1 Motivations

1. Learning to grasp from purely synthetic data.
2. They only consider grasping singulated obj in this paper. Unclear how it will work in clutter.
3. The main component in the approach is a Grasp Quality CNN (GQ-CNN), which predicts binary success label for depth image - grasp configuration pair.
4. To pick grasp, they just use CEM.

2 Learning the GQ-CNN

1. Generate synthetic training data of 6.7 million synthetic point cloud, parallel-jaw grasp, and robust analytic grasp metric from 1500 3D models.
2. The GQ-CNN receives as input a depth map y and a grasp configuration u and predict binary label S.
3. S is computed analytically from the state of environment x and the grasp conf. u using robust epsilon quality, which measures grasp robustness to uncertainty in friction and gripper pose.
4. Each obj is labelled with a set of up to 100 parallel-jaw grasps.
5. Depth image of the 3D model is rendered using a pinhole camera model.
6. Given a rendered depth image and a sampled grasp conf., the img is transformed to align the grasp pixel location with the img center and the grasp axis with the middle row of the img.
7. They argue that this img-gripper alignment removes the need to learn rotational invariances that can be modeled by known, computationally efficient transformations.
8. Such alignment also allows the network to evaluate any grasp orientation in the image rather than a pre-defined discrete set.

3 Experiments

1. For known obj, their methods achieve 99% success rate and 94% precision.
2. For novel obj, their methods achieve 80% success rate and 100% precision.

4 Questions

1. Why is it that depth image is more transferable from synthetic to real than colored image?
2. Is it possible to synthesize novel 3D model for training?