $|\mathcal{G}| < M$ *and* $\mu <= 1.0$ **do**
  $\mathbf{u}, \mathbf{v}$ =UniformRandom($\mathcal{E}, 2$);
  if *Antipodal(*$\mathbf{u}, \mathbf{v}, \mu$*)* **then**
    `// Compute piont in world coordinates`
    $\mathbf{c} = 0.5 * (\mathbf{u} + \mathbf{v})$;
    $\mathbf{p}_c$ =Deproject($K, \mathbf{y}, \mathbf{c}$);
    $\mathbf{p} = T_c * \mathbf{p}_c$;
    $h = \mathbf{p}.z$;
    `// Add all heights`
    while $h > 0$ **do**
      $\mathcal{G} = \mathcal{G} \cup \{\mathbf{g}(\mathbf{u}, \mathbf{v}, h)\}$;
      $h = h - \delta_h$;
    end
  end
  $i = i + 1$, $j = j + 1$;
  `// Update friction coef`
  if $j >= N$ **then**
    $\mu = \mu + \delta_\mu$;
    $j = 0$;
  end
end
return $\mathcal{G}$;

Algorithm 1: Antipodal Grasp Sampling from a Depth Image

B. Derivative Free Optimization

One problem with choosing a grasp from a fixed set of candidates is that the set of candidates may all have a low probability of success. This can be difficult when an object can only be grasped in a small set of precise configurations, such as the example in Fig. 2. Some of these failures can be seen in the right panel of the failure modes figure in the original paper.

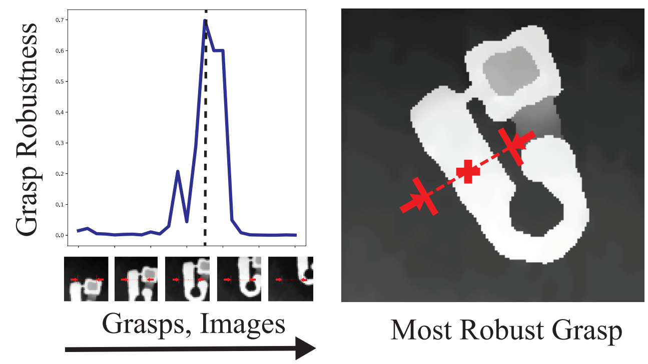

Fig. 2: (Left) Grasp robustness predicted by a Grasp Quality Convolutional Neural Network (GQ-CNN) trained with Dex-Net 2.0 over the space of depth images and grasps for a single point cloud collected with a Primesense Carmine. As the center of the gripper moves from the top to the bottom of the image the GQ-CNN prediction stays near zero and spikes on the most robust grasp (right), for which the gripper fits into a small opening on the object surface. This suggests that the GQ-CNN has learned a detailed representation of the collision space between the object and gripper. Furthermore, the sharp spike suggests that it may be difficult to plan robust grasps by randomly sampling grasps in image space. We consider planning the most robust grasp using the cross-entropy method on the GQ-CNN response.

In our second generalization study we addressed this problem using the cross entropy method (CEM) [2, 3], a form of derivative-free optimization, to optimize for the most robust grasp by iteratively resampling grasps from a learned distribution over robust grasps and updating the distribution. The method, illustrated in Algorithm 2, models the distribution on promising grasps using a Gaussian Mixture Model (GMM) and seeds the initial set of grasps with antipodal point pairs using Algorithm 1 with no iterative friction coefficient updates. The algorithm takes as input the number of CEM iterations m, the number of initial grasps to sample n, the number of grasps to resample from the model c, the number of GMM mixure components k, a friction coefficient μ, and elite percentage γ, and the GQ-CNN Q_θ, and returns an estimate of the most robust grasp $\mathbf{u}$. In our generalization experiment we used $m = 3$, $n = 100$, $c = 50$, $\mu = 0.8$, $k = 3$, and $\gamma = 25\%$. The qualitative performance of our method on several examples from our experiments is illustrated in Fig. 3.

Input: Num rounds m, Num inital samples n, Num CEM samples c, Num GMM mixture k, Friction coef μ, Elite percentage γ, Robustness function Q_θ
Result: $\mathbf{u}$, most robust grasp
$\mathcal{U} \leftarrow$uniform set of n antipodal grasps;
for *i = 1, ..., m* **do**
  $\mathcal{E} \leftarrow$top $\gamma-$percentile of grasps ranked by Q_θ;
  $M \leftarrow$GMM fit to $\mathcal{E}$ with k mixtures;
  $G \leftarrow c$ iid samples from M;
end
return $\underset{\mathbf{u} \in \mathcal{U}}{\text{argmax}}\ Q_\theta(\mathbf{u}, \mathbf{y})$;

Algorithm 2: Robust Grasping Policy using the Cross Entropy Method on a Learned GQ-CNN

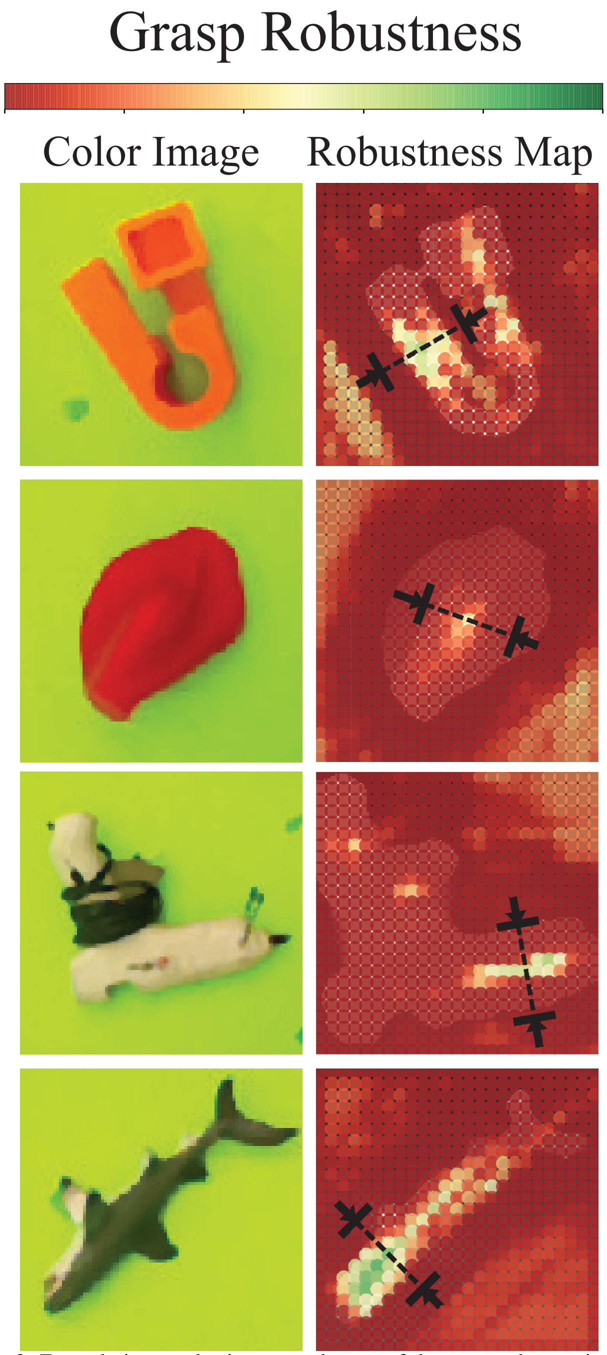

Fig. 3: Example input color images and maps of the grasp robust estimated by the GQ-CNN over grasp centers for a constant grasp axis angle in image space and height above the table, with the grasp planned by our CEM-based robust grasping policy shown in black. CEM is able to find precise robust grasping locations encoded by the GQ-CNN that are very close to the global maximum for the given grasp axis and height. The GQ-CNN also appears to assign non-zero robustness to several grasps that completely miss the object. This is likely because no such grasps are in the training set, and future work could augment the training dataset to avoid these grasps.

References

[1] Ken Goldberg, Brian V Mirtich, Yan Zhuang, John Craig, Brian R Carlisle, and John Canny. Part pose statistics: Estimators and experiments. *IEEE Trans. Robotics and Automation*, 15(5):849–857, 1999.

[2] Sergey Levine, Peter Pastor, Alex Krizhevsky, and Deirdre Quillen. Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. *arXiv preprint arXiv:1603.02199*, 2016.

[3] Reuven Y Rubinstein, Ad Ridder, and Radislav Vaisman. *Fast sequential Monte Carlo methods for counting and optimization*. John Wiley & Sons, 2013.

[4] Daniel Seita, Florian T Pokorny, Jeffrey Mahler, Danica Kragic, Michael Franklin, John Canny, and Ken Goldberg. Large-scale supervised learning of the grasp robustness of surface patch pairs. In *Proc. IEEE Int. Conf. on Simulation, Modeling, and Programming of Autonomous Robots (SIMPAR)*. IEEE, 2016.