Learning Diverse and Physically Feasible Dexterous Grasps with Generative Model and Bilevel Optimization

Anonymous Author(s)
Affiliation
Address
email

Abstract: To fully utilize the versatility of a multi-finger dexterous robotic hand for object grasping, one must satisfy complex physical constraints introduced by hand-object interaction and object geometry during grasp planning. We propose an integrative approach of combining a generative model and a bilevel optimization to compute diverse grasps for novel unseen objects. First, a grasp prediction is obtained from a conditional variational autoencoder trained on merely six YCB objects. The prediction is then projected onto the manifold of kinematically and dynamically feasible grasps by jointly solving collision-aware inverse kinematics, force closure, and friction constraints as one nonconvex bilevel optimization. We demonstrate the effectiveness of our method on hardware by successfully grasping a wide range of unseen household objects, including adversarial shapes challenging to other types of robotic grippers.

Keywords: Dexterous grasping, Grasp planning, Bilevel optimization, Generative model

1 Introduction

Performing diverse grasps on a variety of objects is a fundamental skill in robotic manipulation. Diverse grasps allow flexible interaction with the objects while satisfying requirements demanded by the higher level task of interest. *Dexterous grasping*, which refers to object grasping with a fully actuated, multi-finger dexterous robotic hand, has the potential to achieve diverse grasp configurations with applicability to a wide range of objects. This is in contrast with *simple grasps* achieved by parallel jaw grippers and underactuated multi-finger grippers whose fingertip workspace is restricted to a manifold instead of the full 3D space.

Planning dexterous grasps is significantly more complex than planning simple grasps. On a dexterous hand, the number of degrees of freedom (DoF) in each finger results in a much larger planning space beyond deciding the hand's 6D *wrist pose* in space. Moreover, the range of motion provided by fully-actuated fingers demands explicit consideration of dynamic constraints such as force closure and friction cone. Lastly, the geometric interaction between the object and gripper presents collision and contact constraints. Consequently, the manifold representing physically feasible grasp solutions in a dexterous grasp planning problem is thus highly complex and challenging to identify.

While the complexity of planning dexterous grasps motivates the use of learning-based approaches, requiring the learned policy to produce grasps that precisely lie on the physically feasible manifold is difficult. On the other hand, an analytical approach can enforce physical constraints exactly, yet navigating the combinatorially complex contact choices analytically is computationally intractable.

We propose a systematic approach that combines a learned neural network and a bilevel optimization to produce physically-feasible dexterous grasps for diverse objects. Our method first predicts an initial set of contact points using a conditional variational autoencoder (CVAE). The contact

Submitted to the 6th Conference on Robot Learning (CoRL 2022). Do not distribute.

points are then projected onto the manifold of kinematically and dynamically feasible grasps by solving a bilevel optimization. Our key innovations are summarized below:

- **CVAE-based dexterous grasp prediction network** that takes in an object point cloud and produces diverse grasp finger placement predictions parameterized by the latent variable sample.
- **Bilevel grasp optimization formulation** that enforces fingertip contact, collision-free, inverse kinematics (IK), force closure, and friction cone constraints for a given robotic grasp.
- **Dexterous grasp planning pipeline** that leverages the two aforementioned components to produce diverse, fully-specified dexterous grasps from object point cloud observations.
- **Successful hardware validation** of the proposed planning pipeline on 13 household objects which include novel geometries and adversarial objects challenging for non-dexterous grippers.

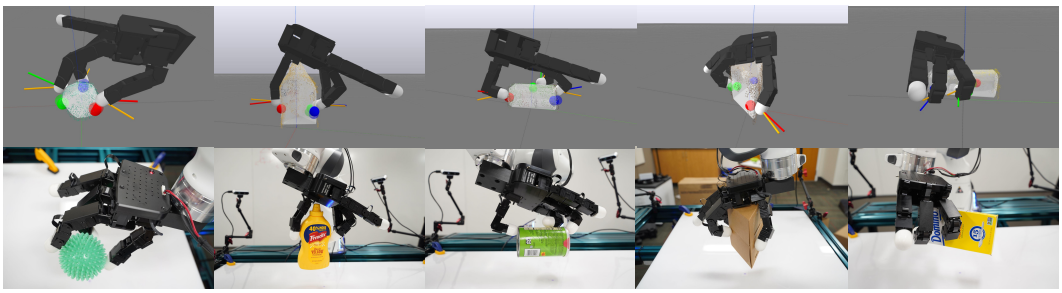


Figure 1: Our method is able to pick up objects with different shapes using a diverse set of grasps.

2 Related Work

Robot manipulation has been an active and broad research area for decades. We limit our discussion to papers that focus on grasping in uncluttered scenes. We exclude work that focuses on object segmentation in cluttered environments or other types of manipulation, such as in-hand manipulation.

2.1 Simple and Dexterous Grasping in Literature

In the simple grasp setting, grasp planning reduces to computing a 6-DoF gripper pose in space. We point the readers to [1, 2, 3, 4] for a thorough review. Generally, analytical simple grasp planners either reason about the object geometry (e.g. [5]) or rank grasps using grasp quality metrics [2]. Recently, learning-based approaches that predict wrist poses have gained significant traction (e.g. [6, 7, 8]). We refer the readers to [3, 4, 9] for a review.

The dexterous grasping literature can also be classified into analytical and learning-based approaches. In additional to object geometry, analytical dexterous grasping methods are often inspired by physical constraints such as force closure, friction cone, robustness to external disturbance wrenches (e.g. [10, 11, 12]). We point the readers to [2, 3] for a detailed review. The main challenges in these approaches include dependency on a high-fidelity object model and lack of diversity in generated grasps [3].

Early learning-based dexterous grasp planners seek to navigate the space of possible grasps (e.g. [10]). More recent papers have shifted to complex model architectures such as generative models [13, 14] and convolutional neural networks [15, 16] to produce grasps directly. The main limitation of these works lies in restricted hardware evaluation. As the sim-to-real transfer of robotic grasps is not obvious due to the difficulty of simulating contact-rich interactions [17, 18], the quality of grasps produced directly by a model trained in simulation is debatable. Moreover, these papers do not emphasize the ability to learn *diverse* grasps, which is a key benefit of dexterous grasping.

Notably, some publications on learning dexterous grasping originate from computer vision and graphics communities (e.g. [19, 20, 21, 22, 23, 24]). While these papers cover diverse grasp generation on fully dexterous hands, their ultimate objective tends to be achieving photorealistic human grasps in simulation rather than satisfying strict real-world physical constraints.

2.2 Bilevel Optimization for Planning

While the theory on bilevel optimization is well established in literature [25], the application of bilevel optimization on motion planning is relatively scarce. Most of these applications are either in continuous systems [26] or robotic locomotion [27, 28]. Other applications include collaborative object manipulation [29] and task and motion planning [30]. To the authors' best knowledge, our work is the first application of bilevel optimization on autonomous robotic manipulation.

3 Method

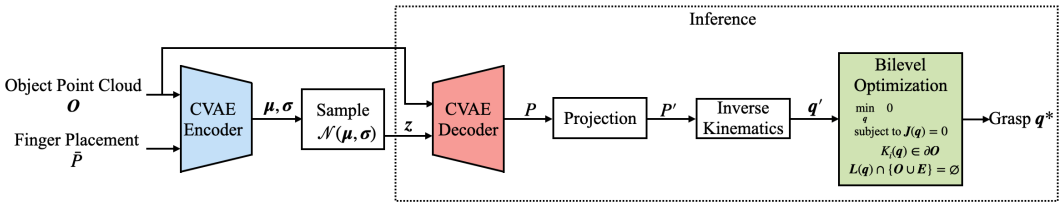


Figure 2: Overview of our method.

Our method consists of a learned model that predicts a plausible finger placement, and an analytical process that computes a physically feasible grasp configuration guided by the finger placement. Figure 2 gives an overview of our method. Given an object $\boldsymbol{O}$, a trained CVAE decoder predicts an initial finger placement $\mathcal{P}$, which is not guaranteed to be physically feasible or lie on the surface of the object. Subsequently, a nonconvex bilevel optimization computes a valid grasping pose $\boldsymbol{q}^*$ by projecting $\mathcal{P}$ onto the surface of the object and enforcing physical constraints.

3.1 Physically Feasible Dexterous Grasp

Before we describe the two main components of our method, we first formally define the criteria of a *physically feasible dexterous grasp*. Given an object $\boldsymbol{O}$ in environment $\boldsymbol{E}$, a grasp configuration $\boldsymbol{q}$, and a desired finger placement $\mathcal{P} = (\boldsymbol{p}_1, \boldsymbol{p}_2, \boldsymbol{p}_3)$, we consider $(\boldsymbol{O}, \boldsymbol{E}, \boldsymbol{q}, \mathcal{P})$ physically feasible if they satisfy following dynamic and kinematic constraints. In our setup, $\boldsymbol{q} \in \mathbb{R}^{22}$ and $\boldsymbol{E}$ is the fixed tabletop (cf. Section 4.2).

3.1.1 Dynamic Constraints: Force Closure and Friction Cone

For a static dexterous grasp, the dynamic constraints include the force closure constraint and the friction cone constraint. The force closure constraint requires the sum of the external wrench from all contact points to be zero,

$$\sum_{i=1}^{3} \boldsymbol{f}_i = \boldsymbol{0} \quad \text{and} \quad \sum_{i=1}^{3} \boldsymbol{p}_i \times \boldsymbol{f}_i = \boldsymbol{0}, \tag{1}$$

where $\boldsymbol{f}_i$ is the unknown contact force applied at position $\boldsymbol{p}_i$ from the dexterous hand to the object.

Given the static friction coefficient μ and the surface normal direction (pointing away from the object) of each contact point $\hat{\boldsymbol{n}}_i$ from $\boldsymbol{O}$, the friction cone constraint requires the normal force to be nonnegative and the contact force to be within the friction cone.

$$0 \leq -\boldsymbol{f}_i \cdot \hat{\boldsymbol{n}}_i \quad \text{and} \quad \|\boldsymbol{f}_i - (\boldsymbol{f}_i \cdot \hat{\boldsymbol{n}}_i)\hat{\boldsymbol{n}}_i\|_2 \leq -\mu \boldsymbol{f}_i \cdot \hat{\boldsymbol{n}}_i. \tag{2}$$

The friction force direction may be approximated with a polyhedral cone spanned by a set of unit vectors perpendicular to $\hat{\boldsymbol{n}}_i$: $\left\{\hat{\boldsymbol{t}}_{i,j} \mid \hat{\boldsymbol{t}}_{i,j} \perp \hat{\boldsymbol{n}}_i, j \in \{1, 2, \dots\}\right\}$ ([31]). In this work, we chose an arbitrary orthogonal basis on the tangent plane of the contact point $(\hat{\boldsymbol{t}}_{i,1}, \hat{\boldsymbol{t}}_{i,2})$, which yields the following approximated friction cone constraints

$$0 \leq -\boldsymbol{f}_i \cdot \hat{\boldsymbol{n}}_i \quad \text{and} \quad |\boldsymbol{f}_i \cdot \hat{\boldsymbol{t}}_{i,j}| \leq -\mu \boldsymbol{f}_i \cdot \hat{\boldsymbol{n}}_i, \forall j \in \{1, 2\}. \tag{3}$$

A grasp is *dynamically feasible* if it satisfies both force closure and friction cone constraints. Leveraging the polyhedral approximation, dynamic feasibility can be cast as a quadratic program (QP):

$$\begin{aligned} \min_{\boldsymbol{f}_1,\boldsymbol{f}_2,\boldsymbol{f}_3} \quad & \|\sum_{i=1}^{3} \boldsymbol{f}_i\|_2^2 + \|\sum_{i=1}^{3} \boldsymbol{p}_i \times \boldsymbol{f}_i\|_2^2 \\ \text{subject to} \quad & 0 < f_{min} \leq -\boldsymbol{f}_i \cdot \hat{\boldsymbol{n}}_i, \;\; |\boldsymbol{f}_i \cdot \hat{\boldsymbol{t}}_{i,j}| \leq -\mu \boldsymbol{f}_i \cdot \hat{\boldsymbol{n}}_i, \; \forall i \in \{1,2,3\}, \forall j \in \{1,2\}. \end{aligned} \tag{4}$$

The optimization in Equation (4) has objective value 0 if and only if the grasp is dynamically feasible. Note that we set an arbitrary lower bound $f_{min} \in \mathbb{R}^+$ on the normal force to avoid the trivial solution of $\boldsymbol{f}_i = \boldsymbol{0}$. The exact value of f_{min} is irrelevant because changing f_{min} represents scaling the optimal contact forces, which does not affect their membership in the friction cone.

3.1.2 Kinematic Constraints: Reachability and Collision

The kinematic constraints include reachability constraints and collision constraints. The reachability constraint enforces the kinematic ability of the robot hand to reach the contact point $\boldsymbol{p}_i$ on the object surface $\partial \boldsymbol{O}$: $\exists \boldsymbol{q} : K_i(\boldsymbol{q}) = \boldsymbol{p}_i \in \partial \boldsymbol{O}, \; \forall i \in \{1,2,3\}$, where $K_i : \mathbb{R}^{22} \mapsto \mathbb{R}^3$ is the forward kinematics function that maps the hand state $\boldsymbol{q}$ to the position of the i-th fingertip. Assuming only fingertip contacts, collision constraints ensure that no robot link $\boldsymbol{L}(\boldsymbol{q})$ is in collision with the object $\boldsymbol{O}$ except at the fingertip or with the environment $\boldsymbol{E}$, i.e. $\boldsymbol{L}(\boldsymbol{q}) \cap \{\boldsymbol{O} \cap \boldsymbol{E}\} = \emptyset$.

Both reachability and collision constraints are nonconvex constraints. Our implementation handles these constraints using the nonlinear optimizer SNOPT [32] and software provided by Drake [33].

3.2 Learning to Predict Finger Placement

We take a machine learning approach to predict a finger placement $\mathcal{P}$ given an arbitrary object observed as a point cloud $\hat{\boldsymbol{O}}$. An immediate challenge faced by the learning approach is the lack of large-scale datasets for dexterous robotic hands. While datasets for real human hands do exist, retargeting human grasping poses to a robot hand with different kinematic structures and joint limits presents numerous challenges. We opt to synthesize a large-scale dataset of physically feasible grasping poses for the Allegro robot hand based on a subset of YCB objects [34]. Our dataset is publicly available and can facilitate other research efforts in the field of dexterous manipulation.

3.2.1 Creating a Physically Feasible Dexterous Grasping Dataset

We generate physically feasible grasps of six YCB objects [34] in simulation. The chosen objects are "cracker box," "sugar box," "tomato soup can," "mustard bottle," "gelatin box," and "potted meat can." We consider a realistic scenario where the object is placed on a flat surface instead of free floating. As such, we need to consider the variations due to different rest poses of on the surface in addition to finger placements. The following summarizes the steps for data generation:

1. Simulate dropping the object on a flat surface from different initial poses to find "rest poses".
2. For each rest pose, produce 256 candidate contact points distributed across the object with Poisson disk sampling.
3. Remove any point in contact with the surface $\boldsymbol{E}$ and store the remaining as a point cloud $\hat{\boldsymbol{O}}$.
4. Evaluate Equation (4) for each of the C_3^{256} 3-point finger placements. Discard if it is not dynamically feasible.
5. For each dynamically feasible finger placement, permute the $3! = 6$ possible assignments for the thumb, index finger, and middle finger. Check each assignment for kinematic feasibility. If a feasible grasping pose $\boldsymbol{q}$ can be found for a finger placement $\bar{\mathcal{P}}$, add $(\hat{\boldsymbol{O}}, \boldsymbol{q}, \bar{\mathcal{P}})$ to the dataset $\mathbb{P}$.
6. At training time, augment the dataset by randomly translating and varying the yaw angle for each grasp in the dataset.

3.2.2 Training a Conditional Variational Autoencoder (CVAE)

To compute diverse grasps for arbitrary objects directly from point cloud observations, we predict fingertip contact points on the object surface with a conditional variational autoencoder

(CVAE) [35]. Our model is adapted from [7] and consists of an encoder E and a decoder D that are based on the PointNet++ architecture [36]. The model ultimately seeks to maximize the likelihood of producing a set of contact points $\mathcal{P}$ deemed feasible in Section 3.2.1 given the point cloud $\hat{\boldsymbol{O}}$.

The encoder $E(\boldsymbol{z} \mid \bar{\mathcal{P}}, \hat{\boldsymbol{O}})$ maps a grasp $\bar{\mathcal{P}}$ and a point cloud $\hat{\boldsymbol{O}}$ to the latent space. We assume the latent variable has a normal distribution $\mathcal{N}(\mathbf{0}, \mathcal{I})$. Meanwhile, given a latent sample $\boldsymbol{z} \sim E$, the decoder attempts to reconstruct the finger placement $\mathcal{P}$. We seek to minimize the element-wise L^1-norm reconstruction loss $L_{rec}(\mathcal{P}, \bar{\mathcal{P}}) \triangleq \|\mathcal{P} - \bar{\mathcal{P}}\|_1$ for a feasible grasp from the dataset $\bar{\mathcal{P}} \in \mathbb{P}$. Additionally, a KL divergence loss $\mathcal{D}_{KL}$ is applied on the latent distribution $E(\cdot \mid \cdot)$ to ensure a normally distributed latent space. The complete loss function of the network is defined as

$$L \triangleq \sum_{\boldsymbol{z} \sim E, \bar{\mathcal{P}} \in \mathbb{P}} L_{rec}(\mathcal{P}, \bar{\mathcal{P}}) + \alpha \mathcal{D}_{KL} \left(E(\boldsymbol{z} \mid \bar{\mathcal{P}}, \hat{\boldsymbol{O}}) \,\middle|\middle|\, \mathcal{N}(\mathbf{0}, \mathcal{I}) \right). \tag{5}$$

At inference time, E is removed and a latent sample $\boldsymbol{z}$ is drawn from $\mathcal{N}(\mathbf{0}, \mathcal{I})$. This is passed into the decoder $D(\mathcal{P} | \hat{\boldsymbol{O}}, \boldsymbol{z})$ along with the point cloud $\hat{\boldsymbol{O}}$ to produce the grasp point prediction $\mathcal{P}$.

3.3 Computing Grasping Pose using Differentiable Bilevel Optimization

While the grasps in the training dataset are physically feasible by construction, there is no guarantee that the CVAE output, $\mathcal{P}$, is physically feasible. In fact, it is very likely that $\mathcal{P}$ is not even in contact with the object surface $\partial \boldsymbol{O}$, i.e. $|\mathcal{P} \cap \partial \boldsymbol{O}| < 3$. However, computing the projection of $\mathcal{P}$ onto the manifold of physically feasible grasps is very challenging due to the complexity of $\partial \boldsymbol{O}$ and the physical constraints described in Section 3.1.

To solve this projection, we propose a nonconvex bilevel optimization to obtain a physically feasible grasping pose $\boldsymbol{q}$ given $\mathcal{P}$, the output from the CVAE. First, we obtain an Euclidean projection of $\mathcal{P}$ onto $\partial \boldsymbol{O}$, denoted as $\mathcal{P}'$:

$$\mathcal{P}' \triangleq \left\{ \boldsymbol{p}_i' \;\middle|\; \boldsymbol{p}_i' = \underset{\boldsymbol{p}_o \in \partial \boldsymbol{O}}{\arg\min} \|\boldsymbol{p}_o - \boldsymbol{p}_i\|_2, \; i \in \{1, 2, 3\} \right\}. \tag{6}$$

Next, we solve an IK problem for a grasp configuration finger placement problem specified by $\mathcal{P}'$ up to a numerical tolerance ϵ.

$$\text{Find } \boldsymbol{q}' : \|K_i(\boldsymbol{q}') - \boldsymbol{p}_i'\|_2 \leq \epsilon, \; K_i(\boldsymbol{q}') \in \partial \boldsymbol{O}, \; \forall i \in \{1, 2, 3\}. \tag{7}$$

$\boldsymbol{q}'$ is used as an initial guess for the bilevel optimization in Section 3.3.1.

3.3.1 Differentiable Bilevel Optimization Formulation

While it is possible to construct a large nonconvex optimization that jointly solves for grasp pose and the contact forces,

$$\begin{aligned} \min_{\boldsymbol{q}, \boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3} \quad & 0 \\ \text{subject to} \quad & K_i(\boldsymbol{q}) \in \partial \boldsymbol{O}, \; \boldsymbol{L}(\boldsymbol{q}) \cap \{\boldsymbol{O} \cup \boldsymbol{E}\} = \emptyset, \; \|\sum_{i=1}^{3} \boldsymbol{f}_i\|_2 = 0, \; \|\sum_{i=1}^{3} K_i(\boldsymbol{q}) \times \boldsymbol{f}_i\|_2 = 0, \\ & f_{min} \leq -\boldsymbol{f}_i \cdot \hat{\boldsymbol{n}}_i, \; |\boldsymbol{f}_i \cdot \hat{\boldsymbol{t}}_{i,j}| \leq -\mu \boldsymbol{f}_i \cdot \hat{\boldsymbol{n}}_i, \; \forall i \in \{1, 2, 3\}, \forall j \in \{1, 2\}, \end{aligned} \tag{8}$$

such an optimization does not solve well in practice due to its complexity. Additionally, the initial choice for $\boldsymbol{f}$ is not obvious. Likewise, relaxing convex constraints of Equation (8), e.g. replacing the objective with $\min_{\boldsymbol{q}, \boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3} \; \|\sum_{i=1}^{3} \boldsymbol{f}_i\|_2^2 + \|\sum_{i=1}^{3} K_i(\boldsymbol{q}) \times \boldsymbol{f}_i\|_2^2$ cannot guarantee the contact forces are solved to optimality for determining force closure possibility of a chosen $\boldsymbol{q}$.

We propose leveraging *bilevel optimization* [25, 28] to offload the dynamic feasibility computation to a lower-level QP. Define $J(\boldsymbol{q})$ as the minimum objective value of the lower-level problem, which is the dynamic constraint QP in Equation (4):

$$\begin{aligned} J(\boldsymbol{q}) \triangleq \min_{\boldsymbol{f}_1, \boldsymbol{f}_2, \boldsymbol{f}_3} \quad & \|\sum_{i=1}^{3} \boldsymbol{f}_i\|_2^2 + \|\sum_{i=1}^{3} K_i(\boldsymbol{q}) \times \boldsymbol{f}_i\|_2^2 \\ \text{subject to} \quad & f_{min} \leq -\boldsymbol{f}_i \cdot \hat{\boldsymbol{n}}_i, \; |\boldsymbol{f}_i \cdot \hat{\boldsymbol{t}}_{i,j}| \leq -\mu \boldsymbol{f}_i \cdot \hat{\boldsymbol{n}}_i, \; \forall i \in \{1, 2, 3\}, \forall j \in \{1, 2\}. \end{aligned} \tag{9}$$

Table 1: Seen and adversarial objects.

Method	Seen Objects				Adversarial Objects				
	Sugar Box	Soup Can	Mustard	Overall	Tetrahedron	Sandwich Box	Massage Ball	Castle	Overall
Ours	83.33%	50.00%	66.67%	66.67%	0.00%	100.00%	66.67%	100.00%	66.67%
CVAE only	33.33%	33.33%	16.67%	27.78%	0.00%	66.67%	0.00%	33.33%	25.00%
Hardcoded	100.00%	83.33%	100.00%	94.44%	0.00%	100.00%	100.00%	0.00%	50.00%

Table 2: Familiar and novel objects.

	Familiar Objects					Novel Objects			
	Webcam	Mask	Chips	Soda	Overall	Glasses	Ramen	Pill Bottle	Overall
Ours	50.00%	66.67%	66.67%	100.00%	70.83%	66.67%	66.67%	33.33%	55.56%

Abstracting away $J(\boldsymbol{q})$ in Equation (8) yields the upper-level problem:

$$\begin{aligned} \min_{\boldsymbol{q}} \quad & 0 \\ \text{subject to} \quad & J(\boldsymbol{q}) = 0,\ K_i(\boldsymbol{q}) \in \partial \boldsymbol{O},\ \boldsymbol{L}(\boldsymbol{q}) \cap \{\boldsymbol{O} \cup \boldsymbol{E}\} = \emptyset. \end{aligned} \tag{10}$$

We denote the solution to Equation (10) as $\boldsymbol{q}^*$. Equation (9) is a constrained QP, it can be solved to optimality numerically with standard convex optimizers. This implies Equation (4) is evaluated exactly with no conservatism in our proposed approach. Moreover, there is no need to provide a good initialization to $\boldsymbol{f}_i$.

To solve Equation (10) with a nonconvex optimizer (e.g. SNOPT [32]), one needs to obtain the gradient of each of the constraints with respect to $\boldsymbol{q}$. The gradient of the kinematic constraints are available from standard IK solvers (e.g. Drake [33]). The gradient of $J(\boldsymbol{q})$, $\nabla_{\boldsymbol{q}} J(\boldsymbol{q})$, may be obtained using recent advances in differentiable optimization [37]. At a high level, differentiable optimization computes $\nabla_{\boldsymbol{q}} J(\boldsymbol{q})$ by taking the matrix differentials of the KKT conditions of the optimization problem at its solution.

We emphasize that the bilevel optimization in Equation (10) is still a complicated nonconvex optimization, and the optimizer may return a locally optimal $\boldsymbol{q}^*$ in the vicinity of $\boldsymbol{q}'$. Intuitively, this could be interpreted as optimizing within the grasp family of the CVAE prediction (e.g. "left side grasp" or "top-down grasp"), which implies the necessity of CVAE in the process. This is supported empirically by our ablation studies in Section 4. In practice, the bilevel optimization provides a certification of physical feasibility for the CVAE prediction. A poor $\boldsymbol{q}'$ choice will likely results in the nonconvex optimizer returning infeasibility. This serves as the condition to reject $\mathcal{P}$ and generate a different CVAE grasp prediction with another latent variable sample.

4 Evaluation

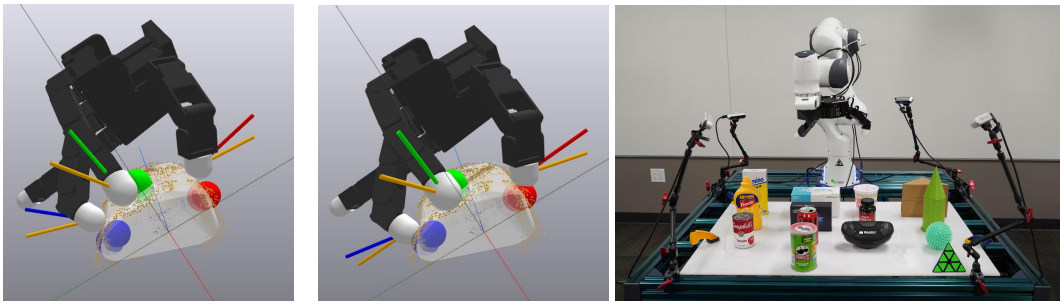
4.1 Pipeline Evaluation

Our trained CVAE achieved a test reconstruction error of approximately 0.5cm, which is sufficient as the precise grasp can be determined by bilevel optimization. More training details are available in the supplementary material.

The effects of applying bilevel optimization is shown in Figure 3. The fingertip placement prior to bilevel optimization clearly cannot form force closure as $\hat{\boldsymbol{n}}_i \cdot \hat{\boldsymbol{n}}_j > 0, \forall i, j \in \{1, 2, 3\}$. Applying bilevel optimization shifted the thumb and middle fingers to an antipodal configuration, which allows for force closure through pinching the two fingers. This behavior validates our approach of projecting a physically infeasible predicted grasp onto the manifold of physically feasible grasps.

4.2 Hardware Setup

The hardware setup and evaluated objects are shown in Figure 3c.



(a) Before optimization ($\boldsymbol{q}'$). (b) After optimization ($\boldsymbol{q}^*$). (c) Hardware setup.

Figure 3: **3a and 3b: Effects of bilevel optimization**, on a bottom-up view of a mustard bottle grasp. Using the color encoding of red: thumb, green: index, and blue: middle, we show $\mathcal{P}$ (solid spheres), $\mathcal{P}'$ (transparent spheres) and $\hat{\boldsymbol{n}}_i$ (colored lines). The direction of the computed contact forces $\boldsymbol{f}_i$ are shown at the respective fingertip with yellow lines. The mismatch between $K(\boldsymbol{q}')$ and $\mathcal{P}'$ is due to the numerical tolerance ϵ. **3c: Hardware setup.** See Section 4.2 for more details.

Hardware. A 16-DoF Allegro v4.0 right hand was used for hardware grasping experiments. The Allegro hand was mounted on a 7-DoF Franka Emika Panda arm to realize the planned wrist pose. The object point cloud was captured by four fixed Intel RealSense D435 depth cameras. The implementation details are available in the supplementary material.

Evaluated Objects. We evaluated our method on 13 household objects resting on the table, each with either one or two rest poses. In total we evaluated 21 object-pose configurations. We categorize our objects into 4 sets. In Figure 3c, each column (left to right) contains an objects set, listed in the order below. Objects within a set are ordered from near to far.

1. Seen objects (2 rest poses). Mustard bottle, soup can, and sugar box from the training set.

2. Familiar objects (2 rest poses). Unseen objects with geometries similar to training objects. Includes boxes (webcam box, mask box) and cylinders (chip can, soda can).

3. Novel objects (1 rest pose). Objects whose geometries are distinct from that of any object in the training set. Includes glasses case, ramen, and pill bottle.

4. Adversarial objects (1 rest pose). A set of objects curated to be challenging or even nearly impossible to grasp when the object is instantiated in certain poses. Includes tetrahedron, sandwich box, massage ball, and castle.

4.3 Evaluation Procedure and Metric

At the start of each trial, the object is placed in a specified pose in front of the robot. After observing the point cloud and computing a grasp, the Franka arm first brings the hand to the desired wrist pose using an *ad hoc* trajectory planner. The finger joint angles from $\boldsymbol{q}^*$ is then executed on the Allegro hand. A fixed, hand-designed squeezing motion is performed after $\boldsymbol{q}^*$ has been achieved to provide contact forces. The Franka then attempts to lift the object to a fixed height approximately 43cm above the table. A trial is considered successful if the object is lifted and all three fingers remain in contact with the object. All trials were repeated for 3 times.

4.4 Baselines and Ablation Studies

We compare our method against the following approaches. As bilevel optimization seldom returns a solution if seeded with a kinematically infeasible solution (e.g. open hand), we excluded "bilevel optimization only" in our ablation studies.

Hardcoded grasp baseline. Using the set of seen objects, we designed a top-down grasp policy that chooses the wrist pose based on the point cloud and executes a fixed grasp.

CVAE-only as an ablation study. We executed $\boldsymbol{q}'$ obtained from CVAE directly without bilevel optimization.

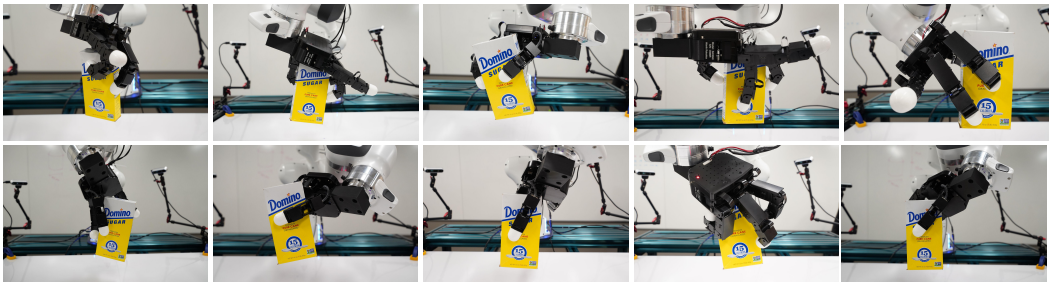


Figure 4: Diverse grasps generated by our method. Each image depicts a unique grasp generated from different sampled latent variables. All grasps were computed for the same initial object rest pose.

4.5 Results

Seen object results are summarized in Table 1. Our method significantly outperforms the CVAE-only ablation, which often generates kinematically feasible but dynamically infeasible grasps. While our hardcoded baseline is capable of achieving near-perfect success rate on seen objects, we emphasize that it is specifically tuned for this object set and is expected to perform well.

Familiar object results are shown in Table 2. Our method's success rate on familiar objects (71%) is comparable to our method's success rate on seen objects (67%). This suggests that our method is capable of generalizing to unseen objects in familiar geometry categories.

Novel object results are shown in Table 2. Our method's success rate only drops slightly (56%) on objects with unseen geometry, which shows its ability to generalize to novel objects.

Adversarial object results are shown in Table 1. The tetrahedron is difficult for any finger-based gripper due to its polyhedral shape. The sandwich box is challenging to parallel jaw grippers as it is triangular. The massage ball is challenging to suction-based grippers as the surface is highly irregular. The castle object requires a side grasp due to its cone-shaped top. Our method demonstrated the merits of diverse dexterous grasps by achieving high grasping success rates on all objects except for the tetrahedron. Note that while our hardcoded baseline nearly always succeed in objects that allow top-down grasps, it fails on objects that require a different grasp type.

Generating diverse grasps is possible with our method through sampling different latent $\boldsymbol{z}$'s. Figure 4 shows 10 distinct successful grasps performed on the upright sugar box.

5 Limitations and Conclusion

Limitations. Our work has a number of limitations. Firstly, our method requires an observation of the full object point cloud. As our choice of physical constraint formulation does not include explicit robustness margins, estimation errors from the vision pipeline is a major source of grasp failures. This may be addressed by introducing point cloud completion (e.g. [38, 39, 40]) and grasp robustness metrics (e.g. [12]) into the pipeline. Secondly, it is currently not possible to specify which of the grasps to generate (e.g. "grasp the object from the left side") with our model. Finally, our method currently assumes that the object to grasp is placed on a flat surface and there are no other objects in the scene.

Conclusion. This work presents a novel pipeline that combines learning-based grasp generation with bilevel optimization for producing diverse and feasible dexterous grasps. Through a comprehensive suite of hardware evaluation, we validated our method on unseen geometries and adversarial objects challenging to non-dexterous grippers. Our integrative approach of projecting learned predictions onto a lower dimensional physically-feasible manifold serves as an example of leveraging model-based techniques to improve learning performance in the contact-rich setting, which can potentially be applied to other robotic manipulation tasks.

References

[1] G. Du, K. Wang, S. Lian, and K. Zhao. Vision-based robotic grasping from object localization, object pose estimation to grasp estimation for parallel grippers: a review. *Artificial Intelligence Review*, 54(3):1677–1734, 2021.

[2] M. A. Roa and R. Suárez. Grasp quality measures: review and performance. *Autonomous robots*, 38(1):65–88, 2015.

[3] J. Bohg, A. Morales, T. Asfour, and D. Kragic. Data-driven grasp synthesis—a survey. *IEEE Transactions on robotics*, 30(2):289–309, 2013.

[4] S. Caldera, A. Rassau, and D. Chai. Review of deep learning methods in robotic grasp detection. *Multimodal Technologies and Interaction*, 2(3):57, 2018.

[5] D. Berenson, R. Diankov, K. Nishiwaki, S. Kagami, and J. Kuffner. Grasp planning in complex scenes. In *2007 7th IEEE-RAS International Conference on Humanoid Robots*, pages 42–48. IEEE, 2007.

[6] A. Saxena, J. Driemeyer, J. Kearns, and A. Ng. Robotic grasping of novel objects. *Advances in neural information processing systems*, 19, 2006.

[7] A. Mousavian, C. Eppner, and D. Fox. 6-dof graspnet: Variational grasp generation for object manipulation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2901–2910, 2019.

[8] 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.

[9] K. Kleeberger, R. Bormann, W. Kraus, and M. F. Huber. A survey on learning-based robotic grasping. *Current Robotics Reports*, 1(4):239–249, 2020.

[10] M. T. Ciocarlie and P. K. Allen. Hand posture subspaces for dexterous robotic grasping. *The International Journal of Robotics Research*, 28(7):851–867, 2009.

[11] F. T. Pokorny and D. Kragic. Classical grasp quality evaluation: New algorithms and theory. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3493–3500. IEEE, 2013.

[12] C. Ferrari and J. F. Canny. Planning optimal grasps. In *ICRA*, volume 3, page 6, 1992.

[13] J. Lundell, F. Verdoja, and V. Kyrki. Ddgc: Generative deep dexterous grasping in clutter. *IEEE Robotics and Automation Letters*, 6(4):6899–6906, 2021.

[14] L. Shao, F. Ferreira, M. Jorda, V. Nambiar, J. Luo, E. Solowjow, J. A. Ojea, O. Khatib, and J. Bohg. Unigrasp: Learning a unified model to grasp with multifingered robotic hands. *IEEE Robotics and Automation Letters*, 5(2):2286–2293, 2020.

[15] Q. Lu, K. Chenna, B. Sundaralingam, and T. Hermans. Planning multi-fingered grasps as probabilistic inference in a learned deep network. In *Robotics Research*, pages 455–472. Springer, 2020.

[16] P. Mandikal and K. Grauman. Learning dexterous grasping with object-centric visual affordances. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6169–6176. IEEE, 2021.

[17] S. Pfrommer, M. Halm, and M. Posa. Contactnets: Learning discontinuous contact dynamics with smooth, implicit representations. *arXiv preprint arXiv:2009.11193*, 2020.

[18] Y. Jiang, J. Sun, and C. K. Liu. Data-augmented contact model for rigid body simulation. In *Learning for Dynamics and Control Conference*, pages 378–390. PMLR, 2022.

[19] T. Zhu, R. Wu, X. Lin, and Y. Sun. Toward human-like grasp: Dexterous grasping via semantic representation of object-hand. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15741–15751, 2021.

[20] S. Brahmbhatt, C. Tang, C. D. Twigg, C. C. Kemp, and J. Hays. Contactpose: A dataset of grasps with object contact and hand pose. In *European Conference on Computer Vision*, pages 361–378. Springer, 2020.

[21] O. Taheri, N. Ghorbani, M. J. Black, and D. Tzionas. Grab: A dataset of whole-body human grasping of objects. In *European conference on computer vision*, pages 581–600. Springer, 2020.

[22] H. Jiang, S. Liu, J. Wang, and X. Wang. Hand-object contact consistency reasoning for human grasps generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11107–11116, 2021.

[23] L. Yang, X. Zhan, K. Li, W. Xu, J. Li, and C. Lu. Cpf: Learning a contact potential field to model the hand-object interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11097–11106, 2021.

[24] E. Corona, A. Pumarola, G. Alenya, F. Moreno-Noguer, and G. Rogez. Ganhand: Predicting human grasp affordances in multi-object scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5031–5041, 2020.

[25] A. Sinha, P. Malo, and K. Deb. A review on bilevel optimization: from classical to evolutionary approaches and applications. *IEEE Transactions on Evolutionary Computation*, 22(2):276–295, 2017.

[26] S. Zimmermann, G. Hakimifard, M. Zamora, R. Poranne, and S. Coros. A multi-level optimization framework for simultaneous grasping and motion planning. *IEEE Robotics and Automation Letters*, 5(2):2966–2972, 2020.

[27] Y. Zhu, Z. Pan, and K. Hauser. Contact-implicit trajectory optimization with learned deformable contacts using bilevel optimization. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9921–9927. IEEE, 2021.

[28] B. Landry, J. Lorenzetti, Z. Manchester, and M. Pavone. Bilevel optimization for planning through contact: A semidirect method. *arXiv preprint arXiv:1906.04292*, 2019.

[29] T. Stouraitis, I. Chatzinikolaidis, M. Gienger, and S. Vijayakumar. Online hybrid motion planning for dyadic collaborative manipulation via bilevel optimization. *IEEE Transactions on Robotics*, 36(5):1452–1471, 2020.

[30] Z. Zhao, Z. Zhou, M. Park, and Y. Zhao. Sydebo: Symbolic-decision-embedded bilevel optimization for long-horizon manipulation in dynamic environments. *IEEE Access*, 9:128817–128826, 2021.

[31] D. E. Stewart and J. C. Trinkle. An implicit time-stepping scheme for rigid body dynamics with inelastic collisions and coulomb friction. *International Journal for Numerical Methods in Engineering*, 39(15):2673–2691, 1996.

[32] P. E. Gill, W. Murray, and M. A. Saunders. Snopt: An sqp algorithm for large-scale constrained optimization. *SIAM review*, 47(1):99–131, 2005.

[33] R. Tedrake and the Drake Development Team. Drake: Model-based design and verification for robotics, 2019. URL `https://drake.mit.edu`.

[34] B. Calli, A. Singh, J. Bruce, A. Walsman, K. Konolige, S. Srinivasa, P. Abbeel, and A. M. Dollar. Yale-cmu-berkeley dataset for robotic manipulation research. *The International Journal of Robotics Research*, 36(3):261–268, 2017.

[35] D. P. Kingma and M. Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[36] 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*, 30, 2017.

[37] B. Amos and J. Z. Kolter. Optnet: Differentiable optimization as a layer in neural networks. In *International Conference on Machine Learning*, pages 136–145. PMLR, 2017.

[38] W. Yuan, T. Khot, D. Held, C. Mertz, and M. Hebert. Pcn: Point completion network. In *2018 International Conference on 3D Vision (3DV)*, pages 728–737. IEEE, 2018.

[39] M. Sarmad, H. J. Lee, and Y. M. Kim. Rl-gan-net: A reinforcement learning agent controlled gan network for real-time point cloud shape completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5898–5907, 2019.

[40] X. Han, Z. Li, H. Huang, E. Kalogerakis, and Y. Yu. High-resolution shape completion using deep neural networks for global structure and local geometry inference. In *Proceedings of the IEEE international conference on computer vision*, pages 85–93, 2017.