2020 IEEE International Conference on Robotics and Automation (ICRA)
31 May - 31 August, 2020. Paris, France

Hierarchical 6-DoF Grasping with Approaching Direction Selection

Yunho Choi, Hogun Kee, Kyungjae Lee, JaeGoo Choy, Junhong Min, Sohee Lee, and Songhwai Oh

***Abstract*— In this paper, we tackle the problem of 6-DoF grasp detection which is crucial for robot grasping in cluttered real-world scenes. Unlike existing approaches which synthesize 6-DoF grasp data sets and train grasp quality networks with input grasp representations based on point clouds, we rather take a novel hierarchical approach which does not use any 6-DoF grasp data. We cast the 6-DoF grasp detection problem as a robot arm approaching direction selection problem using the existing 4-DoF grasp detection algorithm, by exploiting a fully convolutional grasp quality network for evaluating the quality of an approaching direction. To select the best approaching direction with the highest grasp quality, we propose an approaching direction selection method which leverages a geometry-based prior and a derivative-free optimization method. Specifically, we optimize the direction iteratively using the cross entropy method with initial samples of surface normal directions. Our algorithm efficiently finds diverse 6-DoF grasps by the novel way of evaluating and optimizing approaching directions. We validate that the proposed method outperforms other selection methods in scenarios with cluttered objects in a physics-based simulator. Finally, we show that our method outperforms the state-of-the-art grasp detection method in real-world experiments with robots.**

I. Introduction

Grasp detection is one of the most long-studied problems in robot manipulation due to its utility for various robotic applications from industry to service robots [1], [2]. In this problem, a robot observes a novel object and determines the position and orientation of its gripper to pick up the object. The robot relies only on the partial observation of the object, which makes the problem challenging. To make matters worse, the object and gripper's geometry, surface frictions, mass distribution, and kinematic feasibility affect the stability of grasps. Consequently, many approaches dealing with grasp detection simplify the problem by assuming a top-down grasp with a four-dimensional action space [3]–[8] , which consists of 3D positions and a rotation angle about the gripper axis. Especially, recent data-driven approaches [3], [4] which predict the best grasp pose from a depth image have shown successful generalization performance. However, these 4-DoF grasping methods do not generalize well for the arbitrary pose of an object and cluttered objects, and this fact makes them unsuitable for unstructured environments.

Y. Choi, K. Lee, J. Choy, H. Kee and S. Oh are with the Department of Electrical and Computer Engineering and ASRI, Seoul National University, Seoul 08826, Korea (e-mail: {yunho.choi, kyungjae.lee, jaegu.choy}@rllab.snu.ac.kr, {hogunkee, songhwai}@snu.ac.kr). J. Min and S. Lee are with Samsung Electronics Co., Suwon 16677, Korea (e-mail: {junhong1.min, ssohee.lee}@samsung.com). This work was supported in part by Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-01371, Development of Brain-Inspired AI with Human-Like Intelligence) and by a research grant from Samsung Electronics Co., Ltd (No. 0418-20190018).

Fig. 1: The grasp detection method proposed in this paper enables 6-DoF grasping by selecting the best approaching direction for 4-DoF grasping. The proposed method especially yields high grasp success rates for a dense clutter of objects as shown in this figure.

To overcome the limitation of 4-DoF grasping, it is natural to pursue a 6-DoF grasp which consists of the 3D position and 3D orientation of the gripper, giving the robot the maximum flexibility to select the best grasp. Planning a 6-DoF grasp is a much harder problem, since the search space for a 6-DoF grasp pose is larger than that of a 4-DoF grasp and a 6-DoF grasp depends more on the entire geometry of the object while we only have a partial observation from a single view even with possible occlusions. Most of recent methods which deal with 6-DoF grasp detection [9]–[11] propose their own grasp representations based on object point clouds. With the 6-DoF grasp representation, they synthesize a grasp data set to train a grasp quality network which are more demanding compared to the generation of 4-DoF grasp samples.

In order to take advantages of 6-DoF grasping by using only 4DoF grasping data and a simpler grasp detection model, we rather take a hierarchical approach which naturally extends the 4-DoF grasp detection problem into the 6-DoF grasp detection problem. First of all, we assume a depth sensor attached to a robot arm as shown in Figure 1. This is inspired by the fact that humans actively observe an object and manipulate the object with the best approaching direction. Then, we cast the 6-DoF grasp detection problem into the approaching direction selection problem for 4-DoF grasp detection, which is our first contribution. We propose a novel method for solving the grasp approaching direction selection problem which are comprised of three main parts: (1) how to define a good approaching direction; (2) how to generate approaching direction candidates; and (3) how to improve approaching direction candidates. To define a good approaching direction, we utilize a fully convolutional grasp quality network trained with 4-DoF grasp data [4]. We

assume that the 4-DoF grasp quality network has successfully learned the relation between a depth image and qualities of 4-DoF grasps within that image. Then, we evaluate the quality of an approaching direction with top-k grasp qualities predicted from the depth image seen from the chosen approaching direction, where we propose to estimate the depth image by transforming the point cloud, rather than moving the depth sensor every time.

The second and third parts are related with selecting the best approaching direction whose predicted top-k grasp qualities are maximal. Our another main contribution lies in the method we propose for generating good approaching direction candidates and iteratively improving the candidates, which leverages a geometry-based prior and an optimization method. Specifically, the proposed method uses the cross-entropy method to optimize the approaching direction quality, with initial seeds of surface normal directions. With the found best approaching direction, we move the robot arm to the viewpoint aligned with that direction and execute the best 4-DoF grasp found under that viewpoint. With the fully convolutional 4-DoF grasp quality network combined with the proposed approaching direction selection method, we efficiently search the space of 6-DoF grasps and enable the 6-DoF grasping without using any training data of 6-DoF grasps.

We evaluate our method in a physics-based simulator and compare the performance with other approaching direction selection methods. We show that the proposed method especially outperforms in densely cluttered scenes which are challenging for 4-DoF grasping methods. We also deploy our algorithm in real-world grasping scenarios using a robot, and validate that the proposed method takes advantage of 6-DoF grasping. To the best of our knowledge, the proposed method is the first approach solving the general 6-DoF grasp detection problem by replacing it with an approaching direction selection problem, which is more affordable and efficient.

II. Related Work

A. Learning Grasp Detection

The goal of grasp detection is to find a gripper configuration that maximizes the grasp quality metrics. The grasp quality metrics, such as grasp wrench space (GWS) analysis [12] and force-closure [13], physically analyze the geometry of the gripper and the object. Approaches for the grasp detection are either model-based or model-free. Model-based approaches [14]–[16] utilizes a data set of 3d object models labeled with feasible grasps, by matching the sensor inputs with the templates during execution. However, these approaches suffer from poor generalization on novel objects and thus we focus on model-free approaches in this paper.

Model-free methods rather solves the problem in data-driven approaches. Recent methods exploit convolutional neural network architectures to process raw RGBD inputs, and train a neural network with massive data sets of grasps, images, and grasp quality metric labels [3], [5]–[8]. The data sets are collected via either human-labeled [5], [8], self-supervised [6], [7], or synthetic ways [3]. They commonly represent the grasp with a simple 3-DoF oriented rectangle in the image [17] or 4-DoF representation where the grasp depth is added, facilitating the grasp data collection from depth images. However, these simplified rectangular representations of grasps can limit the grasp on arbitrary shapes, and impose a restriction on the workspace a robot.

Most of these model-free methods do not directly generate grasp poses, but rather consist of two cascaded parts, grasp candidate sampling and grasp quality evaluation. They first sample antipodal grasps based on geometry-based heuristics [18], and rank them with the grasp quality metrics predicted by the convolutional neural networks. Since we cannot evaluate all possible grasps, sampling plausible grasp candidates is crucial and often iterative optimization techniques, such as the cross entropy method [3], [19]–[22], are used. However, these techniques impose computational burden since the network must be iteratively queried for a batch of predictions.

Meanwhile, incorporating the advances in the computer vision for pixel-wise prediction using fully convolutional networks (FCNs) [23], recent approaches have enabled dense evaluation of the entire space of possible discretized grasp actions given a depth image. Approaches used by [24], [25] evaluate 3-DoF grasps and FC-GQ-CNN [4] evaluates 4-DoF grasps, both eliminating the need for sampling grasp candidates. Especially, FC-GQ-CNN will play a key role in our proposed method since it can evaluate all 4-DoF grasps within the depth image observed at a viewpoint and it shows the state-of-the-art performance in the 4-DoF grasping.

B. Learning 6-DoF Grasp Detection

To give a robot the maximum degree of freedom to select the grasp, we target the problem of 6-DoF grasp detection which is more challenging since the entire object geometry including the occluded parts affects the grasp quality. Analogous to the cascaded approach of 4-DoF grasp detection methods, most of the 6-DoF approaches have two parts of grasp candidates sampling and grasp quality evaluation. Yan et al. [21] solves the partial observability by training the auxiliary task of reconstructing the object geometry. whereas our method deals with the partial observability by leveraging the second sight at the viewpoint with the selected best approaching direction.

Recent methods, GPD [9] and PointNetGPD [10], sample grasp candidates using a geometry-based method in [26]. They sample a point in the observed point cloud , construct a Darboux frame which is a basis aligned with the estimated surface normal and local principal curvature, and find nearby feasible grasps with local search. The authors of GPD handcrafted several projection features on point clouds for the quality network input, while the authors of PointNetGPD exploited the PointNet architecture [27] to predict grasp qualities from the point cloud within the gripper closing area. 6-DoF grasp data collection for training the quality network of the GPD and PointNetGPD involves processing whole meshes of objects, which are significantly costly

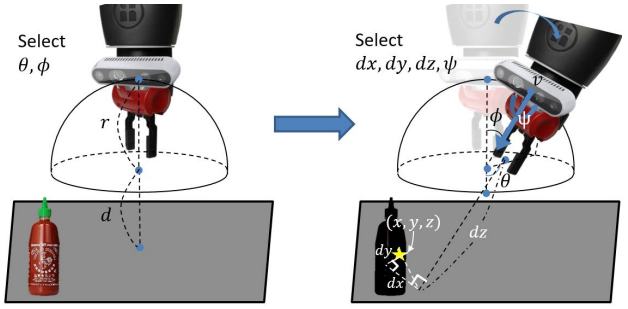


Fig. 2: An illustration of selecting an approaching direction $v = (\theta, \phi)$, which is defined with a point on a hemisphere above the table. The proposed method aligns the robot arm to the best approaching direction v^* and performs 4-DoF grasping under the new viewpoint.

compared to the 4-DoF grasp data collection [3]. Recent work [11] also uses the PointNet architecture for grasp evaluation , and samples diverse grasps from a generative neural network. However, this work is targeted for grasping singulated objects. The abovementioned approaches which learn from 3D point cloud data can suffer from overfitting and performance degradation when the input point cloud is sparse, while our method can efficiently sample grasp candidates in such cases.

Existing study on approaching direction selection for grasping [28] requires a predefined model, and focuses on a viewpoint selection problem near a target grasp to improve the grasp quality. Hence, it cannot be applied to general cases unlike the proposed method.

III. Problem Formulation

We assume a camera is attached to a robot arm, and initially the camera captures a depth map X from the top view. The depth map X is converted into a point cloud P. Given a set of object O, the object point cloud is denoted as P_O which is extracted from P by eliminating points associated with the table. We denote the 6-DoF grasp g by $(R, T) \in SE(3)$ where $R \in SO(3)$ and $T = (x, y, z) \in \mathbb{R}^3$ are the rotation and translation of the gripper in the initial gripper coordinate, respectively. We represent R with Euler angles $(\theta, \phi, \psi) \in \mathbb{R}^3$, where θ, ϕ, ψ implies azimuth, inclination, and rotation about gripper axis, respectively. We only consider parallel jaw grippers in this paper.

We assume a virtual hemisphere of a radius r, floating by d above the table and define a grasp approaching direction as a direction from a point on the hemisphere to the center of the hemisphere as seen in Figure 2. Thus, an approaching direction is uniquely determined by two angles, azimuth and inclination, which respectively correspond to θ and ϕ of R. We denote a grasp approaching direction by $v = (\theta, \phi) \in \mathbb{R}^2$ and a depth map captured from v is denoted by X_v. Q denotes a fully convolutional grasp quality network model which outputs the quality map $Q(X) \in \mathbb{R}^{W \times H \times K \times L}$ for a given depth map X, where W, H, K, and L represent the number of bins discretizing the 3D grasp position and rotation angle, respectively. Then the quality of an approaching direction v is defined as the average of k-largest elements in $Q(X_v)$, which is denoted by $Q_k^{\downarrow}(X_v)$, and can be computed by applying a partition algorithm to the flattened $Q(X_v)$. We consider top-k quality since a high-quality approaching direction must contain a large number of high-quality grasps. The most promising 4-DoF grasp at an approaching direction v is $\arg\max_{x,y,z,\psi} Q(X_v)$, and the quality of a grasp g predicted by Q given a depth map X is denoted by $Q(g|X)$.

Given a depth map X, our original problem of 6-DoF grasp detection is to find $g^* \in SE(3)$ which maximizes the grasp quality $Q_{6d}(g|X)$, where Q_{6d} is a nominal 6-DoF grasp quality evaluation model. For any possible grasp $g = (\theta, \phi, \psi, x, y, z)$, g can be viewed as a 4-DoF grasp from an approaching direction $v = (\theta, \phi)$. After aligning on the hemisphere point corresponding to v, if we let dx, dy, dz be a displacement from the corresponding hemisphere point of v to (x, y, z) in a rotated coordinate as shown in Figure 2, then executing a 4-DoF grasp $g' = (dx, dy, dz, 0, 0, \psi)$ in the rotated coordinate is equivalent with executing g. Therefore, the grasp quality $Q_{6d}(g|X)$ can also be evaluated by $Q(g'|X_v)$. In this paper, we cast the original problem of finding $g^* = (\theta^*, \phi^*, \psi^*, x^*, y^*, z^*)$ into the problem of selecting $v^* = (\theta^*, \phi^*)$ that maximizes the $Q_k^{\downarrow}(X_v)$ and finding a 4-DoF grasp $\arg\max_{g'} Q(g'|X_{v^*})$ at the selected direction, where $k = 1$ makes two problems identical, i.e., $\arg\max_g Q_{6d}(g|X)$ and $\arg\max_{g'} Q(g'|X_{v^*})$ where $v^* = \arg\max_v Q_1^{\downarrow}(X_v)$ are same in the world coordinate.

IV. Grasp Approaching Direction Selection Method

We focus on finding the best approaching direction $v^* = (\theta, \phi)$, assuming that a 4-DoF fully convolutional quality network model Q is given. Accordingly, we propose an efficient approaching direction selection method which generates candidates of good approaching directions with high approaching direction qualities $Q_k^{\downarrow}(X_v)$. The proposed method consists of two main parts, which are evaluation of an approaching direction quality and generation of approaching direction candidates. We iteratively generate better approaching direction candidates based on the evaluation of generated approaching direction candidates.

A. Evaluating Qualities of Grasp Approaching Directions

In order to evaluate the quality of an approaching direction v, i.e., $Q_k^{\downarrow}(X_v)$, the camera on the robot arm should be moved to the point on the hemisphere corresponding to v. Since moving the robot arm to every sampled approaching direction is exhausting, we rather approximate depth images seen at sampled approaching directions. Specifically, we obtain the point cloud P from the depth map X from the top view, transform P according to a camera movement for alignment with v, project the transformed point cloud, and interpolate it to fabricate a new depth image $\widehat{X}_v$. Algorithm 1 shows the overall process to estimate the qualities of approaching direction candidates. To evaluate multiple approaching directions at once, multiprocessing can be used when implementing Algorithm 1.

Using $k > 1$ makes it possible to leverage multi-modality of successful grasps in the process of the approaching direc-

tion selection. Since proposed grasps may violate kinematics and collision constraints, it is important to consider the diversity of high quality grasps and generate approaching direction candidates which lead to diverse choices of high quality grasps.

Algorithm 1 ApproachingDirectionQuality

Require: k for considering the k-largest grasp qualities, camera matrix C, 4-DoF fully convolutional grasp quality network Q

input: depth map X from the top view, approach direction v

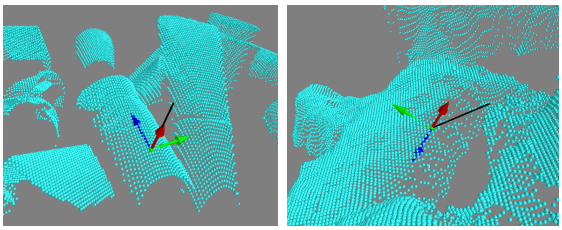
Convert X into the point cloud P
Compute the camera movement (R, T) from the top of the hemisphere to the corresponding point of v
Transform P with the inverse transform of (R, T)
Project the transformed point cloud into the image plane with camera matrix C
Interpolate the regular grid to make a 256×256 depth map $\widehat{X}_v$
Pass $\widehat{X}_v$ into Q
Compute the average of top-k predicted grasp qualities $Q_k^{\downarrow}(\widehat{X}_v)$

output: $Q_k^{\downarrow}(\widehat{X}_v)$

B. Generation of Approaching Direction Candidates

The core of our algorithm is iterative improvement of approaching direction candidates based on the result of the qualities evaluated with Algorithm 1. For the improvement of approaching direction candidates, we exploit the cross entropy method (CEM), which is a derivative-free optimization method with asymptotic convergence properties [19]. For our implementation, we iteratively fit a mixture of Gaussians to the elite set of sampled approaching directions and re-sample from the fitted mixture of Gaussians in order to find maxima of the approaching direction quality. With this simple adaptive procedure, approaching directions of high qualities are found. However, the biggest drawback of the CEM is that it is serial in nature. It repeatedly queries the quality network at every CEM iterations which puts a strain on time constraints for approaching direction selection.

In order to mitigate the computation time issue, we need a better starting point for our CEM iterations so that we can reduce the number of CEM iterations and the number of direction samples which are evaluated. Thus, we leverage the geometry-based prior for initial seeds, rather than starting from random samples of approaching directions. Specifically, we focus on the fact that surface normal vectors usually indicate good approaching directions for grasping, as many point cloud-based work [9], [10] sample grasp candidates using surface normal vectors. Empirically, grasps from the surface normal direction are expected to produce less collisions and insert the gripper deeper which is important for stably grasping round objects as seen in Figure 3-(a). Another important point is that the point cloud data from the real world are noisy that estimating the exact surface normal from the standard principal component analysis (PCA) over a small neighborhood is impossible, as shown in Figure 3-b. Therefore, we do not use the estimate of a surface normal using PCA. Rather, we use a smoothed estimate of the surface normal by considering surface normals of the



(a) (b)

Fig. 3: Visualization of orthogonal bases at surface points sampled from the point clouds of cluttered objects. The red axes, blue axes, green axes indicate surface normal directions, major principal curvature directions, and minor principal curvature directions, respectively. Black lines indicate the surface normal directions computed by the standard principal component analysis. (a) A case of point cloud captured in the MuJoCo simulator [30]. (b) A case of point cloud captured by the RealSense depth sensor in the real world.

Algorithm 2 SampleSmoothedNormals

Require: number of samples m, neighborhood radius ϵ, ϕ_{max}

input: depth map X from the top view

Convert X into the point cloud P and extract the object point cloud P_O
$\hat{n}(p_i)$ = SurfaceNormalEstimation(p_i) for all $p_i \in P_O$
for $j = 0$ to $m - 1$ **do**
  Sample $p_j \in P_O$ randomly uniformly
  Compute ϵ-ball about p_j: $B(p_j) = \{q \in P_O : ||p_j - q|| \leq \epsilon\}$
  $M(p_j) = \sum_{p \in B(p_j)} \hat{n}(p)\hat{n}(p)^T$
  Convert the eigenvector corresponding to the largest eigenvalue of $M(p_j)$ into the approaching direction $v_j = (\theta_j, \phi_j)$
  if $\phi_j > \phi_{max}$ **then** Reject; resample v_j
end for

output: $V = \{v_0, v_1, \cdots, v_{m-1}\}$

neighborhood as in [9], [10], [29]. Specifically, we perform PCA on the neighborhood surface normals without mean subtraction and extract the first principal component as a smoothed estimate of the surface normal. The entire process of sampling smoothed surface normals for initial candidates of the grasp approaching direction is specified in Algorithm 2. In our implementation, multiprocessing is used to generate samples in parallel when implementing Algorithm 2.

After sampling m smoothed surface normal directions for initial seeds, we iteratively improve our grasp approaching direction candidates by running the CEM. By starting from highly informative samples, the mixture of Gaussians in our CEM iterations reaches modes of an optimal region for the grasp approaching direction efficiently with a fewer number of iterations and samples. We name this CEM and surface normal-based approaching direction selection method as the GADS (Grasp Approaching Direction Selection) method and the entire process of GADS is detailed in Algorithm 3.

After selecting the best approaching direction by the GADS method, the robot arm moves to capture the second depth map at the chosen approaching direction, and plan the best 4-DoF grasp with a 4-DoF grasp quality network. Note that our method extensively examine 6-DoF grasps by sampling viewpoint candidates, rather than grasp candidates as in other 6-DoF approaches [9]–[11], and it can be generalized with any 4-DoF grasp quality model Q.

Algorithm 3 GADS: Grasp Approaching Direction Selection

Require: k for considering k-largest grasp qualities, number of CEM iteration r, number of Gaussian mixtures n, elite proportion p, number of initial samples m, number of CEM samples c

input: depth map X from the top view

Sample m initial seeds $\{v_i\}$ = SampleSmoothedNormals(X)
for $j = 0$ to $r - 1$ **do**
  $Q_k^{\downarrow}(\widehat{X}_{v_i})$ = ApproachingDirectionQuality(X, v_i) for all i
  Sort the approaching directions qualities $Q_k^{\downarrow}(\widehat{X}_{v_i})$ for all i
  Fit the mixture of n Gaussians to the p-elite set of the approaching directions
  Replace $\{v_i\}$ with c approaching directions sampled from the fitted mixture of Gaussians
end for
$Q_k^{\downarrow}(\widehat{X}_{v_i})$ = ApproachingDirectionQuality(X, v_i) for all i
Select the best approaching direction v^* among all i

output: $v^* = (\theta^*, \phi^*)$

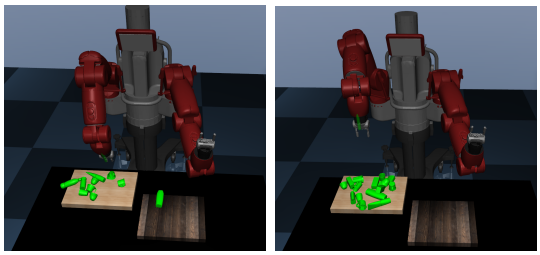


Fig. 4: Examples of cluttered objects scenes in simulation. (Left) Light clutter with 10 objects. (Right) Dense clutter with 20 objects.

V. Experiments

We evaluate the proposed approaching direction selection methods (GADS) both in simulations and robot experiments, for scenes with cluttered objects. We check the validity of GADS by comparing with other selection methods in simulation and get compelling results in the comparison with the state-of-the-art grasping algorithm for the task of picking cluttered objects in real-world experiments. All experiments are using a Baxter robot and the grasp detection algorithms ran on a notebook running Ubuntu 16.04 with a 2.6GHz Intel Core i7-9750H CPU and an NVIDIA GeForce RTX 2060.

A. Simulation for Comparison of Approaching Direction Selection Methods

For the simulation environment, we modified the simulation environment in Surreal suite [31], which is based on the MuJoCo physics engine [30]. To evaluate approach direction selection methods, we made 40 scenes for reproducing object types and positions, which are comprised of 20 scenes where 10 objects are lightly cluttered on a table and 20 scenes where 20 objects are densely cluttered. We aim to see if the proposed method takes advantage of 6-DoF grasping when objects are placed more densely. Objects are randomly chosen from an object set which consists of 15 household objects including a box, cup, and cylindrical shape. The objects are imported from meshes used in the DexNet dataset [3], and decomposed into convex parts using V-HACD algorithm [32] to simulate on MuJoCo.

The protocol of simulation is as follows. After spawning objects of each scene on a simulator, we move a robot arm to an initial position to capture a depth map from the top view, pick an object on the table with the detected grasp and place it on another table. Then we move the arm to the initial position and repeat the process. We try eight grasps for densely cluttered scenes, and six grasps for lightly cluttered scenes. If the robot picks an object up to $10cm$ above the table, then the grasp is declared as a successful grasp.

We compare six different methods, which includes 6-DoF grasping methods with five approaching direction selection methods and the state-of-the-art 4-DoF baseline method. The five selection methods are as follows.

1) GADS: The proposed method.
2) Random Selection: Approaching direction is sampled uniformly from $\theta \in [0, 2\pi], \phi \in [0, \phi_{max}]$.
3) CEM with Random Initial Seeds: GADS modified to start CEM with m uniformly sampled directions from $\theta \in [0, 2\pi], \phi \in [0, \phi_{max}]$.
4) Sampling Smoothed Normals: The best approaching direction is chosen from m samples in Algorithm 2.
5) Surface Normal Histogram: A simpler and faster workaround for sampling smoothed surface normals. This method converts entire $\{\hat{n}(p_i)\}$ in Algorithm 2 into the approaching directions $\{(\theta_i, \phi_i)\}$, and make a 2D histogram of $\{(\theta_i, \phi_i)\}$. In this way, we can find m dominating surface normal directions of the entire point cloud of objects without sampling and post-processing.

For a 4-DoF grasp quality model for all approaching direction selection methods, we use a pretrained FC-GQ-CNN model [4], which is also used for the 4-DoF baseline method. We perform nearest-neighbor interpolation in Algorithm 1 and compute the PCA surface normal estimate in Algorithm 2 with a point cloud library [33].

For the common hyperparameters of the selection methods, we use $k = 10$ to consider top-10 grasp qualities, and put a limit on ϕ by $\phi_{max} = \pi/7$, considering the kinematic feasibility of a Baxter robot. For CEM with Random Initial Seeds, we use $m = 24$, $c = 12$, $r = 2$, $n = 3$, and $p = 0.25$. For Sampling Smoothed Normals, we use $m = 12$, and $\epsilon = 1.5cm$. For Surface Normal Histogram, we use $m = 12$, and bin size of $(2/16\pi, 0.02\pi)$ for $(\theta \in [0, 2\pi], \phi \in [0, \phi_{max}])$. For GADS, we choose $m = 12$, $c = 8$, $r = 2$, $n = 2$, and $p = 0.5$, to see if GADS performs well in spite of using less samples than CEM with Random Initial Seeds.

Simulation results including success rates and average selection times are shown in Table I. Selection time denotes the sum of the time for viewpoint candidate generation and the time for viewpoint quality evaluation. We also tested two variations of the selection methods, differing with k and the use of a second camera capture at the chosen v to detect a exact 4-DoF grasp. Without the second sight, the robot can execute the grasp detected in the course of evaluating the quality of the best viewpoint with Algorithm 1 though this grasp can be inaccurate since Algorithm 1 uses estimated depth maps. We find that choosing $k = 10$ than $k = 1$ and using the second sight improves performance for almost all cases, which validates that considering the multi-modality of good grasps for approaching direction selection is crucial

	k=10		k=1		k=10, no second sight		Average
	light clutter (10 objects)	dense clutter (20 Objects)	light clutter (10 objects)	dense clutter (20 Objects)	light clutter (10 objects)	dense clutter (20 Objects)	selection time
CEM with Random Initial Seeds	85.00%	82.50%	85.00%	75.63%	85.83%	81.25%	3.50s
Surface Normal Histogram	81.67%	78.75%	80.00%	78.13%	71.67%	65.00%	0.97s
Sampling Smoothed Normals	86.67%	82.50%	85.83%	75.63%	82.50%	81.25%	1.15s
GADS	**90.83%**	**86.25%**	88.30%	83.13%	81.67%	79.38%	2.48s
	light clutter (10 objects)		dense clutter (20 objects)				
FC-GQ-CNN [4]	87.50%		76.25%				-
Random Selection	77.50%		77.50%				-

TABLE I: Average success rates of various grasp approaching direction selection methods for 20 lightly cluttered scenes and 20 densely cluttered scenes in the MuJoCo simulation. Average computation times for each approaching direction selection methods are also detailed in the rightmost column.

Fig. 5: Two sets of 10 objects each, used in the robot experiment. (Left) Set 1. (Right) Set 2.

and the second sight from the chosen direction is needed for more accurate grasping, respectively.

We can also see that the 4-DoF grasping performs as well as the presented 6-DoF grasping methods in the lightly cluttered scenarios where most objects are isolated, but in the densely cluttered scenarios, 6-DoF grasping methods with the approaching direction selection methods outperformed the 4-DoF grasping, which implies the importance of utilizing full 6-DoF for picking up the densely cluttered objects. Although the Surface Normal Histogram method is the most time-saving, its performance is not significantly better than the Random Selection method since discretization for the 2D histogram leads to limited diversity of approaching directions and degrades the approaching direction quality. Meanwhile, the CEM with Random Initial Seeds method shows improved performance while taking a longer time of 3.50s due to the serial nature of CEM. Overall, GADS performed the best, well-balancing between the performance and time trade-off by exploiting informative geometry prior of smoothed surface normals and reducing the number of samples needed.

B. Evaluation on the Real Robot

We test the GADS method, which is the best approaching direction selection method for 6-DoF grasping verified in simulation, with a 7-DoF arm of a Baxter robot and a RealSense D435 depth camera attached to the arm as shown in Figure 1. We compare the performance of GADS with the state-of-the-art 4-DoF grasp detection method, FC-GQ-CNN [4], to validate that the hierarchical grasping with GADS exploits the benefit of 6-DoF grasping without using 6-DoF grasp data. We prepare 20 objects in total including household objects, office supplies, and fruits, and divide them into two sets of 10 objects each as shown in Figure 5. We follow the same protocol as the simulation experiments on clutter of 10 objects, and evaluate different grasp detection methods over 15 scenes for each set, where the configuration of each scene is artificially reproduced for fair comparison. Success rates at the isolated condition are measured by three grasp attempts per each object in a stable pose.

	Set 1 isolated	Set 2 isolated	Set 1 clutter	Set 2 clutter	Average
FC-GQ-CNN	70.00%	50.00%	56.67%	52.22%	54.44%
GADS	76.67%	53.33%	67.78%	62.22%	**65.00%**

TABLE II: Average success rates of FC-GQ-CNN and GADS for two sets of objects at isolated and cluttered conditions in robot experiments.

The selected objects are challenging to pick up even in the singulated condition as shown in Table II, and especially Set 2 contains five objects which either of two grasping methods fails to grasp in three attempts while Set 1 contains one such object. Performance degradation in real-world experiment mainly results from the poor quality of depth images in real world, where thin objects in isolation are even undetectable in depth images often. Nevertheless, the GADS method for 6-DoF grasping shows its superiority over the 4-DoF FC-GQ-CNN method for cluttered scenes, even improving success rates over the isolated condition by a margin of 8.89% in Set 2. The proposed method has shown tendency to seek the appropriate approaching direction according to the orientations of objects.

VI. Conclusion

In this paper, we have proposed a novel hierarchical approach for 6-DoF grasping which selects the best approaching direction for 4-DoF grasping. The proposed grasp approaching direction selection method, GADS, enables 6-DoF grasping with a 4-DoF grasp quality model trained with 4-DoF grasp data, eliminating the need for collecting time-consuming and expensive 6-DoF grasp data. By exploiting the fully convolutional grasp quality model which can evaluate entire 4D action space in parallel, GADS gains its efficiency, searching only 2D action space through the cross entropy method with initial seeds of surface normals. 6-DoF grasping with GADS has been shown to be advantageous especially in densely cluttered scenes, outperforming the state-of-the-art 4-DoF grasping method. But GADS has a room for improvement. For example, fine-tuning the 4-DoF quality model with more 4-DoF grasp data from the inclined camera view and utilizing multi-camera view to enhance the quality of point clouds and synthesized depth maps will be topics for our future work.

REFERENCES

[1] H. Ahn, S. Choi, N. Kim, G. Cha, and S. Oh, "Interactive text2pickup networks for natural language-based humanrobot collaboration," *IEEE Robotics and Automation Letters*, vol. 3, no. 4, pp. 3308–3315, 2018.

[2] J. S. Park, C. Park, and D. Manocha, "Intention-aware motion planning using learning based human motion prediction," in *Proc. of the Robotics: Science and Systems (RSS)*, Jul. 2017.

[3] 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 *Proc. of the Robotics: Science and Systems (RSS)*, Jul. 2017.

[4] V. Satish, J. Mahler, and K. Goldberg, "On-policy dataset synthesis for learning robot grasping policies using fully convolutional deep networks," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1357–1364, 2019.

[5] 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.

[6] S. Levine, P. Pastor, A. Krizhevsky, J. Ibarz, and D. Quillen, "Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection," *The International Journal of Robotics Research*, vol. 37, no. 4-5, pp. 421–436, 2018.

[7] L. Pinto and A. Gupta, "Supersizing self-supervision: Learning to grasp from 50k tries and 700 robot hours," in *Proc. of the IEEE international Conference on Robotics and Automation (ICRA)*, May 2016.

[8] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, May 2015.

[9] 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.

[10] 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 *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, May 2019.

[11] A. Mousavian, C. Eppner, and D. Fox, "6-dof graspnet: Variational grasp generation for object manipulation," *arXiv preprint arXiv:1905.10520*, 2019.

[12] F. T. Pokorny and D. Kragic, "Classical grasp quality evaluation: New algorithms and theory," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Nov. 2013.

[13] V.-D. Nguyen, "Constructing force-closure grasps," *The International Journal of Robotics Research*, vol. 7, no. 3, pp. 3–16, 1988.

[14] A. Zeng, K.-T. Yu, S. Song, D. Suo, E. Walker, A. Rodriguez, and J. Xiao, "Multi-view self-supervised deep learning for 6d pose estimation in the amazon picking challenge," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, May 2017.

[15] J. Mahler, F. T. Pokorny, B. Hou, M. Roderick, M. Laskey, M. Aubry, K. Kohlhoff, T. Kröger, J. Kuffner, and K. Goldberg, "Dex-net 1.0: A cloud-based network of 3d objects for robust grasp planning using a multi-armed bandit model with correlated rewards," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, May 2016.

[16] A. Herzog, P. Pastor, M. Kalakrishnan, L. Righetti, T. Asfour, and S. Schaal, "Template-based learning of grasp selection," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, May 2012.

[17] Y. Jiang, S. Moseson, and A. Saxena, "Efficient grasping from rgbd images: Learning using a new rectangle representation," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, May 2011.

[18] I.-M. Chen and J. W. Burdick, "Finding antipodal point grasps on irregularly shaped objects," *IEEE transactions on Robotics and Automation*, vol. 9, no. 4, pp. 507–512, 1993.

[19] R. Y. Rubinstein and D. P. Kroese, *The Cross-Entropy Method*. Springer-Verlag, 2004.

[20] D. Kalashnikov, A. Irpan, P. Pastor, J. Ibarz, A. Herzog, E. Jang, D. Quillen, E. Holly, M. Kalakrishnan, V. Vanhoucke, and S. Levine, "Scalable deep reinforcement learning for vision-based robotic manipulation," in *Proc. of the Conference on Robot Learning*, Oct. 2018.

[21] X. Yan, J. Hsu, M. Khansari, Y. Bai, A. Pathak, A. Gupta, J. Davidson, and H. Lee, "Learning 6-dof grasping interaction via deep geometry-aware 3d representations," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, May 2018.

[22] J. Suh, J. Gong, and S. Oh, "Fast sampling-based cost-aware path planning with nonmyopic extensions using cross entropy," *IEEE Transactions on Robotics*, vol. 33, no. 6, pp. 1313–1326, 2017.

[23] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Jun. 2015.

[24] D. Morrison, P. Corke, and J. Leitner, "Closing the Loop for Robotic Grasping: A Real-time, Generative Grasp Synthesis Approach," in *Proc. of Robotics: Science and Systems (RSS)*, Jun. 2018.

[25] A. Zeng, S. Song, K.-T. Yu, E. Donlon, F. R. Hogan, M. Bauza, D. Ma, O. Taylor, M. Liu, E. Romo, N. Fazeli, F. Alet, N. C. Dafle, R. Holladay, I. Morona, P. Q. Nair, D. Green, I. Taylor, W. Liu, T. Funkhouser, and A. Rodriguez, "Robotic pick-and-place of novel objects in clutter with multi-affordance grasping and cross-domain image matching," in *Proc. of the IEEE International Conference on Robotics and Automation*, May 2018.

[26] A. ten Pas and R. Platt, "Using geometry to detect grasp poses in 3d point clouds," in *Robotics Research*. Springer, 2018, pp. 307–324.

[27] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, Jul. 2017.

[28] M. Gualtieri and R. Platt, "Viewpoint selection for grasp detection," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Sep. 2017.

[29] A. Ten Pas and R. Platt, "Localizing handle-like grasp affordances in 3d point clouds," in *Experimental Robotics*. Springer, 2016, pp. 623–638.

[30] E. Todorov, T. Erez, and Y. Tassa, "Mujoco: A physics engine for model-based control," in *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, Oct. 2012.

[31] L. Fan, Y. Zhu, J. Zhu, Z. Liu, O. Zeng, A. Gupta, J. Creus-Costa, S. Savarese, and L. Fei-Fei, "Surreal: Open-source reinforcement learning framework and robot manipulation benchmark," in *Proc. of the Conference on Robot Learning*, Oct. 2018.

[32] K. Mamou and F. Ghorbel, "A simple and efficient approach for 3d mesh approximate convex decomposition," in *Proc. of the IEEE international Conference on Image Processing (ICIP)*, Nov. 2009.

[33] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, May 2011.