Robotics: Science and Systems 2023
Daegu, Republic of Korea, July 10-July 14, 2023

CHSEL: Producing Diverse Plausible Pose Estimates from Contact and Free Space Data

Sheng Zhong, Nima Fazeli, Dmitry Berenson
University of Michigan, Ann Arbor, MI 48109
Email: {zhsh, nfz, dmitryb}@umich.edu

***Abstract*—This paper proposes a novel method for estimating the set of plausible poses of a rigid object from a set of points with volumetric information, such as whether each point is in free space or on the surface of the object. In particular, we study how pose can be estimated from force and tactile data arising from contact. Using data derived from contact is challenging because it is inherently less information-dense than visual data, and thus the pose estimation problem is severely under-constrained when there are few contacts. Rather than attempting to estimate the true pose of the object, which is not tractable without a large number of contacts, we seek to estimate a plausible set of poses which obey the constraints imposed by the sensor data. Existing methods struggle to estimate this set because they are either designed for single pose estimates or require informative priors to be effective. Our approach to this problem, Constrained pose Hypothesis Set Elimination (CHSEL), has three key attributes: 1) It considers volumetric information, which allows us to account for known free space; 2) It uses a novel differentiable volumetric cost function to take advantage of powerful gradient-based optimization tools; and 3) It uses methods from the Quality Diversity (QD) optimization literature to produce a diverse set of high-quality poses. To our knowledge, QD methods have not been used previously for pose registration. We also show how to update our plausible pose estimates online as more data is gathered by the robot. Our experiments suggest that CHSEL shows large performance improvements over several baseline methods for both simulated and real-world data.** [1]

I. Introduction

Pose registration—the process of estimating the pose of a given rigid object from sensor data, is a fundamental problem in robotics, as it is necessary for manipulation and reasoning. Much research has been done in estimating object pose from visual data, especially laser-range data [6] [5]. However, a clear view of the object may not always be available (e.g. an object in a cupboard, as in Fig. 1, or grocery bag) or the material properties of the object may make it difficult to perceive visually (e.g. transparency).

Partial occlusion in manipulation tasks motivates for rummaging, and researchers have investigated the use of tactile and force feedback for pose registration [27] [8]. However, the nature of this data is quite different from the point-clouds produced by laser-scanners. While point-cloud data is information-dense (e.g. many points on the surface of the object), tactile and force data arising from contact contain much less information (e.g. one contact point per motion) in addition to being time-consuming to collect. This lack of information can be partially mitigated by assuming that a contact sensor moves along the surface of the object [29, 9].

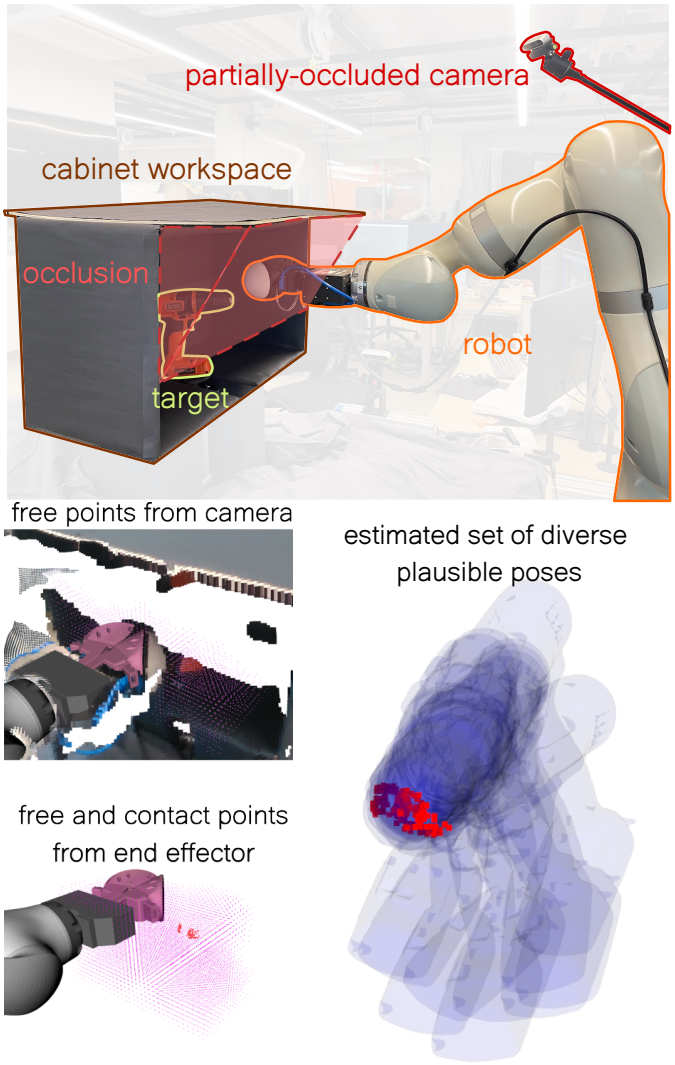

Fig. 1: Top: Set up of a real-world probing experiment where the goal is to estimate the drill's pose. Bottom, left: input to CHSEL, made of known free points in pink (from the camera and swept robot volume) and known surface points in red (from contact). Bottom, right: CHSEL uses these points and the object model to estimate a diverse set of plausible poses.

[1]This work was supported in part by Toyota Research Institute, the Office of Naval Research Grant N00014-21-1-2118 and NSF grants IIS-1750489, IIS-2113401, and IIS-2220876. For code, see https://github.com/UM-ARM-Lab/chsel

However, creating controllers that can do this without moving the object is challenging.

In the context of pose registration problems, the lack of informative data results in a lack of constraints on the set of plausible poses of the object. In such cases, producing an accurate estimate of the true pose is very unlikely, and it is more useful to estimate the set of plausible poses. Especially towards the beginning of contact-based tasks, uncertainty in the object pose is high due to insufficient data, sensor noise, and inherent object symmetries. Characterizing this uncertainty, such as in the form of a set of plausible poses, is useful for object recognition [33], active perception [10], and simultaneous localization and mapping (SLAM) [1] [13].

The most common methods for pose registration are based on the Iterative Closest Point (ICP) algorithm [26] [31], which outputs a single pose estimate for a given initial pose. These methods can be effective for point-cloud data, but producing a set of estimates from random initialization does not yield good coverage of the set of plausible poses for contact data. Bayesian methods that aim to capture the full distribution of plausible object poses use approximation techniques such as Markov Chain Monte Carlo (MCMC) and variational inference [21]. However, such variational methods depend heavily on informative priors, which we do not assume are available.

To overcome the above limitations, we present Constrained pose Hypothesis Set Elimination (**CHSEL**), which has three key attributes: First, we go beyond only considering points on the surface of the object, considering volumetric information instead (similar to Slavcheva et al. [28] and Haugo and Stahl [15]). This allows us to infer more data (and thus more constraints on the pose) from robot motion. For example, when a robot moves into contact with an object, we observe contact points, as well as all the free space the robot traversed before and during contact. Note that this representation can also include free space and object surface points observed by a visual sensor.

Second, to take advantage of powerful gradient-based optimization tools, we construct a differentiable cost function that can be used to efficiently optimize a given pose based on volumetric information. Finally, and most importantly, to estimate a diverse set of poses simultaneously, we adapt methods from the Quality Diversity (QD) optimization literature. To our knowledge, this work is the first application of QD methods to the problem of pose registration. QD methods explicitly optimize for a set of solutions which are both diverse and high-quality, making them a natural choice for pose registration problems that seek to capture the set of plausible poses. We also show how to update our set of estimates online as more data is gathered by the robot.

Our experiments suggest that CHSEL has large performance improvements over several baseline methods for both simulated and real-world contact data. Additionally, we compare against alternatives and show that our cost function is a good QD objective. We also show that real-world visual data can be incorporated seamlessly into our cost function while demonstrating similar performance improvements.

II. Related Work

While our work is the first to use known free space to produce a diverse set of pose registrations, prior work has been done separately in using free space in registration and diverse set (also known as multi-hypothesis) registration. Geometric registration has been extensively studied in robotics and computer vision (see Tam et al. [30] for an overview). In particular, the distinction between free space and surface points can be framed as point semantics or features, and methods such as the 3D Normal Distribution Transform (3D NDT) [20] and its continuous generalization Continuous Visual Odometry (CVO) [34] have been designed with them in mind. We compare against CVO as a baseline. Haugo and Stahl [15] considers free space explicitly, filling it with balls via the medial axis transformation. They then formulate a cost penalizing object-ball penetration while requiring points to lie on the surface. This is a baseline in our experiments.

Specific to SE(2) pose estimation in planar contact problems, the Manifold Particle Filter [16] exploits a robot's contact manifold to estimate an pose. However, it struggles to scale to full SE(3) pose estimation as it is expected to require exponentially more particles.

Deep learning based methods such as SegICP [32] and MHPE [13] have been developed to produce a plausible set of pose estimates. However, they can only use points from the object surface and require relatively dense information.

Related to registration is the problem of object reconstruction. SDF-2-SDF [28] minimizes the difference between pairs of signed distance fields (SDFs). They construct an SDF using observed RGBD images and match it against the target SDF. In cases where a dense view of surface points is not available, such as when the camera is occluded or if the sensing is performed via contact, the constructed SDF will be invalid. Similar to them, we directly work with the target object's SDF, but importantly do not assign SDF values to observed free points. Instead, we only require that known free points be outside the surface (SDF 0-level set).

Diversity in registration has mainly been explored as characterizing the pose uncertainty. Censi [4] provides a closed form estimate for ICP based methods, but require that the initial point-correspondences are correct and that the minimization procedure does not get caught in local minima. This is unlikely to be valid with partial information, and does not utilize known free space. Buch et al. [2] produces high quality uncertainty estimates through MCMC simulation of a depth camera, which is computationally expensive while being restricted in input modality. Maken et al. [21] performs Stein Variational Gradient Descent (SVGD) on a differentiable formulation of the ICP objective. They approximate the distribution of poses by running ICP from different starts, which in general is not the distribution of poses consistent with the data. We compare against SVGD optimization as a baseline.

Generating diverse sets of high quality solutions has been explored explicitly in recent research on evolutionary optimization techniques. In particular, Quality Diversity (QD) [23]

techniques such as MAP-Elites [22] and CMA-MEGA [11] have been developed to optimize objectives while enforcing diversity in some aspect of the solutions. We leverage QD optimization methods with our proposed differentiable cost function to estimate a set of plausible diverse transforms.

III. Problem Statement

For a target object, we have its precomputed object frame signed distance function (SDF) derived from its 3D model, $\texttt{sdf} : \mathbb{R}^3 \to \mathbb{R}$, and are given a set of points $\mathcal{X} = \{(\mathbf{x}_1, s_1), ..., (\mathbf{x}_N, s_N)\}$ with known world positions $\mathbf{x}_n \in \mathbb{R}^3$ and semantics s_n (described below). $\mathcal{X}$ is produced from sensor data. Object registration is the problem of finding transforms $\mathbf{T} \in$ SE(3) that satisfy constraints imposed by $\mathcal{X}$. Let $\mathbf{T}^*$ be the true object transform, then the semantics are

$$s_n = \begin{cases} \texttt{free} & \text{implies } \texttt{sdf}(\mathbf{T}^*\mathbf{x}_n) > 0 \\ \texttt{occupied} & \text{implies } \texttt{sdf}(\mathbf{T}^*\mathbf{x}_n) < 0 \\ v_n & \text{implies } \texttt{sdf}(\mathbf{T}^*\mathbf{x}_n) = v_n \end{cases} \tag{1}$$

We quantify the degree to which the constraints of $\mathcal{X}$ are satisfied by using a cost function (lower is better) $C(\mathcal{X}, \mathbf{T}) = \sum_{n=1}^{N} c(\mathbf{T}\mathbf{x}_n, s_n)$ where $c_m \gg 0$ and

$$c(\mathbf{x}, s) = \begin{cases} c_m \mathbf{1}(\texttt{sdf}(\mathbf{x}) \leq 0) & \text{if } s = \texttt{free} \\ c_m \mathbf{1}(\texttt{sdf}(\mathbf{x}) \geq 0) & \text{if } s = \texttt{occupied} \\ |v - \texttt{sdf}(\mathbf{x})| & \text{else} \end{cases} \tag{2}$$

$\mathbf{1}$ is the indicator function that evaluates to 1 if the argument is true and otherwise evaluates to 0. We then define the plausible set $\mathcal{T}_\epsilon = \{\mathbf{T} \mid C(\mathcal{X}, \mathbf{T}) - C(\mathcal{X}, \mathbf{T}^*) < \epsilon\}$ where $\epsilon > 0$ is the degree of violation we allow in considering the constraints satisfied. Our goal is to produce a hypothesis transform set $\hat{\mathcal{T}}$ that covers $\mathcal{T}_\epsilon$. To quantify how well we cover this set, we use the Plausible Diversity metric [25] [24]:

$$M_c = \frac{1}{|\mathcal{T}_\epsilon|} \sum_{\mathbf{T} \in \mathcal{T}_\epsilon} \min_{\hat{\mathbf{T}} \in \hat{\mathcal{T}}} d(\mathbf{T}, \hat{\mathbf{T}}) \qquad \text{coverage} \tag{3}$$

$$M_p = \frac{1}{|\hat{\mathcal{T}}|} \sum_{\hat{\mathbf{T}} \in \hat{\mathcal{T}}} \min_{\mathbf{T} \in \mathcal{T}_\epsilon} d(\mathbf{T}, \hat{\mathbf{T}}) \qquad \text{plausibility} \tag{4}$$

$$M_{pd} = M_c + M_p \qquad \text{plausible diversity} \tag{5}$$

where d is a distance function on transforms, such as Chamfer Distance between the resulting transformed objects. This metric penalizes $\hat{\mathcal{T}}$ if 1) it does not include a transform that is close to each transform in the plausible set (a lack of coverage); or 2) includes transforms that are far from any transform in the plausible set (such transforms are implausible).

IV. Method

This section presents CHSEL, which consists of a differentiable cost function (relaxation of Eq. 2) that enables gradient-based methods to reduce a transform's cost, and a quality diversity optimization scheme, which uses that cost to estimate the set of plausible transforms. We also show how to update CHSEL's $\hat{\mathcal{T}}$ estimates online as more points are perceived.

A. Relaxation of Semantic Constraints

Eq. 2 has discrete components and is not differentiable. We would like a relaxation $\hat{C}(\mathcal{X}, \mathbf{T})$ of $C(\mathcal{X}, \mathbf{T})$ that is differentiable to be more amenable to optimization. For convenience, when the $\mathbf{T}$ used is unambiguous, we denote point positions in the world frame transformed to an estimated object frame as $\tilde{\mathbf{x}}$, where $\tilde{\mathbf{x}} = \mathbf{T}\mathbf{x}$ in homogeneous coordinates (append 1 to $\tilde{\mathbf{x}}$ and $\mathbf{x}$). Specifically, we want to efficiently compute the gradient $\nabla_{\mathbf{T}} \hat{C}(\mathcal{X}, \mathbf{T})$. For better geometric intuition, we consider the gradient contributed by each known point:

$$\nabla_{\mathbf{T}} \hat{C}(\mathcal{X}, \mathbf{T}) = \sum_{n=1}^{N} \nabla_{\mathbf{T}} \hat{c}(\mathbf{T}\mathbf{x}_n, s_n) = \sum_{n=1}^{N} \nabla_{\tilde{\mathbf{x}}} \hat{c}(\tilde{\mathbf{x}}, s_n) \tag{6}$$

This gradient is spatial and with respect to the transformed point. Intuitively, gradient descent will move the points spatially along their negative gradients through adjusting $\mathbf{T}$. This is visualized in Fig. 2. As it is clear what each gradient is with respect to, we drop the subscript in future usage.

The separate semantic classes motivate us to consider each case separately. We partition $\mathcal{X}$ into $\mathcal{X}_f = \{(\mathbf{x}, s) \mid s = \texttt{free}\}$, $\mathcal{X}_o = \{(\mathbf{x}, s) \mid s = \texttt{occupied}\}$, and $\mathcal{X}_k = \{(\mathbf{x}, s) \mid s \in \mathbb{R}\}$. We then decompose the gradient:

$$\nabla \hat{C}(\mathcal{X}, \mathbf{T}) = \sum_{\mathbf{x}, s \in \mathcal{X}_f} \nabla \hat{c}_f(\tilde{\mathbf{x}}) + \sum_{\mathbf{x}, s \in \mathcal{X}_o} \nabla \hat{c}_o(\tilde{\mathbf{x}}) + \sum_{\mathbf{x}, s \in \mathcal{X}_k} \nabla \hat{c}_k(\tilde{\mathbf{x}}, s) \tag{7}$$

At each point, the cost arises from an SDF value mismatch and thus the gradient must be along the direction of greatest SDF value change. This is provided precisely by $\nabla_{\tilde{\mathbf{x}}} \texttt{sdf}(\tilde{\mathbf{x}})$, the gradient (normalized such that $||\nabla \texttt{sdf}(\tilde{\mathbf{x}})||_2 = 1$) of the SDF at that point. Thus all cost gradients must be parallel or anti-parallel to the SDF gradient. See Section IV-B for how we achieve efficient lookup of SDF values and gradients. Fig. 2 shows our cost applied to points of each semantic class. Arrows indicate the negative cost gradient experienced by that point, which is the spatial direction the points will move along when we perform gradient descent on the cost. We define the gradients directly and assign its magnitude as the cost value.

1) Free space cost: From Eq. 2, points in $\mathcal{X}_f$ achieve 0 cost when $\texttt{sdf}(\tilde{\mathbf{x}}) > 0$. When $\texttt{sdf}(\tilde{\mathbf{x}}) \leq 0$ the negative gradient points towards the SDF 0-level set (surface of the object). To tolerate small degrees of violation due to uncertainty in the point positions, we aim for the α-level set where $\alpha < 0$. We define the magnitude of free space violation as $\max(0, \alpha - \texttt{sdf}(\tilde{\mathbf{x}}))$. This has the effect of only giving non-zero gradients to violations beyond α. Thus we have

$$\nabla \hat{c}_f(\tilde{\mathbf{x}}) = -C \max(0, \alpha - \texttt{sdf}(\tilde{\mathbf{x}})) \nabla \texttt{sdf}(\tilde{\mathbf{x}}) \tag{8}$$

where $C > 0$ is a scaling parameter. In a sense, it controls the degree of relaxation since using smaller values will lead to a smoother optimization path, particularly near the start of the optimization, while a higher value is needed to enforce the high cost from Eq. 2. This scaling parameter can be annealed during the optimization process, i.e. starting with a small value and increasing over optimization iterations.

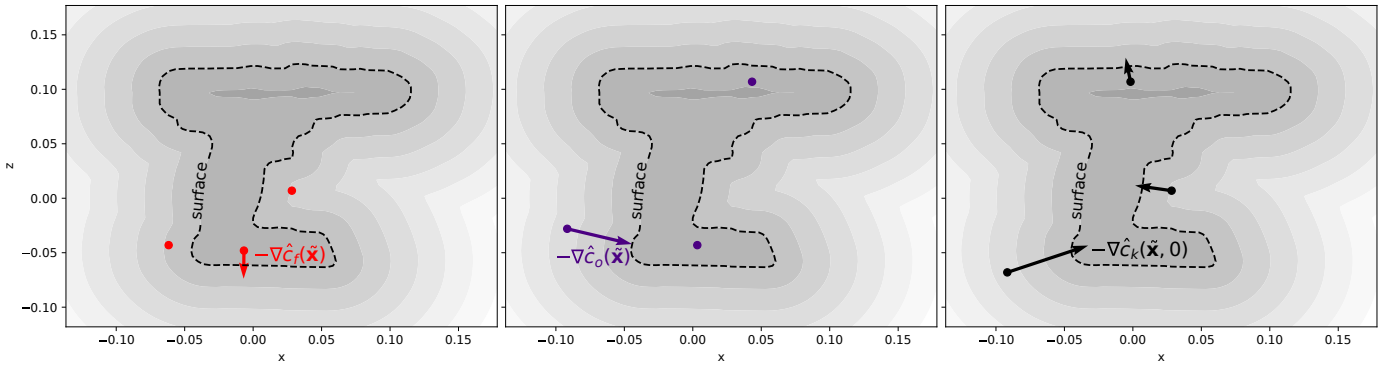

Fig. 2: $\hat{C}(\mathcal{X}, \mathbf{T})$ negative gradients with respect to sampled points shown for a X-Z cross section of the YCB drill. Red points are known free space and $-\nabla\hat{c}_f$ pushes them outside the object. Purple points are known occupied and $-\nabla\hat{c}_o$ pushes them inside the object. Black points have known SDF values (here they are known surface points, $\texttt{sdf}(\tilde{\mathbf{x}}) = 0$) and $-\nabla\hat{c}_k$ pushes them towards the corresponding SDF level set.

2) Occupied space cost: Symmetric to the free space cost, violating occupied points moves along $-\nabla\hat{c}_o$ to the $-\alpha$-level set. In this case, violation occurs when $\texttt{sdf}(\tilde{\mathbf{x}}) > -\alpha$ and has magnitude $-\min(0, -\alpha - \texttt{sdf}(\tilde{\mathbf{x}}))$:

$$\begin{aligned}\nabla\hat{c}_o(\tilde{\mathbf{x}}) &= -C\min(0, -\alpha - \texttt{sdf}(\tilde{\mathbf{x}}))\nabla\texttt{sdf}(\tilde{\mathbf{x}}) \\ &= C\max(0, \alpha + \texttt{sdf}(\tilde{\mathbf{x}}))\nabla\texttt{sdf}(\tilde{\mathbf{x}})\end{aligned} \tag{9}$$

3) Known SDF cost: This cost is a generalization of surface matching present in many registration methods. Known surface points are a special case of $s = 0$, and is commonly perceived through contact and visual perception. The cost's structure is similar to the previous costs, with the difference being that instead of α and $-\alpha$, each point has a separate desired level set given by its semantic value:

$$\nabla\hat{c}_k(\tilde{\mathbf{x}}, s) = (\texttt{sdf}(\tilde{\mathbf{x}}) - s)\nabla\texttt{sdf}(\tilde{\mathbf{x}}) \tag{10}$$

B. SDF Query Improvements

Evaluating $\nabla\hat{c}(\mathcal{X}, \mathbf{T})$ requires computing $\texttt{sdf}(\tilde{\mathbf{x}})$ and $\nabla\texttt{sdf}(\tilde{\mathbf{x}})$ for N known points. Regardless of the structure and efficiency of the given $\texttt{sdf}$, we precompute a voxel-grid approximation of it to enable fast parallel lookup. Each voxel reports the SDF value and gradient at the center of it. Each voxel is cubic with side length (resolution) r_t, with the whole grid being the object's bounding box padded by $\gamma > 0$ on all sides. Queries of points outside the voxel-grid are deferred to the original $\texttt{sdf}$, with $\texttt{sdf}(\tilde{\mathbf{x}}) = ||\tilde{\mathbf{x}} - \mathbf{x}'||_2$ and $\nabla_{\tilde{\mathbf{x}}}(\tilde{\mathbf{x}}) = (1 - 2\mathbf{1}(\tilde{\mathbf{x}} \text{ inside } \mathcal{M}))(\tilde{\mathbf{x}} - \mathbf{x}')$. Where $\mathbf{x}' = \arg\min_{\mathbf{x}\in\mathcal{M}}||\tilde{\mathbf{x}} - \mathbf{x}||_2$ is the closest point on the mesh to $\tilde{\mathbf{x}}$, and a ray is traced from the inside of the object (assuming object-centered origin is interior) to $\tilde{\mathbf{x}}$, with an even number of mesh surface crossings indicating it is inside. Lower r_t (a denser grid) trades higher memory usage for more accurate representation.

Another challenge to the efficiency of evaluating $\nabla\hat{C}(\mathcal{X}, \mathbf{T})$ is the representation of known free points $\mathcal{X}_f$. This is typically a volume, such as the space swept out by a robot's motion or derived from visual data. Representing this volume as a dense set of points makes $\nabla\hat{c}_f$ prohibitively expensive to evaluate. Similar to the 3D Normal Distribution Transform [19], we discretize the free space into a voxel-grid. The voxel-grid has resolution r_f, and the whole grid expands to the range of free points. r_f allows us to set the maximum point density.

C. Quality Diversity Optimization

With Eq. 7 we can optimize an initial $\mathbf{T}$ using stochastic gradient descent (SGD). The optimized $\mathbf{T}$ depends on the starting $\mathbf{T}$ and will achieve a local minima of $\hat{C}$. A naive approach to creating the estimated plausible transform set $\hat{\mathcal{T}}$ is to start with a set of transforms $\hat{\mathcal{T}}_0$ and perform SGD on each $\mathbf{T} \in \hat{\mathcal{T}}_0$ separately. We compare to this approach as an ablation in our experiments, where we find that this method often produces $\hat{\mathcal{T}}$ with poor plausible diversity as it relies only on the diversity of local minima for coverage.

Instead, we turn to Quality Diversity (QD) optimization. At a high level, in addition to the $\mathbb{R}^m$ solution space to search over to maximize an objective $f : \mathbb{R}^m \to \mathbb{R}$, there are k behavior (also known as measure) functions $B_i : \mathbb{R}^m \to \mathbb{R}$, jointly $\mathbf{B} : \mathbb{R}^m \to \mathbb{R}^k$. For the behavior space $\mathcal{B} = \mathbf{B}(\mathbb{R}^m)$ (image of $\mathbf{B}$), the QD objective is to find for each $\mathbf{b} \in \mathcal{B}$ a solution $\boldsymbol{\theta} \in \mathbb{R}^m$ such that $\mathbf{B}(\boldsymbol{\theta}) = \mathbf{b}$ and $f(\boldsymbol{\theta})$ is maximized. See Pugh et al. [23] for an overview of the field.

For our problem, $f(\boldsymbol{\theta}) = -\hat{C}(\mathcal{X}, \mathbf{T})$, and we search over the transforms represented in $\mathbb{R}^9$, with 3 translational components and the 6 dimensional representation of rotation suggested by Zhou et al. [36]. Our $\mathbf{B}$ extracts the translational components of the pose. In a sense, we are searching for the best rotation given some translation to minimize $\hat{C}$. Intuitively, QD's enforced diversity over $\mathcal{B}$ will prevent the collapse of $\hat{\mathcal{T}}$ when $\mathcal{X}$ does not sufficiently constrain our estimation.

In particular, we use CMA-MEGA [11] optimization to take advantage of our cost's differentiability to more efficiently search for good solutions. $\mathcal{B}$ is discretized into a regularly spaced grid, called the archive, with each cell holding the best solution for that cell. Diversity in $\hat{\mathcal{T}}$ is enforced by requiring each $\mathbf{T} \in \hat{\mathcal{T}}$ to come from a different cell in $\mathcal{B}$. This is an evolutionary method in that the lowest cost transforms

Algorithm 1: CHSEL: QD optimization for $\hat{\mathcal{T}}$

Given: $\mathcal{X}$ known points, $\hat{\mathcal{T}}_0$ initial transform set, $\hat{\mathcal{T}}_l$ low cost transform set, b_σ number of standard deviations to consider, n_o number of QD iterations

1 $\hat{\mathcal{T}}' \leftarrow$ SGD on $\hat{\mathcal{T}}_0$ with $\hat{C}(\mathcal{X}, \mathbf{T})$
2 $\mathcal{P} \leftarrow \mathbf{B}(\hat{\mathcal{T}}')$
3 $\boldsymbol{\mu} \leftarrow$ mean($\mathcal{P}$), $\boldsymbol{\sigma} \leftarrow$ std($\mathcal{P}$)
4 $\mathcal{B} \leftarrow$ grid with dimensions $[\boldsymbol{\mu} - b_\sigma \boldsymbol{\sigma}, \boldsymbol{\mu} + b_\sigma \boldsymbol{\sigma}]$
5 $\mathcal{B} \leftarrow$ UpdateCells($\mathcal{B}, \hat{\mathcal{T}}' \cup \hat{\mathcal{T}}_l$) `// initialize the search with low cost transforms`
6 $\mathcal{B} \leftarrow$ CMA-MEGA($\mathcal{B}, \hat{C}, n_o$)
7 $\hat{\mathcal{T}} \leftarrow \{\mathbf{T} | \mathbf{T} \in$ cells from $\mathcal{B}$ with $|\hat{\mathcal{T}}_0|$ lowest costs $\}$

from different cells are iteratively combined to generate new transforms. Thus, it is valuable to populate the archive with low cost transforms $\hat{\mathcal{T}}_l$ to initialize the search. If no prior estimate is available, $\hat{\mathcal{T}}_l = \{\}$, but when we run CHSEL iteratively, $\hat{\mathcal{T}}_l$ contains the estimates from the previous iteration (see Sec. IV-D).

Algorithm 1 describes how we use QD optimization. First, we run SGD on the given initial transform set $\hat{\mathcal{T}}_0$ using Eq. 7 to create an $\hat{\mathcal{T}}'$. We extract its translation components $\mathbf{B}(\hat{\mathcal{T}}) = \mathcal{P}$. Using the mean $\boldsymbol{\mu}$ and standard deviation $\boldsymbol{\sigma}$ of $\mathcal{P}$ along each dimension, we size the grid $\mathcal{B}$ between $[\boldsymbol{\mu} - b_\sigma \boldsymbol{\sigma}, \boldsymbol{\mu} + b_\sigma \boldsymbol{\sigma}]$. The grid is centered on the mean $\boldsymbol{\mu}$ with extents scaled by the standard deviation $\boldsymbol{\sigma}$ along each dimension. A large $\boldsymbol{\sigma}$ suggests that there are low cost solutions with very different values along that dimension, motivating a wider search range. The parameter $b_\sigma > 0$ adjusts how many standard deviations out we search for solutions. We initialize $\mathcal{B}$ with known low cost transforms from $\hat{\mathcal{T}}_l$, along with the SGD solutions $\hat{\mathcal{T}}'$ to seed the QD optimization. Note that sizing the archive defines the region of the behavior space to search over while initializing it populates some grid cells with transform values. We then run CMA-MEGA on $\mathcal{B}$ for n_o iterations to populate $\mathcal{B}$ with the lowest cost $\mathbf{T}$ for each cell. Finally, we select the $\mathbf{T}$ from the $|\hat{\mathcal{T}}_0|$ lowest cost cells as $\hat{\mathcal{T}}$.

Since we initialize $\mathcal{B}$ with $\hat{\mathcal{T}}' \cup \hat{\mathcal{T}}_l$, the QD optimization can be seen as a fine-tuning process. Initially, each $\mathbf{T} \in \hat{\mathcal{T}}' \cup \hat{\mathcal{T}}_l$ is the best solution for their respective cells (assuming they fall into different cells of $\mathcal{B}$). If a lower cost transform exists for a cell, QD optimization will eventually find it and replace the original $\mathbf{T}$ from $\hat{\mathcal{T}}' \cup \hat{\mathcal{T}}_l$.

D. Online Updates to $\hat{\mathcal{T}}$

Registration can be performed iteratively as new sensor data are added to $\mathcal{X}$. Information from the previous registration allows us to more efficiently search for $\hat{\mathcal{T}}$. Our update process is described in Algorithm 2. First, we consider the generation and update of the initial transform set $\hat{\mathcal{T}}_0$. Before any registration, we sample uniformly at random (both position and rotation), within the given workspace $\mathcal{W}$ where the object could possibly be. Assuming the object has not moved, we

Algorithm 2: Online update of $\hat{\mathcal{T}}$

Given: $\mathcal{W}$ workspace dimension, σ_t translation noise, σ_R rotation noise

1 $\hat{\mathcal{T}}_0 \leftarrow \mathcal{P} \sim U(\mathcal{W}) \times U(\text{SO(3)})$
2 $\hat{\mathcal{T}}_l \leftarrow \{\}$
3 **while** *register* **do**
4 $\quad \mathcal{X} \leftarrow$ perceived from environment
5 $\quad \hat{\mathcal{T}} \leftarrow$ CHSEL($\mathcal{X}, \hat{\mathcal{T}}_0, \hat{\mathcal{T}}_l$)
6 $\quad \hat{\mathcal{T}}_l \leftarrow \hat{\mathcal{T}}$
7 $\quad \mathbf{T}' \leftarrow \arg\min_{\mathbf{T} \in \hat{\mathcal{T}}} \hat{C}(\mathcal{X}, \mathbf{T})$
8 $\quad \hat{\mathcal{T}}_0' \leftarrow \{\}$
9 $\quad$ **for** $i \leftarrow 1$ **to** $|\hat{\mathcal{T}}_0|$ **do**
10 $\quad\quad \boldsymbol{\Delta t} \sim \mathcal{N}(\mathbf{0}, \mathbf{diag}([\sigma_t, \sigma_t, \sigma_t]))$
11 $\quad\quad \theta \sim \mathcal{N}(0, \sigma_R)$
12 $\quad\quad \mathbf{e} \sim U(\{\mathbf{x} \mid ||\mathbf{x}||_2 = 1, \mathbf{x} \in \mathbb{R}^3\})$
13 $\quad\quad \boldsymbol{\Delta R} \leftarrow e^{\theta \mathbf{e}}$ `// axis angle to matrix`
14 $\quad\quad \hat{\mathcal{T}}_0' \leftarrow \hat{\mathcal{T}}_0' \cup \{(\boldsymbol{\Delta t} + \text{trans}(\mathbf{T}'), \boldsymbol{\Delta R} \cdot \text{rot}(\mathbf{T}'))\}$
15 $\quad \hat{\mathcal{T}}_0 \leftarrow \hat{\mathcal{T}}_0'$

update $\hat{\mathcal{T}}_0$ as $|\hat{\mathcal{T}}_0|$ perturbations around the best $\mathbf{T}$ of the previous estimated set, $\mathbf{T}' = \arg\min_{\mathbf{T} \in \hat{\mathcal{T}}} \hat{C}(\mathcal{X}, \mathbf{T})$. We perturb its translational components with Gaussian noise $\sigma_t > 0$ along each dimension. Then, in line 12, we uniformly sample a rotation axis, scaled with an angle sampled as Gaussian noise with $\sigma_R > 0$. We take the exponential map of this axis-angle representation and multiply it by the rotational component of $\mathbf{T}'$. This sampling based update of $\hat{\mathcal{T}}_0$ helps with escaping bad local minima.

Secondly, the data update changes $\mathcal{X}$ and invalidates the previously computed $\mathcal{B}$, but the solutions in each cell of the $\mathcal{B}$ may still have low cost with respect to the updated $\mathcal{X}$. We select them as $\hat{\mathcal{T}}_l$ and use them to initialize the next QD optimization in line 5 of Algorithm 1.

V. Experiments

In this section, we first describe our simulated and real robot environments, and how we estimate $\mathcal{X}$ from sensor data. Next, we describe how we generate the plausible set. Then, we describe our baselines and ablations and how we quantitatively evaluate each method on probing experiments of objects in simulation and in the real world. Lastly, we evaluate the value of our cost as a QD objective.

In these experiments, a target object is in a fixed pose inside an occluded cabinet, and we estimate its pose through a fixed sequence of probing actions by a robot (some of which will result in contact). We run all methods after each probe, updating our set of pose estimates using Algorithm 2. Note that all methods receive the same known points $\mathcal{X}$ after each probe. Each probe extends the robot straight into the cabinet for a fixed distance or until contact. The probing locations are configuration specific and designed such that at least some contacts are made. We use YCB objects [3] for both simulated and real experiments as their meshes are

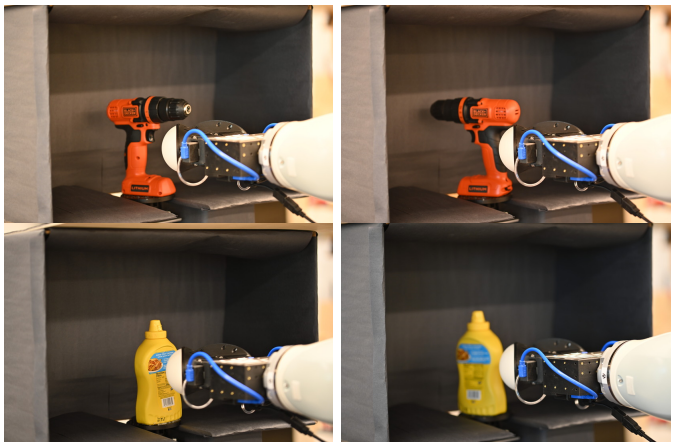

Fig. 3: Real probing configurations for the drill and mustard.

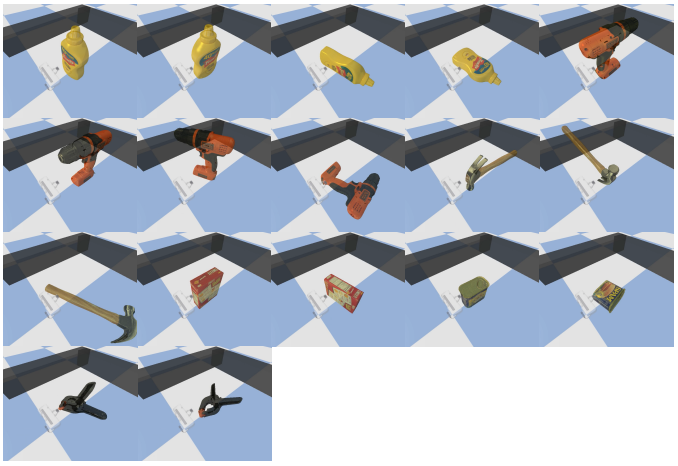

Fig. 4: Simulated probing configurations for multiple YCB objects.

readily available. We precompute the object-frame SDF from these meshes as described in Section IV-B. In all cases, we are estimating a set of 30 transforms ($|\hat{\mathcal{T}}|= 30$), and use parameters $\alpha = -10mm, b_\sigma = 3, C = 20, \sigma_t = 0.05m, \sigma_R = 0.3$. In the sim experiments we use $n_o = 100$ and in the real world $n_o = 500$. For the Plausible Diversity distance function $d(\mathbf{T}, \hat{\mathbf{T}})$, we sample 200 points on the object surface and evaluate the Chamfer Distance between them after being transformed. Note that the 200 points sampled are different per trial. We extract the x and y components of the pose using $\mathbf{B}$ - we found no significant difference in performance from also extracting z.

A. Simulated Environment

We use PyBullet [7] to simulate a Franka Emika (FE) gripper (see Fig. 4) that is position controlled. The workspace is voxelized with resolution $r_f = 25mm$ and spans $[-0.1, 0.5] \times [-0.3, 0.3] \times [-0.075, 0.625]$ in meters. We label the boundary of the workspace as free space. The SDF is voxelized with resolution $r_t = 10mm$ with padding $\gamma = 50mm$. The robot sweeps out voxels in the workspace grid during its probing motions, and $\mathcal{X}_f$ is given by the center of the swept voxels. $\mathcal{X}_o$ is empty as we have have no sensors that detect non-surface occupancy, though such information can be added if available. $\mathcal{X}_k$ is given by the contact points, with each having semantics $s = 0$ since contact can only occur on the object surface. Both the gripper and object are rigid and so only make single-point contacts which we retrieve from the simulator. For different trials, we seed the random number generator with different values.

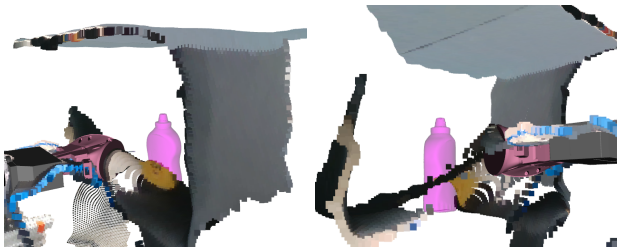

Fig. 5: Unreliable RGBD readings inside the partially occluded cabinet, viewed from both sides, with an approximate pose of the mustard bottle in purple.

B. Real Environment

For our real world experiment, we equip a 7DoF KUKA LBR iiwa arm with two soft-bubble tactile sensors [17] on the gripper (see Fig. 1 and Fig. 3). The soft-bubble sensors allow us to detect patch contact, which we consider as any point with deformation beyond $4mm$ and being in the top 10^{th} percentile of deformations. We use a mean filter to remove noise and downsample such that each contact produces at most 50 surface points.

The workspace is a physical cabinet mock-up and is voxelized with resolution $r_f = 10mm$, spanning $[0.7, 1.1] \times [-0.2, 0.2] \times [0.31, 0.6]$ meters. The SDF is voxelized with resolution $r_t = 5mm$ with padding $\gamma = 50mm$. In addition to populating $\mathcal{X}_f$ with robot swept volume, we utilize a RealSense RGBD camera, partially occluded by the cabinet. The camera is unreliable near occluding edges (see Fig. 5), thus we do not assume the object can be reliably segmented from the camera view, so we only use the free space information derived from the depth data. To that effect, we trace rays from the camera to 95% of each pixel's detected depth and add them to $\mathcal{X}_f$.

C. Computing Plausible Set

In order to evaluate our method, we need to compute $\mathcal{T}_\epsilon$, which is very computationally intensive. We compute $\mathcal{T}_\epsilon$ by densely sampling transforms around $\mathbf{T}^*$ and evaluate each using Eq. 2 with $c_m = 100000$. Specifically, we search over a grid spanning $[-0.1, 0.15] \times [-0.2, 0.2] \times [0, 0.1]$ meters with 15 cells along each dimension. We also uniformly random sample 10000 rotations which we combine with each translation cell. See Table I for the ϵ used to generate the plausible set of each object. They were selected such that most probe trials have around 30 members in $\mathcal{T}_\epsilon$ halfway through.

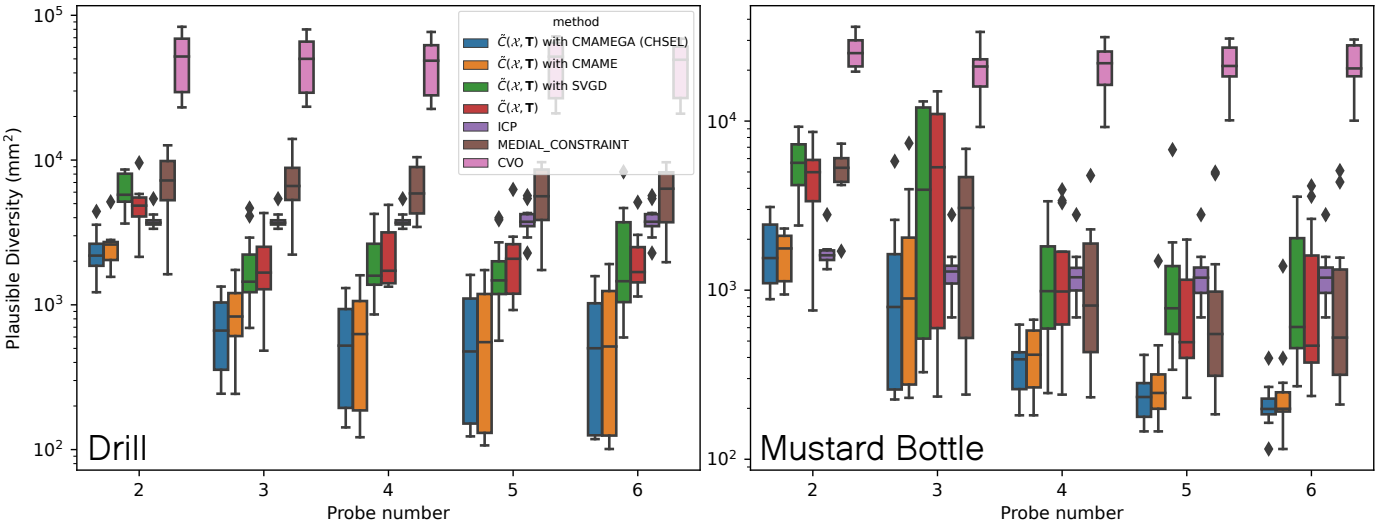

Fig. 6: Plausible Diversity for real drill (left) and mustard (right) probing experiments across 2 configurations and 6 trials each. The bars indicate the 25 to 75 percentile while the whiskers are the min and max with outliers as diamonds. Lower is better.

Object	ϵ
Real Drill	0.001
Real Mustard Bottle	0.0003
Sim Drill	0.001
Sim Mustard Bottle	0.0003
Sim Hammer	0.001
Sim Cracker Box	0.0005
Sim Spam Can	0.0003
Sim Clamp	0.0007

TABLE I: ϵ used to generate the plausible set for each objects.

In simulation, we retrieve $\mathbf{T}^*$ from the simulator, while on the real robot we first manually specify an approximate pose, then search in two passes. The first pass searches around the specified pose to find the optimal transform, which is then used as $\mathbf{T}^*$ for the second pass.

D. Baselines

We compare against **ICP** as a weak baseline that does not use free space information. ICP registers the known surface points against another point set, which we provide as 500 points randomly sampled from the object surface. Note that these points are different for each trial. ICP is run until convergence.

Secondly, we compare against Continuous Visual Odometry (**CVO**) [34], the state of the art in semantic point set registration, and a continuous generalization of 3D NDT. We use 2 dimensional semantics to represent free points as $[0.9, 0.1]$ and surface points as $[0.1, 0.9]$. CVO registers the free and surface points against another semantic point set, which we provide as the center of the precomputed SDF voxels. Voxels with SDF value between $[-r_t, r_t]$ are labelled with surface semantics, and voxels with SDF value greater than r_t are labelled as free. Note that there are many more free points than surface points ($\approx 125:1$).

Next, we consider using Stein Variational Gradient Descent (**SVGD**) [18] of Eq. 7 to enforce diversity. We formulate $p(\mathbf{T}|\mathcal{X}) \propto e^{-\beta \hat{C}(\mathcal{X}, \mathbf{T})}$, and select $\beta = 5$ as a scaling term for how peaked the distribution is. We have $|\hat{\mathcal{T}}_0|$ stein particles, each one initalized with a separate $\mathbf{T} \in \hat{\mathcal{T}}_0$, implicitly defining the prior. We use an RBF kernel with scale 0.01.

Lastly, we compare against Haugo and Stahl [15], which forms free space constraints by covering the free space using balls along the volume's medial axis (we refer to this baseline as **Medial Constraint**). For each ball we have cost $\max(0, B_r - \texttt{sdf}(\tilde{B}_c))^2$ where B_r is its radius and B_c is the center position of the ball. For each surface point we have cost $\texttt{sdf}(\tilde{\mathbf{x}})^2$. The total cost is the sum of the mean ball cost and the mean point costs. We optimize this cost using CMA-ES [14], a gradient-free evolutionary optimization technique.

E. Ablations

We ablate components of our method starting with how useful the gradient is for accelerating QD optimization. Instead of CMA-MEGA, we use **CMA-ME** [12] which does not explore using gradients.

We also consider just gradient descent on Eq. 7 to evaluate the value of additional optimization. We run Adam for 500 iterations with learning rate 0.01, reset to 0.01 every 50 iterations. These are also the parameters used for initializing CMA-ME and CMA-MEGA.

F. Probing Experiments

Qualitatively, we see the progress of a probing experiment and the elimination of hypothesis transforms through gaining known free space points in Fig. 8. From the initial probes along the back of the drill, it could take on many possible upright orientations. Note that after the probe in the second row, the contact points constrain the pose such that the contacts must lie on the back of the drill. As we probe the left side of the drill, without making contact, we eliminate transforms that would conflict with the new free space points. Probing the other side further narrows down the plausible transforms. Note

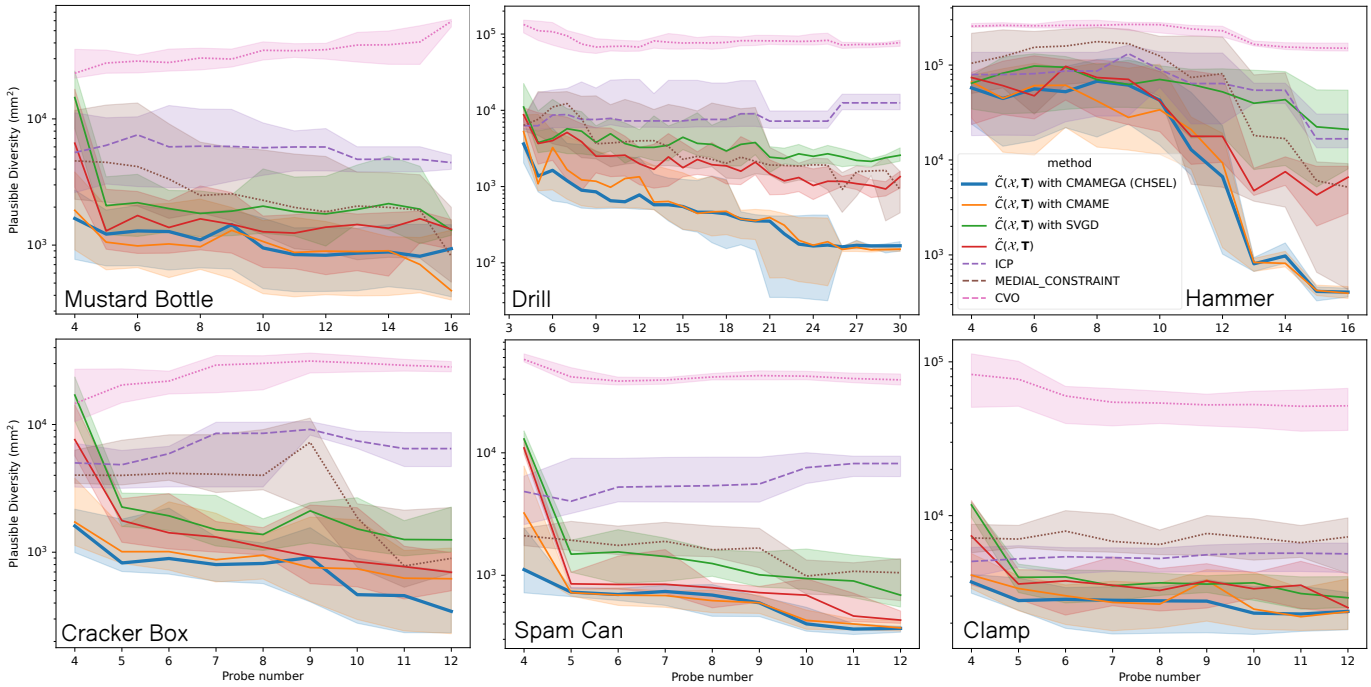

Fig. 7: Plausible Diversity for simulated probing experiments across different YCB objects, with 2 to 4 configurations from Fig. 4 over 10 trials each. The median is in bold while the shaded region represents 25 to 75 percentile. Lower is better.

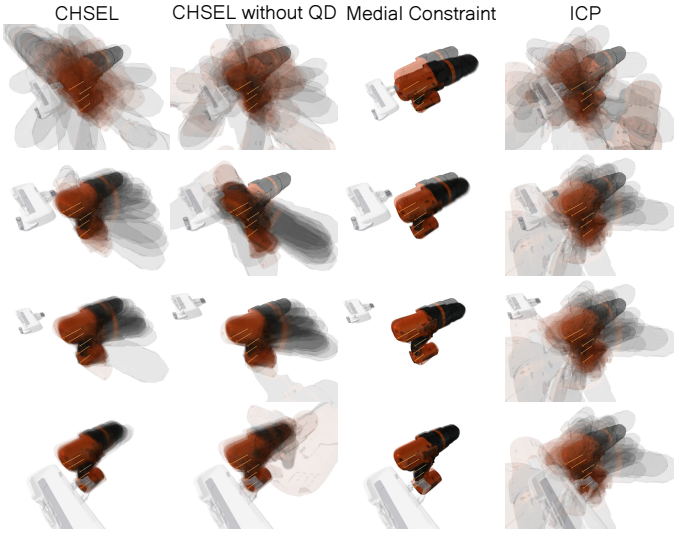

Fig. 8: Reducing uncertainty in estimated pose as a result of additional free space points for selected methods, obtained by probing to the sides of the YCB drill. $\hat{\mathcal{T}}$ is represented as transformed copies of the mesh while contact points are drawn in orange, with the line indicating the direction of the probe.

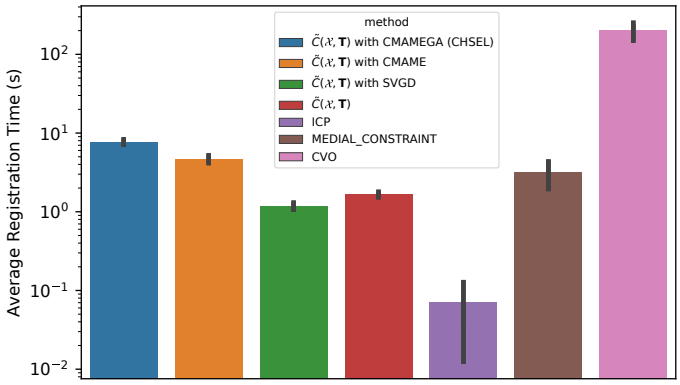

Fig. 9: Average time per registration of 30 transforms on the real probing experiments. Error bars indicate one standard deviation.

the lack of diversity from the Medial Constraint baseline and the poor estimation from ICP since it cannot use free space information.

Fig. 6 summarizes the results of the real probing experiments on the YCB drill and mustard bottle, each in two different configurations (see Fig. 3) over 6 trials. Fig. 7 summarizes the results of the simulation probing experiments on the YCB drill, mustard bottle, hammer, cracker box, spam can, and clamp. Additionally, we show the average time it takes for each method to perform registration on the real experiments in Fig. 9. This involves producing 30 transforms with $|\mathcal{X}| \in [13000, 21000]$, and $|\mathcal{X}_k| \in [0, 150]$. Note that all methods apart from CVO use parallelized implementations. Computations were performed on a NVIDIA GeForce RTX 2080 Ti with 11GB of VRAM.

From Fig. 6 and Fig. 7, we see that applying QD optimization to $\hat{C}$ in general outperforms baselines and the ablations. This is particularly true on more irregular objects such as the

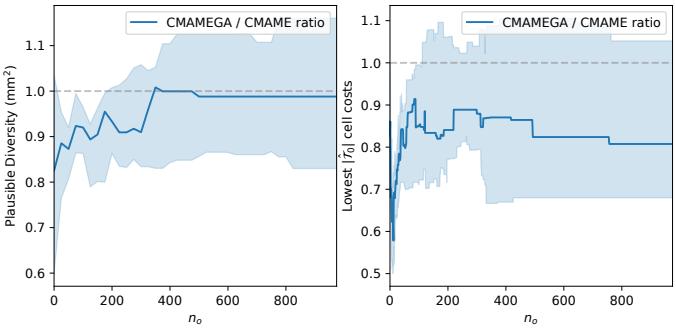

Fig. 10: Comparison of QD optimization progress using $\hat{C}$ on the real drill experiment. Results are averaged across 6 trials and probe numbers 5 and 6. Median is in bold while the shaded region represents 25 to 75 percentile.

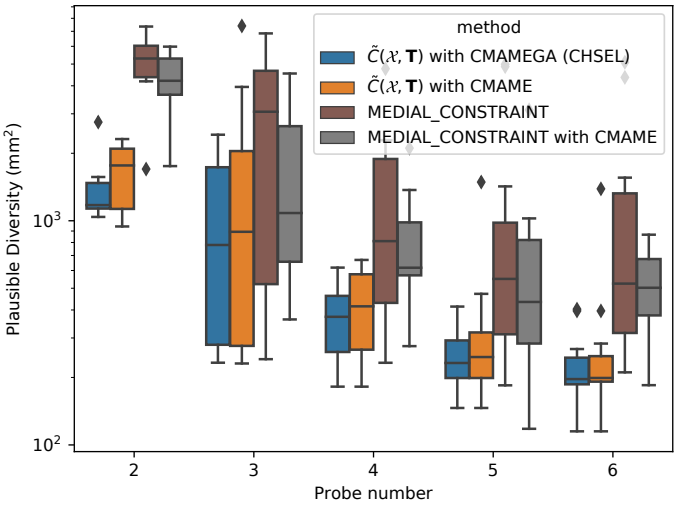

Fig. 11: Plausible Diversity on the two real mustard bottle experiments with a focus on the improvement the Medial Constraint baseline receives from QD optimization.

drill and hammer, and when we have noisy data in the real experiment. Even without QD optimization, gradient descent on $\hat{C}$ outperforms the Medial Constraint baseline. This may be due to the ability of CMA-ES to escape local minima, leading to low coverage, as seen in Fig. 8. All methods, including ICP, outperform CVO. We suspect this is due to the large imbalance of $|\mathcal{X}_f|$ to $|\mathcal{X}_k|$ ($\approx 125:1$). See Appendix A for an investigation of CVO's performance.

G. QD Method Comparison

We investigate the value of our formulated cost's differentiability by considering the QD optimization process in further detail. In Fig. 10, we compare the $\hat{\mathcal{T}}$ performance of using CMA-MEGA and CMA-ME as we increase the number of QD optimization iterations. The results are from the front-facing real drill experiment (Fig. 3 top left), averaged over probes 5 and 6, and across the 6 trials. Both methods are initialized with the same $\hat{\mathcal{T}}_0$ and $\hat{\mathcal{T}}_l$ each trial and probe (see Algorithm 1). We see that CMA-MEGA is able to use our gradients to reach lower Plausible Diversity and average cost of the best cells in fewer iterations, and that they converge and reach parity after around 500 iterations (fewer in simulation due to lack of noise). In Fig. 6 and Fig. 7, both methods have run for enough iterations to converge. On average, each CMA-ME iteration takes 8.37ms while each CMA-MEGA iteration takes 11.5ms.

H. QD Objective Comparison

Lastly, we investigate how well QD optimization works with other objectives. We perform CMA-ME optimization using the Medial Constraint objective, with $\mathcal{B}$ initialized and sized from the $\hat{\mathcal{T}}$ estimated by Medial Constraint using CMA-ES. Fig. 11 shows results on the real mustard bottle experiments, where we see that while QD optimization improves the Medial Constraint performance, our method still significantly outperforms it. This demonstrates the value of $\hat{C}$ as a QD objective.

VI. Discussion and Future Work

In sequential registration problems such as our probing experiments, we assume that the object is stationary and that the updated semantic points are given. However, keeping the object still while probing it is not trivial, as every contact has the potential to move the object. Rapid force and tactile feedback could minimize this issue. Contact could also be made with other objects during the probing motions. Contact point tracking and reasoning over object-contact associations is not within the scope of this paper. However, in future work we will explore using a method such as STUCCO [35] to estimate object-contact associations and add the proper contacts to $\mathcal{X}$.

The experiments in this paper used a fixed sequence of probing motions. This makes for a fair comparison between methods, since the sequence is not dependent on any method's pose estimates. However, in practice, the next probing motion should depend on the current pose estimate. In future work we aim to explore how to reason over the plausible set of poses and plan trajectories that efficiently disambiguate between them, so as to localize the object with as few probing motions as possible.

VII. Conclusion

We presented CHSEL, a pose registration method that utilizes point semantics, such as whether a point is in free space or on the object surface, to impose additional constraints and reduce pose ambiguity. Rather than a single best estimate, it produces a set of diverse plausible estimates given the observed data. We showed that it performs well on both simulated and real data collected from robot probing experiments. In particular, we separately demonstrated the value of performing Quality Diversity (QD) optimization for registration, and the strength of our proposed differentiable cost function as a QD objective. Additionally, we showed how to update the estimated transform set online with updated data, that CHSEL performs well on data with few contact points, and that it is seamless to integrate vision as an input modality.

REFERENCES

[1] Kai O Arras, José A Castellanos, Martin Schilt, and Roland Siegwart. Feature-based multi-hypothesis localization and tracking using geometric constraints. *Robotics and Autonomous Systems*, 44(1):41–53, 2003.

[2] Anders Glent Buch, Dirk Kraft, et al. Prediction of icp pose uncertainties using monte carlo simulation with synthetic depth images. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4640–4647. IEEE, 2017.

[3] Berk Calli, Arjun Singh, James Bruce, Aaron Walsman, Kurt Konolige, Siddhartha Srinivasa, Pieter Abbeel, and Aaron M Dollar. Yale-cmu-berkeley dataset for robotic manipulation research. *The International Journal of Robotics Research*, 36(3):261–268, 2017.

[4] Andrea Censi. An accurate closed-form estimate of icp's covariance. In *Proceedings 2007 IEEE international conference on robotics and automation*, pages 3167–3172. IEEE, 2007.

[5] Liang Cheng, Song Chen, Xiaoqiang Liu, Hao Xu, Yang Wu, Manchun Li, and Yanming Chen. Registration of laser scanning point clouds: A review. *Sensors*, 18(5):1641, 2018.

[6] Alvaro Collet, Dmitry Berenson, Siddhartha S Srinivasa, and Dave Ferguson. Object recognition and full pose registration from a single image for robotic manipulation. In *2009 IEEE International Conference on Robotics and Automation*, pages 48–55. IEEE, 2009.

[7] Erwin Coumans and Yunfei Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. 2016.

[8] Snehal Dikhale, Karankumar Patel, Daksh Dhingra, Itoshi Naramura, Akinobu Hayashi, Soshi Iba, and Nawid Jamali. Visuotactile 6d pose estimation of an in-hand object using vision and tactile sensor data. *IEEE Robotics and Automation Letters*, 7(2):2148–2155, 2022.

[9] Danny Driess, Peter Englert, and Marc Toussaint. Active learning with query paths for tactile object shape exploration. In *2017 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 65–72. IEEE, 2017.

[10] Robert Eidenberger and Josef Scharinger. Active perception and scene modeling by planning with probabilistic 6d object poses. In *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1036–1043. IEEE, 2010.

[11] Matthew Fontaine and Stefanos Nikolaidis. Differentiable quality diversity. *Advances in Neural Information Processing Systems*, 34: 10040–10052, 2021.

[12] Matthew C Fontaine, Julian Togelius, Stefanos Nikolaidis, and Amy K Hoover. Covariance matrix adaptation for the rapid illumination of behavior space. In *Proceedings of the 2020 genetic and evolutionary computation conference*, pages 94–102, 2020.

[13] Jiahui Fu, Qiangqiang Huang, Kevin Doherty, Yue Wang, and John J Leonard. A multi-hypothesis approach to pose ambiguity in object-based slam. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7639–7646. IEEE, 2021.

[14] Nikolaus Hansen, Sibylle D Müller, and Petros Koumoutsakos. Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (cma-es). *Evolutionary computation*, 11 (1):1–18, 2003.

[15] Simen Haugo and Annette Stahl. Iterative closest point with minimal free space constraints. In *International Symposium on Visual Computing*, pages 82–95. Springer, 2020.

[16] Michael C Koval, Nancy S Pollard, and Siddhartha S Srinivasa. Pose estimation for planar contact manipulation with manifold particle filters. *The International Journal of Robotics Research*, 34(7):922–945, 2015.

[17] Naveen Kuppuswamy, Alex Alspach, Avinash Uttamchandani, Sam Creasey, Takuya Ikeda, and Russ Tedrake. Soft-bubble grippers for robust and perceptive manipulation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9917–9924. IEEE, 2020.

[18] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *Advances in neural information processing systems*, 29, 2016.

[19] Martin Magnusson. *The three-dimensional normal-distributions transform: an efficient representation for registration, surface analysis, and loop detection*. PhD thesis, Örebro universitet, 2009.

[20] Martin Magnusson, Achim Lilienthal, and Tom Duckett. Scan registration for autonomous mining vehicles using 3d-ndt. *Journal of Field Robotics*, 24(10):803–827, 2007.

[21] Fahira Afzal Maken, Fabio Ramos, and Lionel Ott. Stein ICP for Uncertainty Estimation in Point Cloud Matching. *IEEE Robotics and Automation Letters*, 7(2):1063–1070, 2021.

[22] Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites. *arXiv preprint arXiv:1504.04909*, 2015.

[23] Justin K Pugh, Lisa B Soros, and Kenneth O Stanley. Quality diversity: A new frontier for evolutionary computation. *Frontiers in Robotics and AI*, page 40, 2016.

[24] Brad Saund and Dmitry Berenson. CLASP: Constrained Latent Shape Projection for Refining Object Shape from Robot Contact. In *Conference on Robot Learning*, pages 1391–1400. PMLR, 2022.

[25] Bradley Saund and Dmitry Berenson. Diverse plausible shape completions from ambiguous depth images. In *Conference on Robot Learning*, pages 1802–1813. PMLR, 2021.

[26] Aleksandr Segal, Dirk Haehnel, and Sebastian Thrun. Generalized-icp. In *Robotics: science and systems*, volume 2, page 435. Seattle, WA, 2009.

[27] Andrea Sipos and Nima Fazeli. Simultaneous contact location and object pose estimation using proprioception and tactile feedback. In *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 3233–3240. IEEE, 2022.

[28] Miroslava Slavcheva, Wadim Kehl, Nassir Navab, and Slobodan Ilic. Sdf-2-sdf: Highly accurate 3d object reconstruction. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14*, pages 680–696. Springer, 2016.

[29] Sudharshan Suresh, Zilin Si, Stuart Anderson, Michael Kaess, and Mustafa Mukadam. Midastouch: Monte-carlo inference over distributions across sliding touch. *arXiv preprint arXiv:2210.14210*, 2022.

[30] Gary KL Tam, Zhi-Quan Cheng, Yu-Kun Lai, Frank C Langbein, Yonghuai Liu, David Marshall, Ralph R Martin, Xian-Fang Sun, and Paul L Rosin. Registration of 3d point clouds and meshes: A survey from rigid to nonrigid. *IEEE transactions on visualization and computer graphics*, 19(7):1199–1217, 2012.

[31] Fang Wang and Zijian Zhao. A survey of iterative closest point algorithm. In *2017 Chinese Automation Congress (CAC)*, pages 4395–4399. IEEE, 2017.

[32] Jay M Wong, Vincent Kee, Tiffany Le, Syler Wagner, Gian-Luca Mariottini, Abraham Schneider, Lei Hamilton, Rahul Chipalkatty, Mitchell Hebert, David MS Johnson, et al. Segicp: Integrated deep semantic segmentation and pose estimation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5784–5789. IEEE, 2017.

[33] Jingxi Xu, Han Lin, Shuran Song, and Matei Ciocarlie. Tandem3d: Active tactile exploration for 3d object recognition. *arXiv preprint arXiv:2209.08772*, 2022.

[34] Ray Zhang, Tzu-Yuan Lin, Chien Erh Lin, Steven A Parkison, William Clark, Jessy W Grizzle, Ryan M Eustice, and Maani Ghaffari. A New Framework for Registration of Semantic Point Clouds from Stereo and RGB-D Cameras. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 12214–12221. IEEE, 2021. doi: 10.1109/ICRA48506.2021.9561929.

[35] Sheng Zhong, Nima Fazeli, and Dmitry Berenson. Soft tracking using contacts for cluttered objects to perform blind object retrieval. *IEEE Robotics and Automation Letters*, 7(2):3507–3514, 2022.

[36] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2019.

APPENDIX A
CVO PERFORMANCE

Qualitatively, we noticed that CVO's estimated transforms tend to place the object such that $\mathcal{X}_f$ is in concentrated regions of observed free points. To check if this free/surface imbalance was the cause of CVO's poor performance, we ran CVO while ignoring $\mathcal{X}_f$ on the real mustard bottle experiments (results shown in Fig. 12). We see that CVO performs comparably to ICP when ignoring $\mathcal{X}_f$. This is significantly better than when considering $\mathcal{X}_f$, suggesting that CVO is not able to effectively use the free space information.

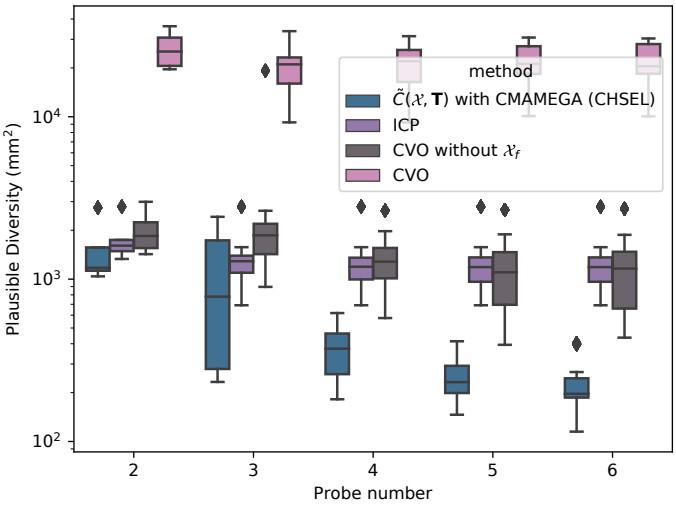

Fig. 12: Plausible Diversity on the two real mustard bottle experiments with a focus on the improvement the CVO baseline receives from ignoring $\mathcal{X}_f$. Lower is better.