

Efficient Representations of Object Geometry for Reinforcement Learning of Interactive Grasping Policies

Malte Mosbach* and Sven Behnke

Abstract—Grasping objects of different shapes and sizes—a foundational, effortless skill for humans—remains a challenging task in robotics. Although model-based approaches can predict stable grasp configurations for known object models, they struggle to generalize to novel objects and often operate in a non-interactive open-loop manner. In this work, we present a reinforcement learning framework that learns the interactive grasping of various geometrically distinct real-world objects by continuously controlling an anthropomorphic robotic hand. We explore several explicit representations of object geometry as input to the policy. Moreover, we propose to inform the policy implicitly through signed distances and show that this is naturally suited to guide the search through a shaped reward component. Finally, we demonstrate that the proposed framework is able to learn even in more challenging conditions, such as targeted grasping from a cluttered bin. Necessary pre-grasping behaviors such as object reorientation and utilization of environmental constraints emerge in this case. Videos of learned interactive policies are available at https://maltemosbach.github.io/geometry_aware_grasping_policies.

I. INTRODUCTION

Grasping is a fundamental capability and the basis for almost all complex manipulation skills, such as pick-and-place or tool use [1]. Consequently, object grasping is of great practical relevance and remains an active area of research. Classical approaches have tackled grasping via analytical, model-based planning and control under the assumption of accurate state estimates and a given dynamics model [2], [3]. Recently, reinforcement learning (RL) has grown in popularity for grasping [4], [5] and object manipulation [6], [7], due to its ability to autonomously discover useful behaviors. The rise of massively parallelized physics simulation [8] and the resulting speedup in RL training has led to a leap in the problem complexity for which proficient controllers can be learned, as evidenced by dexterous in-hand manipulation [9] as well as agile and robust quadruped locomotion [10], [11]. Policies are trained in thousands of environments in parallel from low-dimensional environment states, such as joint or object positions. Surprisingly, skillful in-hand manipulation of geometrically distinct objects can be achieved solely through data-driven learning without access to shape or tactile information [9].

However, in-hand manipulation is so far only concerned with objects of relatively constant size. Objects encountered in grasping and bin-picking scenarios might vary considerably in their extent. It is therefore crucial for a policy to generalize across variations in object shape and size. The effect of object representation on a policy’s robustness to

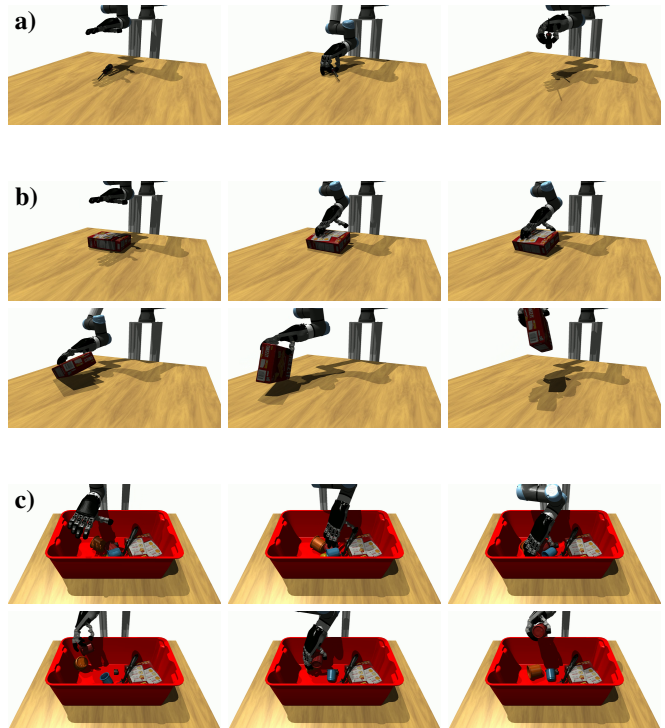


Fig. 1: A single policy learns to grasp objects of vastly different shapes, weights and sizes: (a) Screwdriver that can be readily picked up; (b) Larger cracker box, impossible to grasp from the initial configuration. The policy, nevertheless, learns to lift the object through dexterous pre-grasp manipulation that brings the box into an upright position; (c) Red cup is retrieved from a cluttered container, even as it interacts with other objects and the bin.

variations in shape and size has not been thoroughly investigated, though. Moreover, in-hand manipulation provides dense rewards, as the hand is constantly manipulating and moving the object, achieving states that are closer to or further away from a desired pose. In contrast, object grasping allows the agent to move the hand freely in space. Hence, the majority of random exploration does not change the state of an object and therefore does not provide useful feedback.

In this work, we investigate whether the representation of objects via their center-of-mass is sufficient for our problem, or whether representations that explicitly or implicitly encode object shape can lead to more capable policies. While visual perception, e.g., via RGB-D images, provides detailed information about object geometry, we attempt to find more concise representations. This is motivated because i) learning

*All authors are with the Autonomous Intelligent Systems group of University of Bonn, Germany; mosbach@ais.uni-bonn.de

directly from high-dimensional visual input is more challenging, leading to greatly increased sample complexity, and ii) rendering visual observations in many parallelized environments quickly fills GPU memory, reducing the number of simulations that can be run in parallel and thus undermining the very benefit of massively parallelized training. To avoid these issues, we investigate popular explicit representations of 3D geometries for their performance in RL. Furthermore, we utilize the signed distance function (SDF) to implicitly inform an agent about an object’s shape. We address the sparse reward problem by proposing a novel geometry-aware shaped reward that can achieve the desired reduction of the search space, without biasing the solutions that are found.

We observe that with these representations, existing RL algorithms are able to learn resilient grasping controllers that automatically exhibit pre-grasp manipulation and regrasp behaviors. This allows applying the proposed framework to the more challenging task of targeted picking from a cluttered bin. Notably, even without representation of the context, i.e., other objects and the bin, we are able to learn proficient grasping policies for this scenario. In summary, we make the following contributions:

- A study of explicit shape representations for object manipulation. We investigate different representations of object shape with respect to their RL performance.
- Implicit representation via SDFs is applied to RL of grasping, which has, to the best of our knowledge, not been done in any prior work. From this representation, we infer a geometry-aware reward component encouraging the agent to make contact with an object.
- Picking from cluttered bins is learned without information about the context. We demonstrate that proficient policies can be trained even for this unstructured setting.

II. RELATED WORK

The field of robotic grasping has been an active area of research for many decades. Despite these sustained efforts, the problem remains largely unsolved [12]. Approaches to robotic grasping are typically categorized into analytical and data-driven methods. The former, often also referred to as classical approaches to grasping, is concerned with finding suitable grasp poses for an object, where a kinematically feasible path can then be planned to execute the picking action. The quality of possible grasp poses is assessed through mechanical and geometric conditions, such as force closure or form closure [13]. Data-driven methods have been gaining prominence for robotic grasping due to significant increases in performance as more data and computation becomes available. These subsume approaches for learning-based grasp-synthesis, where only the final grasp pose is learned [14], [15], as well as dexterous manipulation, which typically involves continuous control of the end-effector. Most recent works use reinforcement learning to generate such interactive policies. Rajeswaran et al. [7] combine model-free RL with human demonstrations to learn dexterous manipulation tasks. DQN is used by Deng et al. [16] to learn a policy that picks objects from clutter using a combination

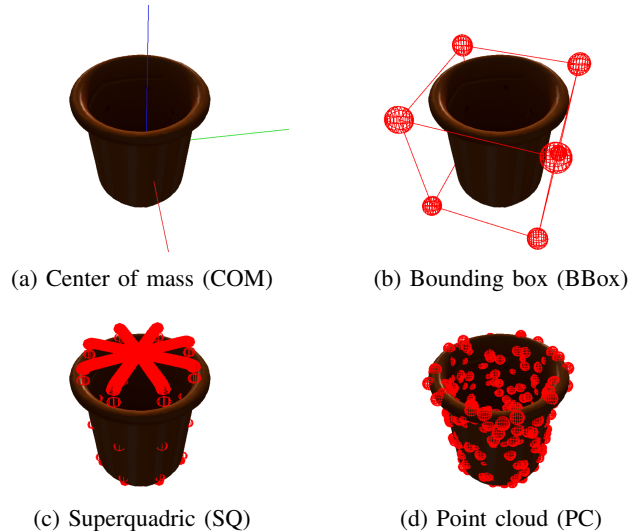


Fig. 2: We examine four explicit representations of object geometry that differ in dimensionality and expressiveness.

of a suction cup and two-fingered gripper. Quillen et al. [17] conduct an empirical evaluation of off-policy model-free RL for vision-based grasping with a parallel gripper.

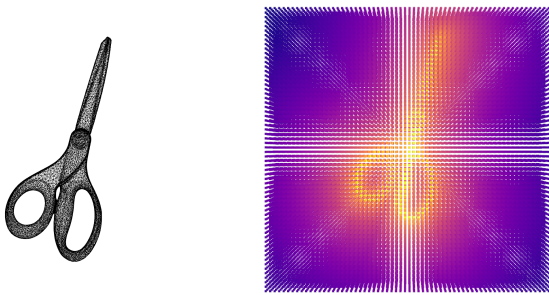
While the representation of object geometry is a prevalent topic in computer vision, few works address it explicitly in the context of RL. Huang et al. [18] study the effect of geometry-awareness for RL of in-hand reorientation and show that a generalist, multi-object policy can be learned that outperforms the single-object baselines. Kumar et al. [1] represent different objects through their bounding boxes. They show that tactile perception is able to compensate for information lost by this approximation as well as measurement noise. In a recent study, Wu et al. [19] use both, an imitation learning and a geometric representation learning objective, to train grasping and manipulation policies from point clouds. However, their framework relies on category-specific grasping demonstrations.

Recent work by Cai et al. [20] utilizes the truncated SDF (TSDF) for contact point detection. They demonstrate visual grasping from a cluttered bin with a two-fingered gripper. Breyer et al. [21] propose a volumetric grasping network that operates directly on a TSDF representation of a cluttered scene and outputs grasp poses and qualities. To the best of our knowledge, no prior work has explored the signed distance function for its potential in RL for object manipulation.

In this work, we perform a comparative study of explicit object representations for RL-based grasping. Further, we propose to implicitly represent object geometry through the SDF, with leads to succinct states amenable to large-scale optimization and which naturally lend themselves to rewarding behaviors useful for object grasping.

III. METHOD

We formalize the task of grasping an object as a Markov decision process $\mathcal{M} = (\mathcal{S}, \mathcal{A}, R, P, \gamma)$, where \mathcal{S} and \mathcal{A}



(a) Detailed triangular mesh of an object (b) SDF representation of the object saved as a voxel grid

Fig. 3: We precompute voxelized SDFs for object meshes to enable fast lookups of distances during training.

are the state and action space, respectively. The transition function $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}_+$ represents the probability of transitioning to the next state $s_{t+1} \in \mathcal{S}$ when taking action $a_t \in \mathcal{A}$ in the current state $s_t \in \mathcal{S}$. $R : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ is a reward function and $0 \leq \gamma < 1$ is a discount factor.

Our goal is to learn a policy π_θ that can grasp a large number N of geometrically diverse objects. This induces a multi-task problem, the goal of which can be formalized as optimizing the expected sum of discounted rewards over all N objects:

$$\mathbb{E}_{\pi_\theta} \left[\sum_{i=1}^N \sum_{t=0}^{T-1} \gamma^t R(s_t, \pi_\theta(s_t)) \right]. \quad (1)$$

As we will show, naive optimization for this objective without incorporating object-specific information can be surprisingly successful if sufficient amounts of experience can be generated. There is, however, one major drawback to this approach. As the policy cannot distinguish different objects from each other, a common, universally effective strategy must be learned [18]. This puts an upper limit to the dexterity that can emerge during optimization. Instead, we aim to learn a policy that adapts to the different objects it encounters and is able to employ specialized, geometry-aware strategies to grasp them. Thus, we investigate representations of object shape for their performance in RL of grasping policies.

A. Explicit Representation of Object Geometry

The explicit representations of an object’s geometry we explore in this work are depicted in Figure 2.

1) *Center of Mass*: The naive representation of objects via their center of mass position or 6D pose is prevalent in reinforcement learning of single-object manipulation policies [18], [22]. We employ this representation as a geometry-agnostic baseline.

2) *Bounding Box*: 3D bounding boxes represent the cuboid hull in which an object lies. We extract the oriented bounding box along the three principal axes of a given mesh. While prior work [1] has studied the effect of bounding box representations for grasping, this has only been investigated for contextual RL. To provide a succinct observation to the

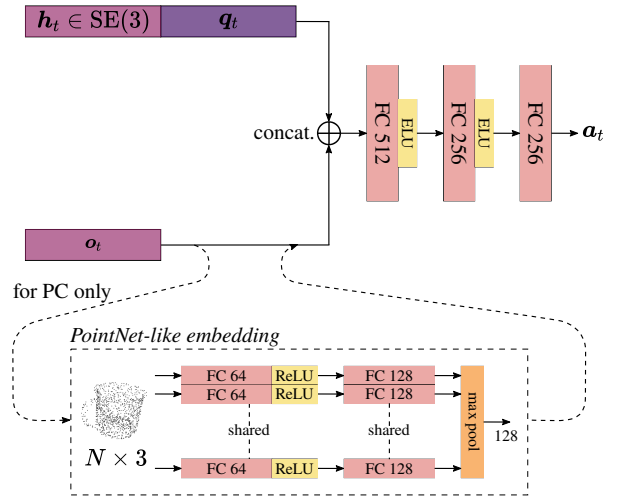


Fig. 4: Policy architecture. h_t is the hand pose, q_t are the joint positions of the hand, and o_t is the object representation at time step t . When the object is represented as a point cloud (PC), we preprocess the input via a PointNet-like architecture [27] to generate a permutation invariant embedding first. The three-layer fully connected neural network outputs actions a_t that are changes to the 6DoF hand pose and to the five controllable DoF of an anthropomorphic hand.

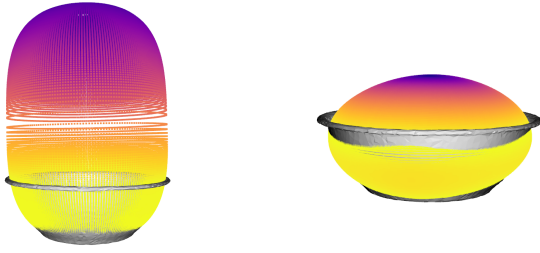
policy, we represent the bounding box via its central pose and extent.

3) *Superquadric*: Superquadrics go beyond the standard 3D cuboid representation. They are an 11-dimensional parametric family of shapes able to represent cubes, cylinders, spheres, etc. [23]. Due to their capacity to represent a variety of shapes with very few parameters, superquadrics have been used as object representations for manipulation [24], [25] and shape parsing [26]. We are, however, not aware of prior work using superquadrics in the context of RL. The section III-B outlines how we recover superquadric parameters for an object.

4) *Point Cloud*: Point clouds are a natural representation of 3D objects as they can be created by laser scanners or RGB-D cameras. However, they are irregular in the sense that there is no natural ordering of the points. To respect the permutation invariance of the input, we preprocess a point cloud with a PointNet-like [27] architecture before feeding the embedding into the MLP policy. We perform experiments for point clouds of different sizes, sampling 32, 128, and 512 points on the mesh surface, respectively.

B. Superquadric Recovery

Accurate recovery of superquadrics is a difficult problem. Liu et al. [28] have recently proposed a robust algorithm for fitting superquadrics to point clouds. They formulate the recovery of superquadrics as a maximum likelihood estimation problem and propose a variation of the expectation-maximization algorithm that exploits the geometric features of superquadrics to solve it. Their approach is called Expectation, Maximization, and Switching (EMS). While excellent



(a) Solution found by the original EMS algorithm (b) Solution found using our regularization

Fig. 5: We regularize the recovery of superquadrics by minimizing the distance of the points lying on the superquadric surface to the object mesh. The color indicates the distance of each point from the mesh.

results can be obtained for objects whose shape is approximately within the geometric scope of superquadrics, objects outside this domain can produce misleading solutions. This can be seen in the top row of Figure 6. While the cuboid object is recovered very well, the other objects lead to unsatisfactory results. This is because the solutions are found solely by optimizing the probability $p(\mathbf{x}_i; \theta)$ of points on the mesh surface \mathbf{x}_i under the superquadric parameters θ . As can be seen for the second object, although a large region of the superquadric is far from the bowl, all points on the mesh surface are very close to the surface of the superquadric, which defines optimality in this case. While the authors propose to use an additional term in the loss function expressing a preference for superquadrics with smaller surface area, we have found that this negatively affects performance on objects that could otherwise be recovered accurately, as it biases the solution towards spherical parameterizations. Instead, we introduce a second loss term that minimizes the distance of points on the superquadric surface from the original mesh surface. The regularization effect can be seen in Figure 5. Figure 6 highlights the difference our adaption makes for recovering superquadrics on the YCB objects. It can be seen that the representation of the cuboid object remains accurate, while the approximations of objects that cannot be fitted precisely converge to more representative superquadric parameterizations.

C. Implicit SDF Representation

In addition to the explicit representations mentioned above, complex geometries can also be represented implicitly. A signed distance function is a continuous mapping $\phi(\mathbf{x}) : \mathbb{R}^3 \rightarrow \mathbb{R}$ that projects a 3D point \mathbf{x} to its distance to a surface, where the sign indicates whether the point is inside (-) or outside (+) the surface. To make this information available to a RL policy, we evaluate the SDF on a small set of points. Specifically, we query the distance to the object surface at the fingertip positions of the robot hand, resulting in a 5-dimensional observation $(d(\mathbf{x}_1^{ft}, \delta\mathcal{O}), \dots, d(\mathbf{x}_5^{ft}, \delta\mathcal{O}))$, where \mathbf{x}_i^{ft} is the position of the i -th fingertip and $\delta\mathcal{O}$ is the

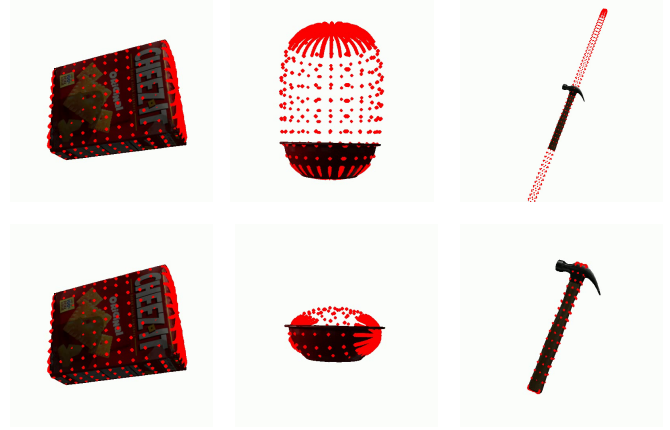


Fig. 6: Superquadrics recovered for YCB objects: Top: Original EMS algorithm; Bottom: Our regularized version.

object surface. Hence, the agent can perceive the distance of its fingers to the object surface and is thus implicitly informed about the geometry it is interacting with.

Computing the signed distance of a set of points to a high-fidelity mesh is expensive [29], however, and would be required in every environment at every step of the simulation. To nevertheless make this observation type feasible for massively parallel reinforcement learning, we precompute a discrete voxel-based representation of the SDF for all object meshes (see Figure 3). Specifically, we create a $200 \times 200 \times 200$ grid filling the space $[-0.5m, 0.5m]^3$ around the object center. This creates a spatial resolution of 5 mm, which we found to be sufficient in our experiments. When query points lie outside the volume of the voxel-grid, we return the result from the closest voxel. With this approach, the lookup of signed distances to the robot fingertips takes only 0.31 ms for all 16,384 parallel environments, leading to negligible impact on the overall training performance. SDF voxel-grids computed once for an object mesh are stored permanently to enable fast reloading in future training runs.

D. SDF-based Reward Shaping

Using our SDF-based observation of the distance between end-effector and object, we introduce a reward term that incentivizes the agent to make maximum contact with whatever object it interacts with. Specifically, we reward minimizing the distance of the fingertips to the object surface:











$$r_{sdf} = \frac{1}{d_t + \epsilon_{sdf}}, \quad (2)$$

where $d_t = \sum_i d(\mathbf{x}_i^{ft}, \delta\mathcal{O})$ and ϵ_{sdf} is a constant. We hypothesize that this will help reducing the search space for the policy, while being unbiased and object-independent, since no particular position or orientation is imposed for grasping an object, only that the agent should make contact.

E. Policy Optimization

We use proximal policy optimization (PPO) [30] to learn the agents, since the algorithm has been shown to converge robustly on problems with high-dimensional action spaces, and

TABLE I: Variations in lifting success rates for different representations of object geometry. For point clouds, the number after the dash indicates the number of points sampled on the object surface.

Object	Success rate						
	COM	SDF	BBox	SQ	PC-32	PC-128	PC-512
 007_tuna_fish_can	0.99 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
 013_apple	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
 029_plate	0.74 ± 0.14	0.95 ± 0.0	0.93 ± 0.02	0.95 ± 0.04	0.5 ± 0.39	0.94 ± 0.03	0.94 ± 0.04
 032_knife	0.77 ± 0.07	0.81 ± 0.07	0.96 ± 0.02	0.85 ± 0.1	0.71 ± 0.02	0.9 ± 0.02	0.79 ± 0.06
 053_mini_soccer_ball	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
 062_dice	0.65 ± 0.3	0.85 ± 0.08	0.95 ± 0.02	0.94 ± 0.01	0.55 ± 0.07	0.86 ± 0.0	0.6 ± 0.12
 065-a_cups	0.99 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
 072-a_toy_airplane	0.93 ± 0.02	0.99 ± 0.01	0.99 ± 0.01	0.97 ± 0.03	0.95 ± 0.01	0.99 ± 0.0	0.96 ± 0.02
 073-g_lego_duplo	0.9 ± 0.01	0.97 ± 0.01	0.99 ± 0.0	0.96 ± 0.02	0.95 ± 0.01	0.99 ± 0.0	0.98 ± 0.01
 077_rubiks_cube	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0
All 78 YCB objects	0.95 ± 0.02	0.96 ± 0.01	0.97 ± 0.01	0.96 ± 0.02	0.92 ± 0.03	0.97 ± 0.0	0.95 ± 0.02

make use of the high-performance implementation provided by Makoviichuk and Makoviyuchuk [31]. We run 16,384 parallel environments to collect data. After every 32 steps, we update the policy for 5 epochs on the rollout data with a batch size of 32,768. All experiments are run on a single NVIDIA RTX A6000 GPU with 48GB of memory. The policy is computed by a three-layer fully connected neural network, as detailed in Figure 4.

IV. EXPERIMENTAL SETUP

We use NVIDIA Isaac Gym [8] as our physics simulator. The robot consists of a UR5e arm and the Schunk SIH hand, which features 11 joints, the actuation of which is coupled, leading to five controllable degrees of freedom (DoF).

A. Observations, Actions, and Rewards

Observations consist of the current pose of the robot hand and its joint positions as well as a representation of the object. For targeted picking from cluttered bins, \mathbf{o}_t represents the observation of the target object only. Hence, we do not inform the policy about other objects or the walls of the bin. In this scenario, a random object in the bin is selected as the target at the beginning of each episode.

The agent controls the robot via 11-dimensional actions, representing the desired change in position and rotation of the hand pose and the controllable DoF positions of the hand. From the desired pose of the robot hand, we compute the UR5e joint position targets using the Jacobian transpose method. The action frequency is 30 Hz.

An episode terminates when the object falls off the table, is lifted to the target height, or the policy runs out of time after 300 actions taken. We use the following reward function

to train the lifting policies:

$$r(s_t, a_t) = c_1 \frac{1}{|\bar{h} - \Delta h_t| + \epsilon_h} + c_2 \mathbb{1}(\Delta h_t \geq \bar{h}) + c_3 r_{sdf} \quad (3)$$

where Δh_t is the height an object has been lifted by, relative to its starting position, and \bar{h} is the target value for Δh_t . The first two reward terms incentivize the policy to lift the object up, while the last term reduces the search space by encouraging the agent to make contact with the object. We set \bar{h} to 0.2 m for our experiments and weigh the terms with $c_1 = 0.5$, $c_2 = 5000$, and $c_3 = 1$, where $\epsilon_h = 0.02$ and $\epsilon_{sdf} = 0.025$. We evaluate the success of a policy rollout by whether an object was raised by \bar{h} , whereupon the episode terminates. We report the mean and standard deviation of three seeds for all experiments.

B. Object Dataset

We consider the problem of grasping and lifting objects of variable shape and size. To this end, we evaluate our policies on all 78 YCB object models [32] for which laser scans exist. The weights of all objects are set to realistic values and their inertia tensors are computed from the geometry in Isaac Gym. The objects are spawned in random poses and then dropped onto a table to create realistic initial configurations.

V. RESULTS

We evaluate the impact of geometry-awareness and the type of object representation on the performance of RL-based grasping policies. Further, we analyze the effect of our geometry-aware shaped reward on the learning behavior. Finally, because we observe that PPO is able to learn robust grasping policies that naturally exhibit useful behaviors such as pre-grasp manipulation and re-grasping, we test whether

TABLE II: Sample efficiency of different object representations measured by the number of environment steps (in million) and the runtime in minutes required to reach success rate thresholds.

Object repr.	90 %		95 %	
	Steps	Runtime [min]	Steps	Runtime [min]
COM	149 ± 30	47 ± 9	550 ± 161	116 ± 34
SDF	134 ± 5	72 ± 3	276 ± 51	152 ± 28
BBox	165 ± 8	69 ± 3	274 ± 44	127 ± 23
SQ	178 ± 18	58 ± 6	436 ± 52	220 ± 26
PC-32	196 ± 15	52 ± 4	474 ± 170	178 ± 63
PC-128	173 ± 4	47 ± 2	232 ± 1	65 ± 1
PC-512	219 ± 19	62 ± 6	277 ± 12	87 ± 4

our method can be straightforwardly applied to picking objects from cluttered bins.

A. Role of Object Representation

In the following, we discuss the role of object representations in learning grasping policies. We compare the performance that the different representations achieve during training both in terms of sample efficiency and final performance to capture the strengths and weaknesses of both succinct and expressive representations.

Table I reports the final test success rate, both overall and for specific objects, after training for a total of 500 epochs, where each epoch consists of 32 steps in 16,384 parallel environments. It can be seen that the geometry-independent policies provide a very strong baseline, even for grasping objects with significantly different shapes and sizes. Nevertheless, we were able to demonstrate a performance advantage from geometry awareness. Especially for very small objects, such as the dice, or objects far from cubic shape, such as the plate or knife, geometry-awareness enables learning of specific grasping strategies. While observations of the distance of the fingertips from the object surface learn more quickly, their final performance falls short of that reached via bounding boxes or point clouds. Bounding box representations showed strong final performance while maintaining high convergence speed. They introduce the geometric information necessary to learn a proficient object-specific policy for the studied task. Superquadrics reach a final performance just short of the bounding box policies. It should be noted that superquadrics subsume the cuboid representation of objects via bounding boxes. While they are in principle strictly more expressive, using superquadrics introduces the added challenge of fitting them to a given object mesh. While they are able to fit many common objects precisely, superquadric representations may be misleading for objects that cannot be approximated well. Despite our efforts to ensure that meaningful superquadric parameterizations were recovered for all objects, the resulting policies performed worse overall than the bounding box policies. For point cloud-based observations, we encountered a strong dependence of performance on the number of sampled points. For only 32 points, this observation type did not benefit the policy and actually leads to a reduction

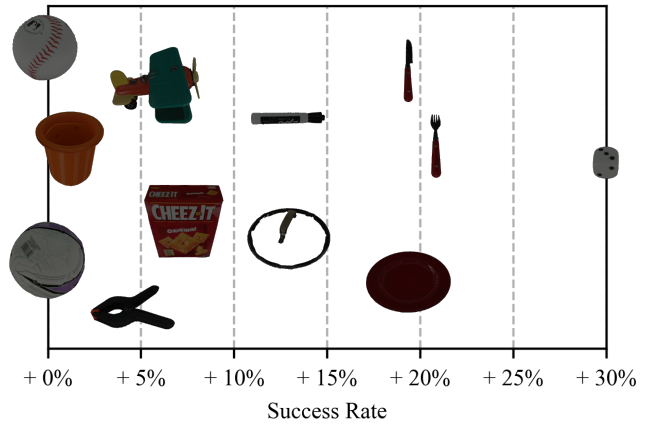


Fig. 7: Performance gains of geometry-aware (BBox) policy over the COM baseline. While there are little to no improvements for cubic or spherical objects of average extent, significant performance improvements can be achieved for objects of irregular shape and size. Visualization inspired by Huang et al. [18].

in final performance, compared to the geometry-independent baseline. As more points are sampled, training performance improves, but speed of convergence goes down. We found the best trade-off to be at 128 points sampled per object, resulting in a policy that converges quickly and to the shared highest overall performance.

Table II reports the number of samples required for the policies to achieve an average success rate of 90 % or 95 % across all objects during training. It can be seen, that the proposed implicit representation of object geometry via signed distances reaches the success rate threshold of 90 % first. The representation of object via 128 sampled points already outperforms all others when it comes to reaching a success rate of 95 %. Learning from point clouds with only 32 samples does not adequately capture object geometry, resulting in unstable training performance and thus significantly worse performance compared to the other point cloud runs. While the geometry-agnostic representation by the object’s center of mass is able to reach the first threshold quickly, improving performance above 90 % becomes challenging, given the limited information available.

While the gains achieved by incorporating object shape may seem minor in terms of the success rate average across the entire object set, it is important to note that the COM baseline works very well for the majority of objects studied. The most interesting comparisons can be made for objects of irregular shape and size. We find that geometry-awareness provides significant improvements for these object types where a universal, object-independent strategy fails. Figure 7 shows this performance gain for specific objects. It highlights the correlation of performance gained through shape-awareness and object irregularity. The dice, for which the biggest improvement can be made is the smallest object in the entire dataset. The fork and knife, which also profit

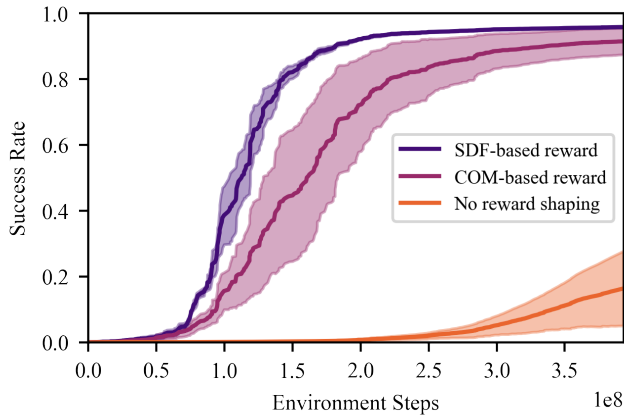


Fig. 8: Average success rate achieved during training with different reward functions.

significantly from shape-awareness, have some of the largest aspect ratio of their extent along the three principal axes.

B. SDF-based Reward Shaping

Next, we benchmark our proposed SDF-based reward component against two ablations. The comparative training performance is shown in Figure 8.

Not rewarding the policy based on the position of its end-effector relative to the object, i.e. setting $c_3 = 0$ in Eq. 3, leads to a drastic deterioration in performance, as most transitions do not receive a meaningful reward signal, which hinders learning progress. In a second ablation, we replace the distance of the fingers to the object surface by their distance to its center of mass. This represents a naive formulation of our contact-inducing reward. The results show that this leads to accelerated learning, since the agent is encouraged to move its hand close to the object. However, this simplification introduces bias in how to grasp an object, motivating the policy to always reach for the center of mass. Especially for larger or more irregularly shaped objects, this strategy becomes infeasible. Consequently, the COM-based reward shaping performs inferior to our proposed SDF-based reward both in terms of sample efficiency and final performance. Thus, we can confirm that the novel SDF-based reward induces the desired effect of guiding the search effectively, while being agnostic to how an object is grasped by the policy. This leads to faster convergence and improved final performance when grasping the YCB objects.

C. Application to Cluttered Bin-Picking

To analyze whether the interactive policies can utilize pre-grasp manipulation and regrasping to solve more demanding tasks, we apply our method to targeted picking from cluttered bins. To initialize a scenario, we drop five randomly chosen YCB objects into a bin. The agent is tasked to retrieve a specific object from this cluttered bin. We provide the policy with the same information as for the unobstructed grasping of objects lying on the table. Hence, only proprioceptive measurements of the robot’s state and the state of the target

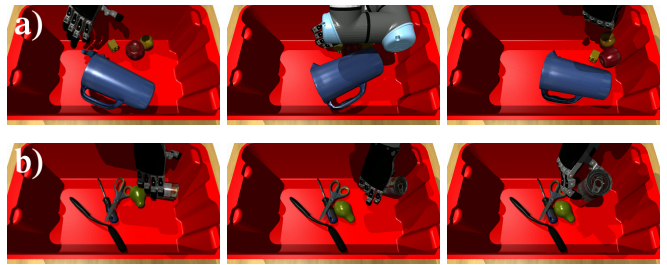


Fig. 9: Strategies learned for targeted bin picking: (a) The agent lifts the tiny airplane part by sliding it up the near wall; (b) The tomato soup can is initially in a position that is difficult to reach. Through continuous manipulation, the policy succeeds in bringing it into an upright pose from which it can be grasped.

object are included in the observation. We represent the target object via its bounding box, since strong performance was obtained for this low-dimensional observation type in the single object case. The trained policy achieves a success rate of 0.85 ± 0.02 , averaged over all objects. When we increase task difficulty by dropping ten random objects in the bin, the same policy still achieves a success rate of 0.81 ± 0.04 , underlining the robustness of the strategies that have been learned. Interesting behaviors that emerged in this setting are shown Figure 9. They include extended pre-grasp interactions to reorient objects and the utilization of environmental constraints, such as the bin walls.

Naturally, the policy cannot reach quite the same performance as for the single object case. Nevertheless, the results show that our method is able to learn a proficient controller even for this challenging setting. This approach can serve as a strong baseline for future work on targeted picking from cluttered bins that explicitly takes the context into account.

VI. CONCLUSIONS AND DISCUSSION

Our work showed that regular on-policy optimization can train a single agent that is able to grasp various geometrically diverse objects. We confirm that learning with shape-independent observations can bring about a universal strategy that achieves strong performance. Nevertheless, simple geometry-conditioned observations can boost the performance significantly for objects of irregular shape and size, leading to more generally capable policies. Basic observations such as bounding boxes proved sufficient for this task and offered a strong trade-off between expressiveness and compactness in our experiments. Point clouds showed very strong training performance when the right number of surface points is sampled. While a higher number of sampled points requires increased training times, combining reinforcement learning with a pre-trained point-cloud encoder might alleviate this issue and presents an interesting avenue for future work. Since the learned agents exhibit robust interactive grasping strategies, we evaluated the performance of our framework for the unstructured scenario of retrieving a specific object from a cluttered bin. Even without any context

information we were able to learn proficient interactive policies for this task that reorient objects and utilize environmental constraints. Our framework can therefore serve as a proficient baseline for future work that explicitly considers object interactions and environment representations. Robust inference of succinct object descriptions from data available in the real world, such as RGB-D images, remains an open problem and is left to future work.

Furthermore, we introduced a novel geometry-aware reward function that benefits both the speed of convergence as well as final performance of the interactive grasping policies. Minimizing the distance between the end-effector and object surface proved to be an essential addition, as it effectively guides the search without imposing limitations on the specific grasping strategy the policy finds. We demonstrated that signed distances required to evaluate this reward can be approximated efficiently, making it readily applicable for high-performance RL. Hence, this formulation of the reward can potentially be a valuable addition for reinforcement learning of grasping policies in general, for example, for different gripper morphologies or to promote stable grasps when using tools.

ACKNOWLEDGMENT

This work has been funded by the German Ministry of Education and Research (BMBF), grant no. 01IS21080, project "Learn2Grasp: Learning Human-like Interactive Grasping based on Visual and Haptic Feedback".

REFERENCES

- [1] V. Kumar, T. Hermans, D. Fox, S. Birchfield, and J. Tremblay, "Contextual reinforcement learning of visuo-tactile multi-fingered grasping policies," *arXiv preprint arXiv:1911.09233*, 2019.
- [2] V. Kumar, Y. Tassa, T. Erez, and E. Todorov, "Real-time behaviour synthesis for dynamic hand-manipulation," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 6808–6815.
- [3] I. Mordatch, Z. Popović, and E. Todorov, "Contact-invariant optimization for hand manipulation," in *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, 2012, pp. 137–144.
- [4] S. Joshi, S. Kumra, and F. Sahin, "Robotic grasping using deep reinforcement learning," in *IEEE 16th International Conference on Automation Science and Engineering (CASE)*, 2020, pp. 1461–1466.
- [5] M. Q. Mohammed, K. L. Chung, and C. S. Chyi, "Review of deep reinforcement learning-based object grasping: Techniques, open challenges, and recommendations," *IEEE Access*, vol. 8, pp. 178450–178481, 2020.
- [6] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, *et al.*, "Solving Rubik's cube with a robot hand," *arXiv preprint arXiv:1910.07113*, 2019.
- [7] A. Rajeswaran, V. Kumar, A. Gupta, G. Vezzani, J. Schulman, E. Todorov, and S. Levine, "Learning complex dexterous manipulation with deep reinforcement learning and demonstrations," in *Robotics: Science and Systems (RSS)*, 2018.
- [8] V. Makoviychuk, L. Wawrzyniak, Y. Guo, M. Lu, K. Storey, M. Macklin, D. Hoeller, N. Rudin, A. Allshire, A. Handa, and G. State, "Isaac Gym: High performance GPU based physics simulation for robot learning," in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 2021.
- [9] T. Chen, J. Xu, and P. Agrawal, "A system for general in-hand object re-orientation," in *Conference on Robot Learning (CoRL)*, PMLR, 2022, pp. 297–307.
- [10] N. Rudin, D. Hoeller, P. Reist, and M. Hutter, "Learning to walk in minutes using massively parallel deep reinforcement learning," in *Conference on Robot Learning (CoRL)*, PMLR, 2022, pp. 91–100.
- [11] G. B. Margolis, G. Yang, K. Paigwar, T. Chen, and P. Agrawal, "Rapid locomotion via reinforcement learning," *arXiv preprint arXiv:2205.02824*, 2022.
- [12] K. Kleeberger, R. Bormann, W. Kraus, and M. F. Huber, "A survey on learning-based robotic grasping," *Current Robotics Reports*, vol. 1, no. 4, pp. 239–249, 2020.
- [13] H. Zhang, J. Tang, S. Sun, and X. Lan, "Robotic grasping from classical to modern: A survey," *arXiv preprint arXiv:2202.03631*, 2022.
- [14] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, and S. Birchfield, "Deep object pose estimation for semantic robotic grasping of household objects," in *Conference on Robot Learning (CoRL)*, 2018.
- [15] Z. Dong, S. Liu, T. Zhou, H. Cheng, L. Zeng, X. Yu, and H. Liu, "Ppr-net: Point-wise pose regression network for instance segmentation and 6d pose estimation in bin-picking scenarios," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 1773–1780.
- [16] Y. Deng, X. Guo, Y. Wei, K. Lu, B. Fang, D. Guo, H. Liu, and F. Sun, "Deep reinforcement learning for robotic pushing and picking in cluttered environment," in *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 619–626.
- [17] D. Quillen, E. Jang, O. Nachum, C. Finn, J. Ibarz, and S. Levine, "Deep reinforcement learning for vision-based robotic grasping: A simulated comparative evaluation of off-policy methods," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2018, pp. 6284–6291.
- [18] W. Huang, I. Mordatch, P. Abbeel, and D. Pathak, "Generalization in dexterous manipulation via geometry-aware multi-task learning," *arXiv preprint arXiv:2111.03062*, 2021.
- [19] Y.-H. Wu, J. Wang, and X. Wang, "Learning generalizable dexterous manipulation from human grasp affordance," *arXiv preprint arXiv:2204.02320*, 2022.
- [20] J. Cai, J. Su, Z. Zhou, H. Cheng, Q. Chen, and M. Y. Wang, "Volumetric-based contact point detection for 7-dof grasping," 2022.
- [21] M. Breyer, J. J. Chung, L. Ott, S. Roland, and N. Juan, "Volumetric grasping network: Real-time 6 dof grasp detection in clutter," in *Conference on Robot Learning*, 2020.
- [22] M. Plappert, M. Andrychowicz, A. Ray, B. McGrew, B. Baker, G. Powell, J. Schneider, J. Tobin, M. Chociej, P. Welinder, *et al.*, "Multi-goal reinforcement learning: Challenging robotics environments and request for research," *arXiv preprint arXiv:1802.09464*, 2018.
- [23] A. H. Barr, "Superquadrics and angle-preserving transformations," *IEEE Computer Graphics and Applications*, vol. 1, no. 1, pp. 11–23, 1981.
- [24] E. C. e. Silva, M. F. Costa, W. Erlhagen, and E. Bicho, "Superquadrics objects representation for robot manipulation," in *AIP Conference Proceedings*, vol. 1738, 2016, p. 300004.
- [25] G. Vezzani, U. Pattacini, and L. Natale, "A grasping approach based on superquadric models," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 1579–1586.
- [26] D. Paschalidou, A. O. Ulusoy, and A. Geiger, "Superquadrics revisited: Learning 3D shape parsing beyond cuboids," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 10344–10353.
- [27] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 652–660.
- [28] W. Liu, Y. Wu, S. Ruan, and G. S. Chirikjian, "Robust and accurate superquadric recovery: A probabilistic approach," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 2676–2685.
- [29] Y. Narang, K. Storey, I. Akinola, M. Macklin, P. Reist, L. Wawrzyniak, Y. Guo, A. Moravanszky, G. State, M. Lu, *et al.*, "Factory: Fast contact for robotic assembly," *arXiv preprint arXiv:2205.03532*, 2022.
- [30] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
- [31] D. Makoviychuk and V. Makoviychuk, *RI-games: A high-performance framework for reinforcement learning*, https://github.com/Denys88/rl_games, 2022.
- [32] 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 *International Conference on Advanced Robotics (ICAR)*, IEEE, 2015, pp. 510–517.