

# Edge Grasp Network: A Graph-Based SE(3)-invariant Approach to Grasp Detection

Haojie Huang   Dian Wang   Xupeng Zhu   Robin Walters   Robert Platt  
 Khoury College of Computer Science, Northeastern University  
 {huang.haoj; wang.dian; zhu.xup; r.walters; r.platt} @northeastern.edu

**Abstract**—Given point cloud input, the problem of 6-DoF grasp pose detection is to identify a set of hand poses in SE(3) from which an object can be successfully grasped. This important problem has many practical applications. Here we propose a novel method and neural network model that enables better grasp success rates relative to what is available in the literature. The method takes standard point cloud data as input and works well with single-view point clouds observed from arbitrary viewing directions.

## I. INTRODUCTION

Grasp detection [1], [2], [3] is a critical robotic skill. The robot first observes a scene containing objects in the form of images, voxels, or point clouds, and detects a set of viable grasp poses from which an object may be grasped stably. There are two general approaches to grasp detection: SE(2) methods where the model reasons in terms of a top-down image of the scene (e.g. [4], [5], [6], [7], [8]), and SE(3) methods where the model reasons in terms of a point cloud or voxel grid (e.g. [1], [3], [9], [10]). SE(3) methods have a distinct advantage over SE(2) methods because they can find side grasps more easily and they are easier to apply in general robotics settings where only a point cloud is available. Unfortunately, SE(3) methods are generally much more complex, so SE(2) models are often preferred.

This paper tackles the problem of SE(3) grasping with a novel grasp detection model that we call the *Edge Grasp Network*. The model is based on a novel representation of a 6-DoF grasp that uses a pair of vertices in a graph. Given a single approach point (a position the hand will approach), we define a KNN graph that contains all the points in the point cloud that are within a fixed radius of the approach point. Each point in this KNN graph corresponds to an orientation of the gripper and, when paired with the approach point, defines a distinct 6-DOF grasp pose. We infer the quality of all such grasps simultaneously using a graph neural network to compute point features. These are used to compute features on the edges connecting point pairs and ultimately evaluate the probability of grasp success for each grasp candidate associated with the approach point. In order to evaluate the qualities of grasps defined with respect to several approach points in a single forward pass through the model, we can simply form a batch comprised of several individual KNN graphs. This approach is novel relative to the literature in three ways: 1) First, our method of defining unique grasp candidates in terms of a pair of vertices in a

graph is new; 2) Second, our inference model using a graph neural network defined with respect to a single approach point is novel; 3) Third, our model is the first SE(3) grasp method that incorporates SO(3) equivariance.

Our approach has several advantages over prior work. First, and perhaps most importantly from a practical perspective, our method works well with single-view point clouds of a scene taken from arbitrary directions. The ability to use a single arbitrary direction is especially relevant because many methods like VGN [11] or GIGA [9] are trained specifically for multiple camera views or a particular viewpoint direction and do not generalize well to point clouds generated from a novel viewing direction. The consequence is that it is difficult to apply these methods out of the box to practical problems. Another key advantage of our method is that we can easily provide the approximate position of a desired grasp as an *input* to the model. If we want to grasp a tool by its handle, for example, this is easily achieved by only considering approach positions and contact locations on the handle. Finally, this work goes to great lengths to perform fair comparisons with other grasp detection methods in the literature, both in simulation and on physical systems. The results indicate that our method has better grasp success rates than several strong baselines including VGN [11], VPN [10], GIGA [9], and the method of Zhu *et al.* [12].

## II. RELATED WORK

### A. 6-DoF grasping methods

There are two main types of 6-DoF grasping methods in recent research. **Sample-based methods** like GPD [2], PoinetNetGDP [13], GraspNet [3] that are often comprised of a grasp sampler module and a grasp evaluator module. These methods often require long training time and execution time since each grasp is represented and evaluated individually. In contrast, our method uses shared features to represent different grasps and achieve more computation efficiency. **Element-wise prediction methods** include point-based methods [10], [14], [15], [16] and volumetric-based methods [11], [9]. They estimate grasp qualities for all interesting points or voxels with a single feed-forward propagation. For instance, S4G [14] generates each point feature through PointNet++ [17] and predicts the grasp quality and the grasp pose together. REGNet [16] considers the geometry of radius sphere around the sampled points and regresses

the orientations. However, the grasp distribution is a multi-modal function and regression methods only predict one grasp pose for a single point, which may cause ambiguity when multiple graspable poses are valid in that position. Classification methods can generate the distributions over multiple grasps at a single point, but copious amounts of data are often required. Volumetric-based methods [11], [9] use well-structured voxels instead of an unordered set of points. The memory requirements for voxel grids or SDFs are cubic in the resolution of the grid and therefore severely limit the resolution at which the method can be applied.

### B. Grasp Pose Representation

Grasp representation matters in evaluating and refining grasp poses. Most sample-based methods have a clear representation of grasp pose. GPD [2] projects the points around the gripper into canonical planes; PoinetGPD [13] feeds the points inside the gripper to PointNet; GraspNet [3] represents the grasp pose with a set of points of the gripper. On the other hand, element-wise methods [10], [14], [15], [16], [11], [9] often avoid representing grasp explicitly. Since the relative pose between the gripper and the point/voxel is unclear, they have to do regressions or classifications of some elements of the grasp pose. Our method has a clear representation of the grasp pose and satisfies the multi-modal property of the grasp distribution and the friction constraint [18] of the contact point.

### C. Symmetries in Manipulation

Symmetries and equivariance have been shown to improve learning efficiency and generalization ability in many manipulation tasks [12], [19], [20], [21]. *Zhu et al.* [12] decouples rotation and translation symmetries to enable the robot to learn a planar grasp policy within 1.5 hours; *Huang et al.* [20] achieve better sample efficiency and faster convergence speed in planar pick and place tasks with the use of  $C_n \times C_n$  equivariance; *Simeonov et al.* [21] use Vector Neurons to get SE(3)-equivariant object representations so that the model can manipulate objects in the same category with a few training demonstrations. Our method also leverages SE(3) symmetry to learn faster and generalize better on 6-DoF grasping.

## III. PROBLEM STATEMENT

The grasp detection problem is to locate a set of grasp poses in SE(3) for a parallel-jaw gripper given input about the scene in the form of a point cloud. Denote the point cloud observation as  $P = \{p_i \in \mathbb{R}^3\}_{i=1}^n$ , where  $n$  is the number of points. For each point  $p \in P$ , we will assume that an estimate of the object surface normal  $n_p \in S^2$  can be calculated. Although it is not required, we generally assume that this point cloud is generated by a single depth camera. A grasp pose of the gripper is parameterized  $\alpha = (C, R) \in \text{SE}(3)$ , where  $C \in \mathbb{R}^3$  is the location of the center of the gripper and  $R \in \text{SO}(3)$  represents its orientation. The grasp detection problem is to find a function

$$S: P \mapsto \{\alpha_i \in \text{SE}(3)\}_{i=1}^m, \quad (1)$$

that maps  $P$  onto  $m$  grasp poses detected in the scene. The grasp evaluation problem is to find a function  $\Phi: (P, \alpha) \mapsto [0, 1]$ , that denotes the quality of grasp  $\alpha$ . Notice that  $\Phi$  is invariant to translation and rotation in the sense that  $\Phi(g \cdot P, g \cdot \alpha) = \Phi(P, \alpha)$  for an arbitrary  $g \in \text{SE}(3)$ . In other words, the predicted quality of a grasp attempt should be invariant to transformation of the object to be grasped and the grasp pose by the same rotation and translation.

## IV. METHOD

### A. Grasp Pose Representation

We represent a grasp as a pair of points in the cloud,  $(p_a, p_c) \in P^2$ .  $p_a$  is considered to be the *approach point* and  $p_c$  is the *contact point*. Assuming that we can estimate the object surface normal  $n_c$  at point  $p_c$ ,  $(p_a, p_c)$  defines a grasp orientation  $R$  where the gripper fingers move parallel to the vector  $n_c$  and the gripper approaches the object

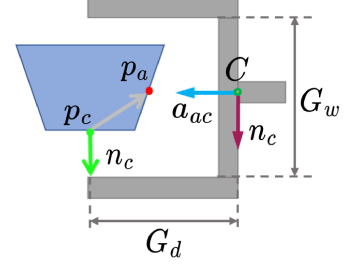


Fig. 1. Grasp pose defined by the edge grasp  $(p_a, p_c)$ . The reference frame of the gripper is illustrated by the RGB coordinate system.  $G_w$  and  $G_d$  are the gripper width and gripper depth.

along the vector  $a_{ac} = n_c \times (n_c \times (p_a - p_c))$ . This is illustrated in Figure 1. The gripper center  $C$  is positioned such that  $p_a$  is directly between the fingers and  $p_c$  is at a desired point of contact on the finger,  $C = p_a - \delta a_{ac}$ . Here,  $\delta = G_d + (p_a - p_c)^T a_{ac}$  denotes the distance between the center of the gripper and  $p_a$  and  $G_d$  denotes gripper depth. We will sometimes refer to a grasp defined this way as an *edge grasp*. To sample edge grasps, we will generally sample the approach point  $p_a$  first and then for each approach point sample multiple contact points  $p_c$  from the neighbors of  $p_a$  within the distance of  $\frac{G_w}{2}$ , where  $G_w$  denotes the aperture of the gripper, i.e. the distance between the fingers when the gripper is open.

### B. Model Architecture

Our model, which we call the *Edge Grasp Network*, evaluates the grasp quality for a set of edge grasps that have a single approach point  $p_a \in P$  in common. We evaluate multiple approach points by cropping separately and then placing them in a batch. There are four steps, as illustrated in Figure 2.

**Step 1: Crop Point Cloud.** Given a point cloud  $P$  and an approach point  $p_a$ , only a set of neighboring points of  $p_a$  affects the edge grasp. We crop the point cloud to a ball around  $p_a$ :

$$S_a = \{p \in P : \|p - p_a\|_2 \leq G_w/2\},$$

**Step 2: PointNetConv ( $\psi$ ).** We compute a feature at each point using a stack of PointNetConv layers [17], denoted  $\psi$ . Each layer calculates a new feature  $f_i^{(l+1)}$  at each point

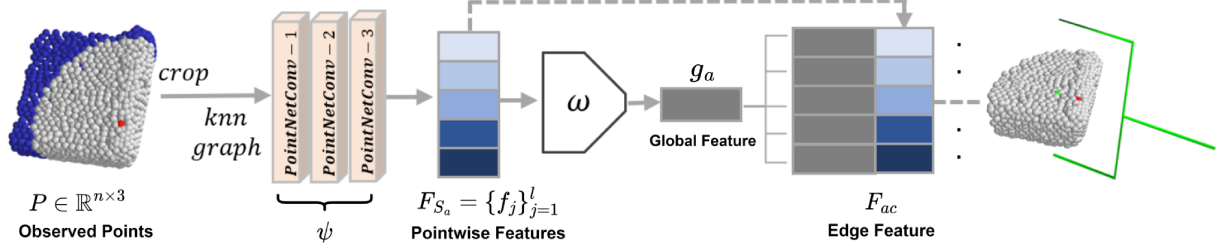


Fig. 2. Encoding process of edge grasps. After sampling  $p_a$  colored in the red from the partial point cloud, we first crop the neighboring points  $S_a$  colored in grey around  $p_a$ . Then,  $S_a$  is fed to a graph neural network  $\psi$  to generate pointwise feature  $F_{S_a}$ . The global feature  $g_a$  is extracted with  $\omega$  from  $F_{S_a}$  and the edge feature is the concatenation of the point feature and the shared global feature. The rightmost part shows the represented grasp of one edge feature.

$p_i \in S_a$  using

$$f_i^{(\ell+1)} = \max_{j \in \mathcal{N}(i)} \text{MLP} \left( f_j^{(\ell)}, p_j - p_i \right), \quad (2)$$

where  $\mathcal{N}(i)$  denotes the  $k$ -nearest neighbors to  $p_i$ . Here,  $f_j^{(\ell)}$  denotes the feature at point  $p_j$  prior to the layer,  $\max$  denotes max-pooling where the max is taken over features (like in PointNet [22]). MLP is a 2-layer multi-layer perceptron that takes both parameters as input. The input features at the first layer are the positions and surface normals of the points. Notice that this is slightly different from a standard PointNet++ model. Whereas PointNet++ has PointNetConv layers with successively coarser resolution, our model applies multiple successive layers over the same graph – without successive abstraction. Let  $F_{S_a}$  denote the set of features for the points in  $S_a$  at the output of Step 2.

**Step 3: Compute Global Feature ( $\omega$ ).**  $\omega$  takes  $F_{S_a}$  as input and generates a single global feature  $g_a$  that describes  $S_a$ . First,  $F_{S_a}$  is passed to an MLP followed by a max-pooling layer (over features) to generate a first-level global feature. This is concatenated with each feature  $f \in F_{S_a}$  and passed to a second MLP and max-pooling layer to output  $g_a$ . Finally, for each edge grasp  $(p_a, p_c) \in P^2$  associated with  $p_a$ , we calculate an edge feature  $f_{ac} \in F_{ac}$  by concatenating  $g_a$  with the point feature  $f_c \in F_{S_a}$  corresponding to  $p_c$ . This edge feature will represent the edge grasp to the classifier (next step). The combination of local features with global features has shown improvements in 3D geometric learning [23], [24], [25]. Similarly, the edge representation allows our model to have a good awareness of the local context of  $p_c$  and the global information of  $S_a$ .

**Step 4: Grasp Classification.** After calculating the edge features  $F_{ac}$ , we are now ready to evaluate grasp quality. We predict grasp success using a four-layer MLP with a sigmoid function which takes an edge feature  $f_{ac}$  as input and infers whether the corresponding edge grasp will succeed.

### C. SO(3) Invariance of Edge Grasp Network

In Section III, we noted that the grasp quality function  $\Phi(P, \alpha)$  is invariant to translation and rotation, i.e.  $\Phi(g \cdot P, g \cdot \alpha) = \Phi(P, \alpha)$  for arbitrary  $g \in \text{SE}(3)$ . As presented above, the Edge Grasp Network is invariant to translation because each  $S_a$  is centered at the approach point  $p_a$  (we translate  $p_a$  to the origin of the world frame). However, additional methodology is required to create invariance to

rotations. Rotational invariance allows the model to generalize grasp knowledge from one orientation to another. We enable rotational invariance with two different approaches. The first approach is to apply data augmentation on  $S_a$  to learn  $\text{SO}(3)$  invariance during training. Our second approach is to use an  $\text{SO}(3)$ -equivariant model, Vector Neurons [26]. Vector Neurons can be applied to nearly any neural model architecture by encoding the  $\mathbb{R}^3$  along which  $\text{SO}(3)$  acts as a separate tensor axis. In the Edge Grasp Network, for example, the initial feature for a single point  $p_i \in \mathbb{R}^3$  with its normal  $n_i \in \mathbb{R}^3$  is regarded as a two-channel feature  $f_i^{\ell=0} \in \mathbb{R}^{2 \times 3}$ . As a result, it is *equivariant* with a linear neural network layer. That is, for a linear layer  $f^{(\ell+1)} = W^\ell f^{(\ell)}$  with weights  $W^\ell$ , rotation of the input  $f^{(\ell)}$  by an arbitrary  $R \in \text{SO}(3)$  corresponds to a rotation of the output, i.e.  $f^{(\ell+1)} R = W^\ell (f^{(\ell)} R)$ . It turns out that with additional modifications to nonlinearities and max-pooling layers, Vector Neurons can transform any neural model into an  $\text{SO}(3)$  equivariant model. Since invariance is a special case of equivariance, we can now create an end-to-end invariant model. As we show in Section V-H, leveraging  $\text{SO}(3)$  symmetries is beneficial to learn a grasp function.

### D. Grasp Sampling

Edge Grasp Network enables us to evaluate a large number of edge grasps that share a single approach point with a single forward pass through the model. However, each different approach point necessitates evaluating the model separately. Therefore we adopt the following grasp sample strategy. First, we sample a small number of approach points  $P_a \subset P$ . These approach points can be sampled uniformly at random from the cloud, or they can be focused on parts of the cloud where a grasp is preferred. Then, we evaluate the model once for all approach points by forming a minibatch of  $|P_a|$  inputs and performing a single forward pass. The output of this is a set of sets of edge grasp features,  $F_{(ac)_1}, F_{(ac)_2}, \dots, F_{(ac)_{|P_a|}}$ . One can take the union of these sets, sample  $m$  edge grasps uniformly at random or select grasps with preferred gripper approach directions and gripper contact locations, and then run the grasp classifier on these sampled grasps to produce the final output.

## V. SIMULATIONS

We benchmarked our method in simulation against three strong baselines, PointNetGPD [13], VGN [11], and GIGA [9]. To make the comparison as fair as possible, we

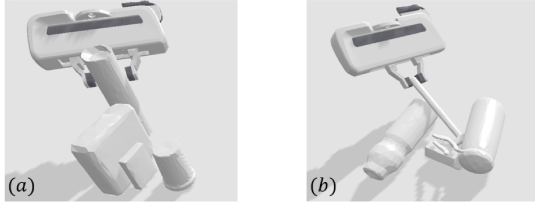


Fig. 3. Left: the packed scenario; Right: the pile scenario.

used the same simulator developed by Breyer et al. [11] (the authors of VGN) and used by Jiang et al. [9] (the authors of GIGA). The results indicate that our method outperforms all three methods by significant margins.

#### A. Simulation Environment

The grasp simulator developed by Breyer et al. [11] is implemented in PyBullet [27] and includes a Franka-Emika Panda gripper. There are 303 training objects and 40 test objects drawn collectively from YCB [28], BigBird [29] and other sources [30], [31]. There are two types of simulated grasp environments, PACKED and PILED. In PACKED, objects are placed randomly in an upright configuration in close proximity, e.g. as shown in Figure 3(a). In PILED, objects are dumped randomly from a box into a pile, e.g. as shown in Figure 3(b).

#### B. Experimental Protocol

We evaluate our model over several rounds of testing. During each round, a pile or packed scene with 5 test objects is generated inside of a  $30 \times 30 \times 30 \text{ cm}^3$  workspace and the system begins grasping one object at a time. Prior to each grasp, we take a depth image of the scene from a direction<sup>1</sup> above the table. We add pixelwise Gaussian noise ( $\mathcal{N} \sim (0, 0.001)$ ) to the depth image, extract the point cloud or TSDF (Truncated Signed Distance Function) from the depth image, and pass it to the model. After receiving grasp scores from the model, we execute the grasp with the highest quality score. If the grasp is successful, the object is removed from the workspace. A round of testing ends when either all objects are cleared or two consecutive grasp failures occur.

#### C. Model

We implemented the Edge Grasp Network model described in Section IV-B. The input to the model is a downsampled point cloud created by voxelizing the input with a 4mm voxel dimension. The PointNetConv layers in  $\psi$  are implemented using a KNN graph with  $k = 16$ , i.e. with 16 nearest neighbors.  $\psi$  is implemented as a sequence of three PointNetConv layers with a 2-layer MLP as the message passing function. The grasp classifier is implemented as a 4-layer MLP with ReLUs [32] and a sigmoid layer at the end. We evaluate both conventional and Vector Neuron versions of our model.

#### D. Training

We created training data by generating both packed and piled scenes with a random number of objects in simulation, adding Gaussian noise to the depth images captured from

random camera views, voxelizing the point cloud, generating up to 2000 edge grasp candidates per scene, and labeling each of those candidates by attempting a grasp in simulation. To generate the 2000 edge grasp candidates, we sample 32 approach points uniformly at random from the voxelized cloud. In total, we generated 3.36M labeled grasps based on 3,317 scenes, 85% of which were used for training and 15% were used for testing. We train our model with the Adam [33] optimizer and an initial learning rate of  $10^{-4}$ . The learning rate is reduced by a factor of 2 when the test loss has stopped improving for 6 epochs. It takes about 0.5 seconds to complete one SGD step with a batch size of 32 on a NVIDIA Tesla V100 SXM2 GPU. We train the model for 150 epochs and balance the positive and negative grasp labels during training. Both VN-EdgeGraspNet and EdgeGraspNet converge in less than 10 hours.

#### E. Data Augmentation

Extensive data augmentation is applied to the conventional version of our model to force it to learn the  $SO(3)$  invariance from training. Before loading the point cloud  $P$  from the training dataset, we randomly sample a  $g \in SO(3)$  to rotate  $P$ . This results in rotations on the 32 cropped point clouds corresponding to each approach point, i.e.,  $\{g \cdot S_{a_1}, g \cdot S_{a_2}, \dots, g \cdot S_{a_{32}}\}$ . Since  $S_a$  is centered at  $p_a$ , we then translate  $p_a$  to the origin. A batch of 32 rotated and translated  $S_a$  is fed to our model as the input during training. Since the Vector Neurons version of our model obtains  $SO(3)$  invariance by mathematical constraint, in this case only a translation is applied to each  $S_a$ .

#### F. Baselines

We compare our method against three strong baselines. *PointNetGPD* [13] is a sample-based method that represents a candidate grasp pose by the canonicalized points inside the gripper and infers grasp quality using a PointNet [22] model. *VGN* [11] (Volumetric Grasping Network) takes a TSDF of the workspace as input and outputs the grasp orientation and quality at each voxel. *GIGA* [9] (Grasp detection via Implicit Geometry and Affordance) uses a structured implicit neural representation from 2D feature grids and generates the grasp orientation and quality for each point trained with a auxiliary occupancy loss. Both VGN and GIGA receive a  $40 \times 40 \times 40$  TSDF based on output from a single depth image. We also evaluate a variation of GIGA with a  $60 \times 60 \times 60$  resolution TSDF, which we refer to as *GIGA-High*. We use the pretrained models<sup>2</sup> of VGN and GIGA from Jiang et al. [9] and uniformly sample 64 approach points and 4000 grasps for our method and *PointNetGPD*. As shown in Table II, the pretrained VGN and GIGA models have fewer parameters than our method due to their TSDF input. While our model requires more parameters to operate on point clouds, all compared models are relatively lightweight.

<sup>1</sup>All methods are tested using the particular viewpoint direction reflected in the training data of Jiang et al. [9].

<sup>2</sup>Our trained models for VGN and GIGA on the dataset described above in Section V-D did not perform as well as the pretrained models from Jiang et al. [9]. It is probably because they train separate models for the PACKED and PILE scenarios with a larger dataset (4M labeled grasps for each scenario). We used their pretrained models to do the evaluations.



TABLE I. Quantitative results of clutter removal. We report mean and standard deviation of grasp success rates (GSR) and declutter rates (DR). Edge-sample randomly sample edges that do not collide with the table. GIGA-High query at a higher resolution of  $60 \times 60 \times 60$ .

| Method          | Packed                           |                                  | Pile                             |                                  |
|-----------------|----------------------------------|----------------------------------|----------------------------------|----------------------------------|
|                 | GSR (%)                          | DR (%)                           | GSR (%)                          | DR (%)                           |
| PointNetGPD     | 79.3 $\pm$ 1.8                   | 82.5 $\pm$ 2.9                   | 75.6 $\pm$ 2.3                   | 77.0 $\pm$ 2.8                   |
| VGN             | 80.2 $\pm$ 1.6                   | 86.2 $\pm$ 2.0                   | 64.9 $\pm$ 2.2                   | 69.1 $\pm$ 3.2                   |
| GIGA            | 85.3 $\pm$ 1.9                   | 91.2 $\pm$ 1.7                   | 69.9 $\pm$ 1.8                   | 75.2 $\pm$ 2.2                   |
| GIGA-High       | 88.5 $\pm$ 2.0                   | 93.9 $\pm$ 1.4                   | 74.1 $\pm$ 1.5                   | 80.1 $\pm$ 0.5                   |
| Edge-Sample     | 44.0 $\pm$ 4.0                   | 39.7 $\pm$ 4.5                   | 40.2 $\pm$ 2.5                   | 30.9 $\pm$ 3.2                   |
| EdgeGraspNet    | 92.0 $\pm$ 1.4                   | 94.8 $\pm$ 0.8                   | 89.9 $\pm$ 1.8                   | 92.8 $\pm$ 1.6                   |
| VN-EdgeGraspNet | <b>92.3 <math>\pm</math> 1.2</b> | <b>95.2 <math>\pm</math> 0.6</b> | <b>92.3 <math>\pm</math> 1.5</b> | <b>93.5 <math>\pm</math> 1.8</b> |

| Method          | PointNetGPD | VGN   | GIGA  | GIGA-High | EdgeGraspNet | VN-EdgeGraspNet |
|-----------------|-------------|-------|-------|-----------|--------------|-----------------|
| # of Parameters | 1.6 M       | 0.3 M | 0.6 M | 0.6 M     | 3.0 M        | 1.7 M           |
| Inference time  | 382 ms      | 10 ms | 21 ms | 50 ms     | 28 ms        | 89 ms           |

TABLE II. Number of parameters and inference time for proposed methods and baselines. Evaluated on one NVIDIA-GeForce RTX 3090.

### G. Results

We compare the three baselines above with three variations of our own method. First, we compare against *Edge-Sample* which denotes the version of our method where we disregard the grasp quality inferred by the model and select a sampled edge grasp uniformly at random. Second, we compare against *EdgeGraspNet*, which is the version of our method trained with data augmentation. Finally, we compare with *VN-EdgeGraspNet* which is the version with Vector Neurons. We report the results in Table I. Performance is measured in terms of: 1) Grasp Success Rate ( $GSR = \frac{\# \text{successful grasps}}{\# \text{total grasps}}$ ) that measures the ratio of successful grasps to total grasps; and 2) Declutter Rate ( $DR = \frac{\# \text{grasped objects}}{\# \text{total objects}}$ ) that measures the ratio of objects removed successfully to the number of total objects presented. In all cases, performance is averaged over 100 simulation rounds with 5 different random seeds.

We draw several conclusions from Table I. First, our sample strategy unadorned with grasp quality inference (Edge-Sample) already performs surprisingly well with a grasp success rate of between 40% and 44%. This suggests our edge grasp representation and sample strategy provides a helpful bias. Second, both EdgeGraspNet and VN-EdgeGraspNet outperform all the baselines in all performance categories by a significant margin, particularly in the PILE category. Finally, the performance gap between the packed and piled scenarios is smaller for our method than that for the baselines, which suggests that our model adapts to different object configurations better.

### H. Performance Considerations

**Inference Time:** Table II shows the time needed by various models to infer grasp qualities. At 28ms per 4,000 grasps, our EdgeGraspNet model is slightly slower than both VGN and GIGA but still much faster than PointNetGPD and GIGA-High. The Vector Neurons version of our model is about three times slower than the EdgeGraspNet model.

**Performance of different sample sizes:** The speed and performance of our model is closely tied to the number of approach points (which determines batch size) and the number of classified grasps. Table III shows that fewer approach

| Method                  | Packed         |                | Pile           |                |
|-------------------------|----------------|----------------|----------------|----------------|
|                         | GSR (%)        | DR (%)         | GSR (%)        | DR (%)         |
| EdgeGraspNet (16-1k)    | 88.5 $\pm$ 1.7 | 92.6 $\pm$ 1.4 | 84.8 $\pm$ 2.1 | 86.7 $\pm$ 3.3 |
| EdgeGraspNet (32-2k)    | 91.4 $\pm$ 1.5 | 94.0 $\pm$ 2.0 | 89.4 $\pm$ 1.3 | 91.2 $\pm$ 2.5 |
| EdgeGraspNet (64-4k)    | 92.0 $\pm$ 1.4 | 94.8 $\pm$ 0.8 | 89.9 $\pm$ 1.8 | 92.8 $\pm$ 1.6 |
| VN-EdgeGraspNet (16-1k) | 89.7 $\pm$ 2.4 | 92.2 $\pm$ 1.6 | 87.1 $\pm$ 0.8 | 88.5 $\pm$ 2.3 |
| VN-EdgeGraspNet (32-2k) | 91.4 $\pm$ 1.3 | 93.8 $\pm$ 2.0 | 89.3 $\pm$ 0.5 | 92.1 $\pm$ 1.8 |
| VN-EdgeGraspNet (64-4k) | 92.3 $\pm$ 1.2 | 95.2 $\pm$ 0.6 | 92.3 $\pm$ 1.5 | 93.5 $\pm$ 1.8 |

TABLE III. Grasp performance for different numbers of approach points (16, 32, and 64) and grasp samples (1000, 2000, and 4000).

points and grasp samples reduce grasp success somewhat, but not by a huge amount.

### Vector Neurons and Data Augmentation:

To investigate the role of  $SO(3)$  invariance, we compared our base version of EdgeGraspNet with a variation that omits data augmentation (EdgeGraspNet-NoAug) and VN-EdgeGraspNet. As shown in Figure 4, the Vector Neurons version performs best and learns fastest and the base EdgeGraspNet converges to approximately the same level. However, without either Vector Neurons or data augmentation, the model overfits. This demonstrates that leveraging  $SO(3)$  symmetry is beneficial to learning the grasp function.

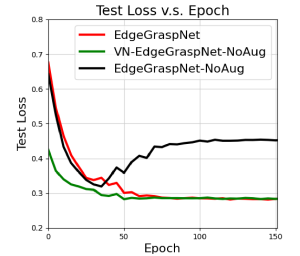


Fig. 4. Test loss functions showing the effect of data augmentation and Vector Neurons.

## VI. EVALUATION ON A ROBOT

In this paper, we measure physical grasp performance in three different setups, two of which are directly comparable to at least one other method from the literature.

### A. Setup

We used a UR5 robot equipped with a Robotiq-85 Gripper, as shown in Figure 6. An Occipital Structure Sensor was mounted on the arm to capture the observation. Prior to each grasp, we move the sensor to a randomly selected viewpoint<sup>3</sup> (pointing toward the objects to be grasped, as shown in Figure 6(a)), take a depth image, and generate a point cloud. We detect and remove the table plane with RANSAC and we denoise and downsample the point cloud using Open3D [34]. For each observed point cloud, we sample 40 approach points and 2000 grasps total. After running inference, we filter out the grasps with a grasp quality score below 0.9. As is the procedure in [11] and [1], we select the highest (largest  $z$ -coordinate) above-threshold candidate for execution.

### B. Household Objects in the Packed and Pile Settings

This experiment evaluates our method in the *packed* and *piled* settings described in Section V-A. In the packed environment, the robot grasps objects that are placed upright as shown in the lower part of Figure 5(a). In the piled environment, objects are piled as shown in Figure 5(b). In each round, 5 objects are randomly selected from 10 objects.

<sup>3</sup>We randomly select a viewpoint and repeatedly use it.

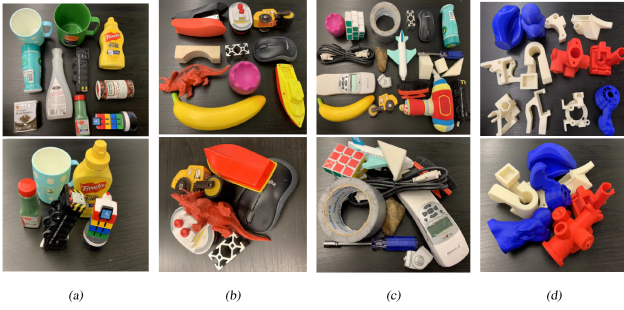


Fig. 5. Object sets and test configurations used for real robot experiments. From left column to right column: packed scene with 10 objects; pile scene with 10 objects; 20 test hard objects [12]; 12 Berkeley adversarial objects [35].

| Method          | Packed       |              | Pile         |              |
|-----------------|--------------|--------------|--------------|--------------|
|                 | GSR (%)      | DR (%)       | GSR (%)      | DR (%)       |
| EdgeGraspNet    | 91.9 (80/87) | 100 (80/80)  | 93.0 (80/86) | 100 (80/80)  |
| VN-EdgeGraspNet | 91.7 (78/85) | 98.7 (79/80) | 92.9 (79/85) | 98.7 (79/80) |

TABLE IV. Results of real-robot experiments for packed and piled grasp settings. The packed and pile objects are from Figure 5(a) and (b). Experiments are performed using the robot setup shown in Figure 6(a).

We repeatedly plan and execute grasps till all objects inside the workspace are cleared. Table IV reports grasp success rates (GSR) and declutter rates (DR) from 16 rounds (80 objects total) on the packed and pile scenes, respectively. GSRs vary between 91.7% and 93% – a result that closely matches our simulated results shown in Table I. Notice that our method performs similarly well in both packed and piled settings which is important because the packed setting favors side grasps while the piled setting favors top grasps. Qualitatively, most of our failures seem to be caused by collisions with other objects during grasping.

### C. Comparison with Zhu et al. [12] on a Pile of Household Objects

| Method          | GSR (%)        | DR (%)         |
|-----------------|----------------|----------------|
| Zhu et al. [12] | 89.0 (138/155) | 94.0 (141/150) |
| EdgeGraspNet    | 91.8 (146/159) | 98.0 (147/150) |
| VN-EdgeGraspNet | 93.6 (148/159) | 98.6 (148/150) |

TABLE V. Comparison with the method of Zhu et al. [12] using exactly the same objects and setup – a piled setting where 10 objects are selected from the 20 shown in Figure 5 (c). (Sometimes, two objects were grasped together in one trial.)

This experiment compares our method against the method of Zhu et al. [12], a strong baseline from the literature. We use exactly the same objects and robotic setup as in that paper. The object set is shown in Figure 5 (c) – these are the 20 “hard” objects from [12]. In each round, 10 objects are randomly selected and dumped on the table randomly. We parameterize the method of Zhu et al [12] using a model made available online by the authors [12]. Table V shows the results from 15 runs of each method. VN-EdgeGraspNet outperforms [12] by about four percentage points both in terms of the grasp success rate and the declutter rate – a significant improvement against a strong baseline.

### D. Comparison with [10] on the Berkeley Adversarial Pile

We also baselined our method using the 12 Berkeley Adversarial Objects described in [35]. These objects, shown

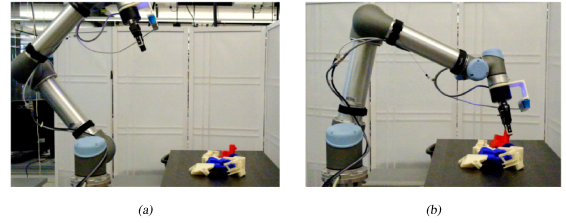


Fig. 6. Robot setup. Left: the robot takes a depth image of the scene from a random view point. Right: the robot grasps the red adversarial object from a localized graspable part.

| Method                | GSR (%)       | DR (%)       |
|-----------------------|---------------|--------------|
| Gualtieri et al. [1]* | 70.91 (39/55) | 97.5 (39/40) |
| Breyer et al. [11]*   | 41.56 (32/77) | 80 (32/40)   |
| Cai et al. [10]*      | 78.4 (40/51)  | 100 (40/40)  |
| EdgeGraspNet          | 84.4 (38/45)  | 95.0 (38/40) |
| VN-EdgeGraspNet       | 83.0 (40/48)  | 100 (40/40)  |

TABLE VI. Comparison with VPN [10], GPD [1], and VGN [11] for the Berkeley Adversarial Objects in a pile setting. We performed five rounds of grasping with piles of eight objects in each. \* Results for VPN [10], GPD [1], and VGN [11] are copied directly from [10].

in Figure 5 (d), have complex shapes that make them challenging to grasp. Here, we compare our method to the work of Cai et al. [10], called *Volumetric Point Network* (VPN). VPN is a very recent 6-DoF grasp pose detection method that predicts the point-wise grasp qualities by mapping 3D features from TSDF to points. To ensure our results are comparable to the VPN results reported by [10], we used the same experimental procedure that was used by those authors – five rounds of grasp testing, each with eight adversarial objects deposited randomly on a tabletop, e.g. as shown in Figure 6. For comparison, we report the result obtained by Cai et al. for their own method, as well as the results obtained by Cai et al. for GPD [1] and VGN [11]. Table VI shows the performance comparison. The results indicate that our method outperforms all the baselines. The most competitive baseline is VPN [10], which our grasp success rate outperforms by between four and six percentage points. Our final grasp success rate is 84.4%, a very good performance for the Berkeley adversarial object set. It indicates our model has a better understanding of graspable features of complex geometries. Videos of clutter removal of different object sets can be found in the supplementary videos.

## VII. CONCLUSION

This paper proposes a novel edge representation in 6-DoF grasp detection problem. By formulating the grasp pose with an approach point, a contact point and its surface normal, we represent edge grasps by local features of contacts and global features of the related points. We explore the SE(3) symmetry of our representation and propose EdgeGraspNet and VN-EdgeGraspNet to leverage SE(3) invariance in two different ways. Finally, We evaluate our models on various simulated and real-world object sets against several strong baselines. Experiments shows the small sim-to-real gap, the high grasping success rate, and the generalization ability to different object sets of our method. A clear direction for future work is to integrate more on-policy learning, which we believe would enable us to improve the performance.

## REFERENCES

- [1] M. Gualtieri, A. Ten Pas, K. Saenko, and R. Platt, "High precision grasp pose detection in dense clutter," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2016, pp. 598–605.
- [2] A. ten Pas, M. Gualtieri, K. Saenko, and R. Platt, "Grasp pose detection in point clouds," *The International Journal of Robotics Research*, vol. 36, no. 13-14, pp. 1455–1473, 2017.
- [3] A. Mousavian, C. Eppner, and D. Fox, "6-dof grasnet: Variational grasp generation for object manipulation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 2901–2910.
- [4] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.
- [5] 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," *arXiv preprint arXiv:1703.09312*, 2017.
- [6] D. Morrison, P. Corke, and J. Leitner, "Closing the loop for robotic grasping: A real-time, generative grasp synthesis approach," *arXiv preprint arXiv:1804.05172*, 2018.
- [7] S. Kumra, S. Joshi, and F. Sahin, "Antipodal robotic grasping using generative residual convolutional neural network," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9626–9633.
- [8] X. Zhu, D. Wang, O. Biza, G. Su, R. Walters, and R. Platt, "Sample efficient grasp learning using equivariant models," *Proceedings of Robotics: Science and Systems (RSS)*, 2022.
- [9] Z. Jiang, Y. Zhu, M. Svetlik, K. Fang, and Y. Zhu, "Synergies between affordance and geometry: 6-dof grasp detection via implicit representations," *arXiv preprint arXiv:2104.01542*, 2021.
- [10] J. Cai, J. Cen, H. Wang, and M. Y. Wang, "Real-time collision-free grasp pose detection with geometry-aware refinement using high-resolution volume," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1888–1895, 2022.
- [11] M. Breyer, J. J. Chung, L. Ott, R. Siegwart, and J. Nieto, "Volumetric grasping network: Real-time 6 dof grasp detection in clutter," *arXiv preprint arXiv:2101.01132*, 2021.
- [12] X. Zhu, D. Wang, O. Biza, G. Su, R. Walters, and R. Platt, "Sample efficient grasp learning using equivariant models," *arXiv preprint arXiv:2202.09468*, 2022.
- [13] H. Liang, X. Ma, S. Li, M. Görner, S. Tang, B. Fang, F. Sun, and J. Zhang, "Pointnetgpd: Detecting grasp configurations from point sets," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3629–3635.
- [14] Y. Qin, R. Chen, H. Zhu, M. Song, J. Xu, and H. Su, "S4g: Amodal single-view single-shot se (3) grasp detection in cluttered scenes," in *Conference on robot learning*. PMLR, 2020, pp. 53–65.
- [15] C. Wu, J. Chen, Q. Cao, J. Zhang, Y. Tai, L. Sun, and K. Jia, "Grasp proposal networks: An end-to-end solution for visual learning of robotic grasps," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 174–13 184, 2020.
- [16] B. Zhao, H. Zhang, X. Lan, H. Wang, Z. Tian, and N. Zheng, "Regnet: Region-based grasp network for end-to-end grasp detection in point clouds," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 474–13 480.
- [17] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.
- [18] A. Bicchi, "On the closure properties of robotic grasping," *The International Journal of Robotics Research*, vol. 14, no. 4, pp. 319–334, 1995.
- [19] D. Wang, R. Walters, X. Zhu, and R. Platt, "Equivariant  $q$  learning in spatial action spaces," in *Conference on Robot Learning*. PMLR, 2022, pp. 1713–1723.
- [20] H. Huang, D. Wang, R. Walter, and R. Platt, "Equivariant transporter network," *arXiv preprint arXiv:2202.09400*, 2022.
- [21] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, and V. Sitzmann, "Neural descriptor fields: Se (3)-equivariant object representations for manipulation," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 6394–6400.
- [22] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 652–660.
- [23] Y. Zhang and M. Rabbat, "A graph-cnn for 3d point cloud classification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6279–6283.
- [24] H. Huang, Z. Yang, and R. Platt, "Gascn: Graph attention shape completion network," in *2021 International Conference on 3D Vision (3DV)*. IEEE, 2021, pp. 1269–1278.
- [25] S. Huang, Z. Gojcic, M. Usvyatsov, A. Wieser, and K. Schindler, "Predator: Registration of 3d point clouds with low overlap," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2021, pp. 4267–4276.
- [26] C. Deng, O. Litany, Y. Duan, A. Poulenard, A. Tagliasacchi, and L. J. Guibas, "Vector neurons: A general framework for so (3)-equivariant networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 200–12 209.
- [27] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," 2016.
- [28] B. Calli, A. Singh, A. Walsman, S. Srinivasa, P. Abbeel, and A. M. Dollar, "The ycb object and model set: Towards common benchmarks for manipulation research," in *2015 international conference on advanced robotics (ICAR)*. IEEE, 2015, pp. 510–517.
- [29] A. Singh, J. Sha, K. S. Narayan, T. Achim, and P. Abbeel, "Bigbird: A large-scale 3d database of object instances," in *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2014, pp. 509–516.
- [30] A. Kasper, Z. Xue, and R. Dillmann, "The kit object models database: An object model database for object recognition, localization and manipulation in service robotics," *The International Journal of Robotics Research*, vol. 31, no. 8, pp. 927–934, 2012.
- [31] D. Kappler, J. Bohg, and S. Schaal, "Leveraging big data for grasp planning," in *2015 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2015, pp. 4304–4311.
- [32] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.
- [33] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
- [34] Q.-Y. Zhou, J. Park, and V. Koltun, "Open3d: A modern library for 3d data processing," *arXiv preprint arXiv:1801.09847*, 2018.
- [35] J. Mahler, M. Matl, V. Satish, M. Danielczuk, B. DeRose, S. McKinley, and K. Goldberg, "Learning ambidextrous robot grasping policies," *Science Robotics*, vol. 4, no. 26, p. eaau4984, 2019.

## VIII. APPENDIX

### A. Simulator Setting

Here, we provide a more detailed description of our simulator settings. To generate the training data, we selected a random number of objects from training object sets. We set the mass of each object as 0.5 kg and the friction ratio between gripper and the object as 0.75. We label up to 2000 edge grasp candidates per scene by attempting a grasp in simulation. To sample 2000 grasps, we sample 32 approach points from the observed point clouds through Farthest Point Sampling. Edge grasps whose minimum  $z$  value is smaller than the height of the table are filtered out to avoid colliding with the table. A *True* label of a grasp candidate must satisfy the following conditions: 1) the gripper should not collide with any objects when moving from the “prergrasp” pose to the grasp pose; 2) the object must be hold by the gripper after a sequence of gripper shaking motions.

### B. $SO(3)$ Equivariance to $SO(3)$ Invariance

Based on Vector Neurons [26], we implement the equivariant PointNetConv to realize the  $SO(3)$  equivariant feature. We maintain the equivariance of our network until getting the edge feature  $f_{ac}$ . Invariance is a special case of equivariance and can be achieved by multiplying a matrix  $T_{ac} \in \mathbb{R}^{3 \times 3}$  generated from  $f_{ac}$  by a network:

$$(f_{ac}R)(T_{ac}R)^\top = f_{ac}RR^\top T_{ac}^\top = f_{ac}T_{ac}^\top \quad (3)$$

Equation 3 transforms the  $SO(3)$ -equivariant edge feature to  $SO(3)$ -invariant edge feature. Combined with the translational invariance described in Section IV-C, we finally realize the  $SE(3)$  invariance of edge features. Once the edge features are  $SE(3)$  invariant, the entire network becomes  $SE(3)$  invariant, i.e., the invariant feature could be fed to a conventional MLP without breaking its invariant property.

### C. The importance of cropping $S_a$

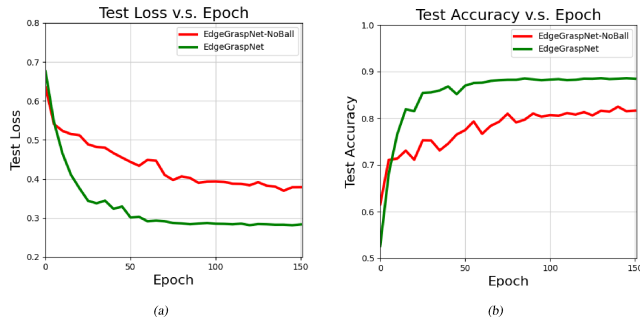


Fig. 7. Ablation Study on cropping  $S_a$ . Left Figure: Test loss v.s. Epoch; Right Figure: Test Accuracy v.s. Epoch. The results show the effect of cropping  $S_a$ .

We compare our EdgeGraspNet with a variation that skips cropping point cloud around the approach point  $p_a$ . After getting the observed point cloud  $P$ , we build a KNN graph on  $P$  and feed it to  $\psi$  directly to get the point features  $F_P$ . Then, we extract the global feature  $g_a$  corresponding to  $p_a$  from  $\{f_p \in F_P | p \in S_a\}$ . Instead of translating  $p_a$  to the origin of the world coordinate, we center  $P$  at the origin. Except these variations, other operations are the same as Section IV-B.

Let’s denote the variation as EdgeGraspNet-NoBall. Figure 7 shows the results of our model and the variation version. It indicates that implementing on  $S_a$  is better on implementing on  $P$ . There are some reasons why  $S_a$  is better than  $P$ . First,  $P$  is a special case of  $S_a$  when we set the radius of the sphere as infinity. Second,  $S_a$  includes all the related points that affect the grasp quality without redundant information. Last but not least, the invariant property on  $S_a$  is more generalized than that on  $P_a$ . Given a  $g \in SO(3)$ , a grasp action  $\alpha$ , and a grasp evaluation function  $\Psi$ , the invariance of EdgeGraspNet could be defined as

$$\Psi(g \cdot S_a, g \cdot \alpha) = \Psi(S_a, \alpha) \quad (4)$$

However, EdgeGraspNet-NoBall could only be invariant to rotations on the entire point cloud:  $\Psi(g \cdot P, g \cdot \alpha) = \Psi(P, \alpha)$ , which is less generalized.

### D. Inference Time

TABLE VII. Inference time v.s. # of approach points. We sample different numbers of approach points (16, 32 and 64) with the same number (2000) of edge grasps. Evaluated on one NVIDIA-GeForce RTX 3090.

|              | 16-2k  | 32-2k   | 64-2k   |
|--------------|--------|---------|---------|
| EdgeGraspNet | 9.6 ms | 15.8 ms | 27.4 ms |

|              | 32-500  | 32-1k   | 32-2k   |
|--------------|---------|---------|---------|
| EdgeGraspNet | 15.8 ms | 15.7 ms | 15.8 ms |

TABLE VIII. Inference time v.s. # sampled edge grasps. We sample different numbers of edge grasps (500, 1000 and 2000) with the same number (32) of approach points. Evaluated on one NVIDIA-GeForce RTX 3090.

The inference time of our method depends on the number of approach points. As shown in Table VII, when we double the number of approach points, the inference time increases about 1.7 times. As shown in Table IX, when we fix the number of approach point and increase the sampled edge grasps, the inference time almost does not change.

### E. Failure Case Analysis

| Method               | EdgeGraspNet   |                | VN-EdgeGraspNet |                |
|----------------------|----------------|----------------|-----------------|----------------|
|                      | GSR (%)        | DR (%)         | GSR (%)         | DR (%)         |
| Household Packed     | 91.9 (80/87)   | 100 (80/80)    | 91.7 (78/85)    | 98.7 (79/80)   |
| Household Pile       | 93.0 (80/86)   | 100 (80/80)    | 92.9 (79/85)    | 98.7 (79/80)   |
| Test Hard objects    | 91.8 (146/159) | 98.0 (147/150) | 93.6 (148/159)  | 98.6 (148/150) |
| Berkeley Adversarial | 84.4 (38/45)   | 95.0 (38/40)   | 83.0 (40/48)    | 100 (40/40)    |

TABLE IX. Summary of real Robot experiments. We report grasp success rates (GSR) and declutter rates (DR).

Almost half of our failures are caused by colliding with other objects when executing the grasp. It could be mitigated by considering collision when selecting grasps. However, there are some other cases we think readers might want to notice. 1). Occlusion due to partial observation, e.g., a single camera view could only capture a plane of a complex object. 2). Sensor noise. Our model is robust to small noises and leverage the bilateral symmetry of a parallel jaw gripper,



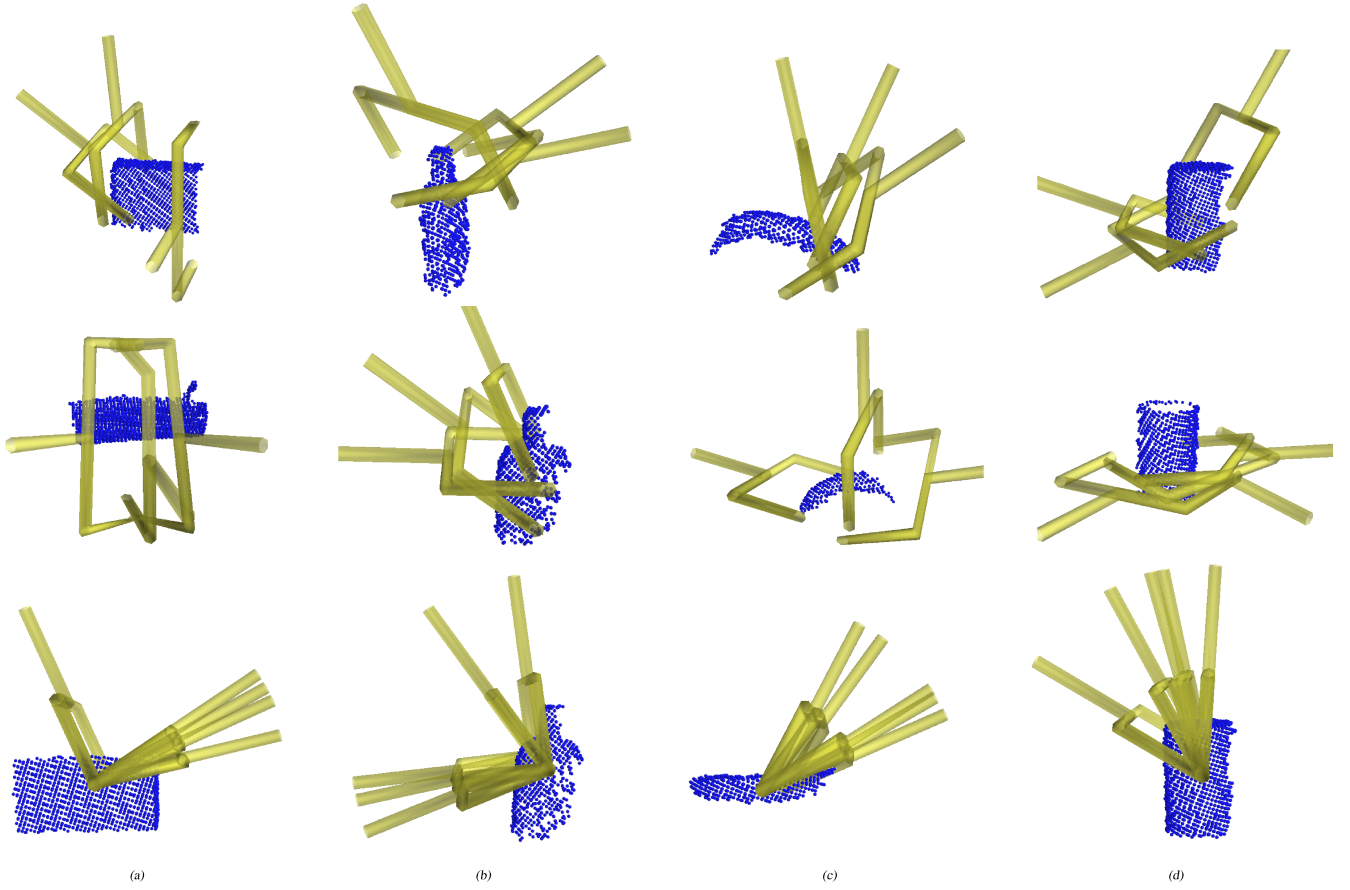


Fig. 8. Illustrations of grasp candidates found using our algorithm. The first two rows show three examples of a gripper placed at randomly sampled grasp candidate configurations. The last row shows five grasps that share the same contact point.

i.e., a flip of the calculated surface normal<sup>4</sup> results in a  $180^\circ$  rotation of the gripper along the approach direction. However, if the observation is largely distorted, the proposed edge grasp could be inaccurate since our sampling strategy is closely related to the observed points. There is a trade-off between the precise grasping and the robust grasping. 3). Grasp label of training data. Our binary label of the training data is described in Section VIII-A, but it does not prohibit *true dangerous* grasps. A dangerous grasp could be defined as there is a large change of the pose of the target object when being grasped regardless a successful outcome or not. We believe the true dangerous grasp could cause false-positive predictions when the observation is noisy. Last but not least, failures are the stepping stones to better algorithms in robotics.

#### F. Visualization of Grasps

We show grasp candidates found using our algorithm in Figure 8. The first two rows show three examples of randomly sampled grasp poses for each observed object. The diversity of grasp poses demonstrates our model can provide a high coverage of possible stable grasps. The last row of Figure 8 shows five grasps that share the same contact point. It indicates our model is beneficial to grasping tasks involved with specific contact locations.

<sup>4</sup>A flip of the calculated surface normal happens frequently.