Sliding Touch-Based Exploration for Modeling Unknown Object Shape with Multi-Fingered Hands

N.B. When citing this work, cite the original published paper.

(article starts on next page)

Sliding Touch-based Exploration
for Modeling Unknown Object Shape with Multi-finger Hands

Yiting Chen[1], Ahmet Ercan Tekden[1], Marc Peter Deisenroth[2], Yasemin Bekiroglu[1,2]

Abstract— Efficient and accurate 3D object shape reconstruction contributes significantly to the success of a robot's physical interaction with its environment. Acquiring accurate shape information about unknown objects is challenging, especially in unstructured environments, e.g. the vision sensors may only be able to provide a partial view. To address this issue, tactile sensors could be employed to extract local surface information for more robust unknown object shape estimation. In this paper, we propose a novel approach for efficient unknown 3D object shape exploration and reconstruction using a multi-finger hand equipped with tactile sensors and a depth camera only providing a partial view. We present a multi-finger sliding touch strategy for efficient shape exploration using a Bayesian Optimization approach and a single-leader-multi-follower strategy for multi-finger smooth local surface perception. We evaluate our proposed method by estimating the 3D shape of objects from the YCB and OCRTOC datasets based on simulation and real robot experiments. The proposed approach yields successful reconstruction results relying on only a few continuous sliding touches. Experimental results demonstrate that our method is able to model unknown objects in an efficient and accurate way.

I. INTRODUCTION

Robotic physical interaction tasks, such as assembly, grasping and dexterous manipulation, benefit from knowing the 3D shape of the object. Though visual-only (RGB and depth) 3D object perception methods have been well studied during past years, robotic visual perception accuracy is reduced in presence of noisy, incomplete, or occluded data. Approaches rely on sufficient prior knowledge from pre-trained neural network models, which often fail due to uncertainties regarding unknown objects and unstructured environments [1].

Humans make use of both visual and tactile information for unknown objects' perception and modeling [2]. Even when visual perception is heavily confined, humans can still explore target objects' size, shape, and texture using fingertips tactile sensing. Thanks to the advances in robotic tactile sensors, nowadays robots are enabled to have human-like tactile perception [3]–[5]. Optical tactile sensors [6]–[10] are popular given their high-frequency, high resolution, and ability to capture fine details about the local contact. The development of tactile simulators [11]–[15] has also contributed greatly to include tactile sensing in robotic applications.
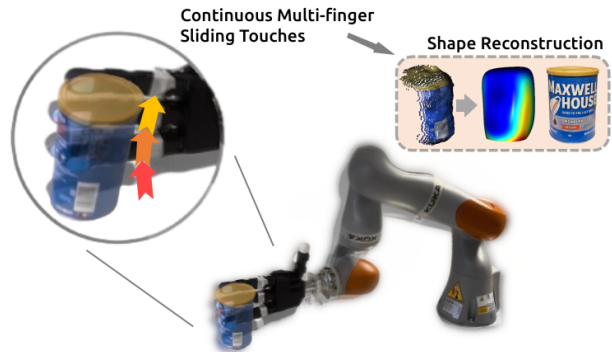
Fig. 1. We propose a visuo-tactile perception approach based on multi-finger sliding touch for unknown object modeling. A few sliding touches are illustrated in the upper left corner, using a multi-finger hand visualized in gray for the first touch and black for the final touch. The hand is continuously moved in between these touches following the surface normal, reducing the uncertainty and increasing the accuracy of the initial object model that is fit based on visual data only. The initial incomplete point cloud obtained from the vision sensor, the final complete reconstructed object model (where blue color corresponds to low uncertainty), and the ground truth model respectively are seen in the upper right corner.

However, how to efficiently fuse both tactile and visual information for reconstructing an unknown object's shape sufficiently well for subsequent manipulation planning still remains an open problem. Tactile perception has been complementary to visual perception in most multi-modal settings. Existing methods represent tactile feedback as discrete points or a few pixels [16]–[20] during manipulation tasks. [21], [22] propose to use tactile images to predict local shape information, which further contributes to global shape reconstruction. However, all of the above methods require frequent re-planning of the robotic arm motion and are time-consuming in the exploration process because the robot needs to constantly re-establish new contact with objects.

Compared with discretely sampled interactions, tactile servoing aims at continuously applying a certain pressure on target objects to perform interaction tasks [23], which allows robots to extract the geometric features smoothly. Tactile servoing includes important tasks, such as sliding a fingertip across an object's surface [24], which brings inspiration to robotic surface exploration. Previous works [25]–[28] have demonstrated that the robot can robustly follow the target object surface with one tactile sensor and further provides shape information such as the object's contour. Though tactile servoing with multi-finger hand has been studied in robotic grasping [29], the aforementioned tactile servoing methods for objects' surface exploration are only using one single tactile sensor.
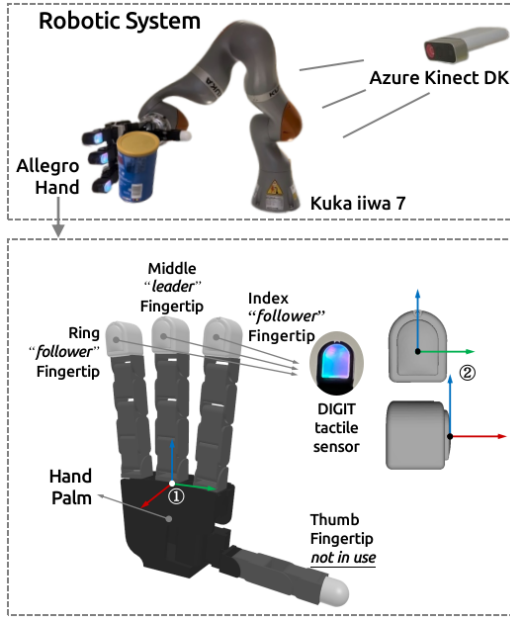
Fig. 2. Robot setup, composed of a KUKA IIWA arm, Allegro Hand equipped with Digit tactile sensors, and an Azure Kinect scene camera. Frame 1 is rigidly attached to the hand palm to represent the palm pose and Frame 2 shows how local frames are rigidly attached to each fingertip.

One of the key challenges still remaining is how to fully exploit the potential of robotic multi-finger hands for human-like tactile perception. Moreover, under the condition that the coverage of the tactile sensor is limited by its size, how to make full use of the dexterity of each robotic fingertip to sense the target object will contribute to perception efficiency significantly. To tackle the above issues, we propose a single-leader-multi-follower sliding touch strategy for multi-finger hands, which takes every fingertip into account for surface exploration.

This paper presents a novel framework for reconstructing the 3D shape of unknown objects through visuo-tactile perception using a multi-finger robotic hand. First, we leverage the power of domain randomization from the tactile sensor simulator to learn a direct mapping between the tactile images from the DIGIT tactile sensors to the corresponding height maps. Then we propose a single-leader–multi-follower strategy for efficient tactile perception based on tactile servoing. The "leader" fingertip is responsible for providing tactile features to lead the robotic hand sliding servoing movement while the "follower" fingertips are responsible for enlarging the contact field. A fingertip adaptation strategy is adopted for "follower" fingers, aiming at obtaining a larger sensing area during contact. We model the target object using a probabilistic representation, Gaussian process implicit surfaces (GPIS) [30], which can be further combined with query point guidance from Bayesian optimization (BOpt) [20], [28], [31] for sliding touch exploration. We show that combined visual perception and multi-finger tactile sensing can be performed efficiently (shown in Fig. 1), which significantly reduces the robotic arm motions.

To the best of our knowledge, this is the first study that considers the continuous motion of a multi-finger hand in visuo-tactile perception for the purpose of unknown object modeling through exploration. We present both simulated and real experiments, validating our proposed method of modeling unknown objects' global shapes with continuous robotic arm motion minimizing the required actions to reconstruct object models similar to the ground truth data.

Our main contributions can be summarized as follows: **i)** We propose an efficient visuo-tactile perception approach based on continuous sliding touches with multi-finger hands. Our approach provides an efficient way for surface exploration and significantly reduces the robotic arm motion and the required execution time. **ii)** We present a hand-agnostic single-leader-multi-follower hand control strategy to combine tactile servoing and fingertip adaptation for smooth tactile sensing. This strategy fully exploits multi-finger hand dexterity for maximizing the local contact area being perceived. **iii)** We demonstrate that the visuo-tactile perception for unknown object modeling can be performed with limited arm motion in a smooth manner without re-establishing contact.

II. System Setup and Problem Formulation

A. System Setup

Fig. 2 shows the multi-finger hand used in our experiments to evaluate the proposed sliding touch approach. Each of the index, middle, and ring fingertips of the Allegro hand is equipped with DIGIT tactile sensor. We define the middle fingertip as the "leader" and the index and ring fingertip as the "follower", a detailed strategy will be introduced in Section IV. The thumb fingertip is not in use for tactile sensing because we do not construct any force-closure configurations during exploration and only explore from the side of the object that is not visible to the scene camera. An impedance controller [32] is deployed for finger control, which provides passive position-force adjustment during the interaction. Fingers are independent during contact and do not affect each other. The contact from fingertips cannot affect the hand palm's pose. The initial hand finger configuration is shown in Fig. 3, a small joint position offset is set to the second joint of each finger. After getting in contact with the target object, the hand finger will passively move near the zero position which causes an additional force to press on the target object.
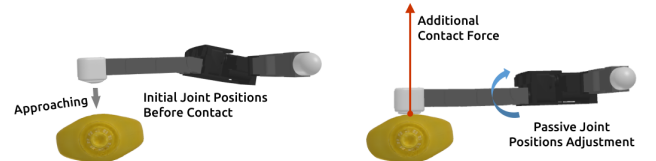

Fig. 3. The initial hand configuration provides passive joint space for additional contact force between fingertips and the target object. During the sliding process, the hand configuration is remained around zero-position to apply a constant contact force.

B. Problem Formulation

We address the problem of how to use a partial view of the depth camera and the sequential robotic tactile sliding touches to model the shape of a given unknown object. We consider that the underlying true shape can be represented using an implicit "black-box" scalar function f, which defines a manifold

$$\mathcal{S} := \{x \in \mathbb{R}^3 | f(x) = 0\}. \quad (1)$$

We cannot evaluate the derivatives of f, but only observe $f(x)$ from visual and tactile perception. As for the visual perception, we use the depth image $\mathbf{D} \in \mathbb{R}^{H_1 \times W_1}$ from the partial view of the target object and the camera intrinsic and extrinsic matrix, by transferring $\mathbf{D}$ into point cloud $\mathbf{P_d^w} = \{x_1^d, x_2^d, ..., x_k^d\} \in \mathbb{R}^3, k \leq H_1 \cdot W_1$ w.r.t the world frame after filtering with object segmentation mask, $f(x)$ can be observed as

$$f(x_i^d) = 0, \quad x_i^d \in \mathbf{P_d^w}. \quad (2)$$

The unseen part of the object is perceived based on touch sensing. Given the tactile image $\mathbf{I} \in \mathbb{R}^{H_2 \times W_2 \times 3}$ from the tactile sensor, we can estimate the height map of local contact $\mathbf{I} \to \mathbf{M} \in \mathbb{R}^{H_2 \times W_2}$. Based on robotic forward kinematics and sensor intrinsic matrix, the height map is transferred from the fingertip's image frame to the tactile point cloud $\mathbf{M} \to \{x_1^w, x_2^w, ..., x_n^w\}, n \leq H_2 \cdot W_2$ w.r.t the world frame after filtering with the binary contact area. We represent the motion of one sliding touch $\mathbf{T} = \{I_1, I_2, I_3, ..., I_m\} \to \{M_1, M_2, M_3, ..., M_m\}$ as a series of consecutive touch frames. The tactile point cloud $\mathbf{P_t^w}$ from one complete sliding touch is obtained as

$$\mathbf{P_t^w} = \{x_1^t, x_2^t, ..., x_k^t\} \in \mathbb{R}^3, k \leq m \cdot H_2 \cdot W_2. \quad (3)$$

The value of $f(x)$ is given by

$$f(x_j^t) = 0, \quad x_j^t \in \mathbf{P_t^w}. \quad (4)$$

Based on the observations $f(x)$, we use Gaussian Process Regression (GPR) to build an implicit surface (GPIS) representation to estimate our target function f. The GPIS also provides a probabilistic description to further combine with BOpt to guide our robotic fingertip tactile perception. The sliding touch strategy will output sequential tactile point clouds $\mathbf{P_t^w}$ after each BOpt iteration for continuous $f(x)$ observation and leads to GPIS with lower uncertainty and more thorough modeling.

The rest of the paper is organized as follows: Section III presents the fingertip perception method for contact shape estimation and tactile feature extraction. Section IV introduces the sliding touch strategy for efficient and smooth visuo-tactile perception. Section V presents experiments using sliding touches with simulation and a real robot, followed by conclusions in Section VI.

III. FINGERTIP TACTILE PERCEPTION

A. Local Shape Estimation from Tactile Image

The vision-based tactile sensor mainly consists of an embedded camera, an inner light source, and its surface gel [6], [10]. The output tactile images reflect the surface deformation of the gel pad under special light conditions. We consider the task of recovering the height map $\mathbf{M} \in \mathbb{R}^{H_2 \times W_2}$ from the gel image $\mathbf{I} \in \mathbb{R}^{H_2 \times W_2 \times 3}$ as a similar problem of monocular depth prediction.
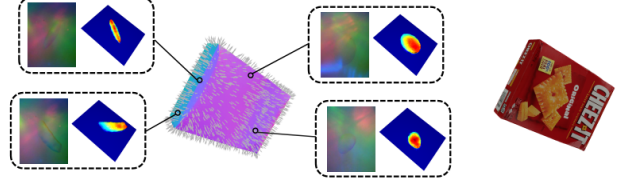

Fig. 4. Four rendered tactile image-height map pairs with different calibrated backgrounds from real-world DIGIT sensors. We randomly select surface normal as our reference direction and sample the contact surface with angle noise and randomized depth. In total, we generate 75,000 pairs of labeled data from 25 different objects for depth estimator training.

Data Collection in Simulation: We learn the depth estimator in a supervised fashion with paired tactile images and their corresponding height maps. Since it is difficult to collect ground truth height maps in the real world, we collect the training data with TACTO [11], a flexible tactile sensor simulator with background calibration. We first calibrate the tactile simulator with different sensor backgrounds and uniformly sample tactile images across the object's surface. To simulate the real tactile perception situation, we sample the contact with different depth values $X \sim U[-0.2\,\mathrm{mm}, -1.2\,\mathrm{mm}]$ along surface normals with an orientation noise angle $\theta \sim \mathcal{N}(0, 30°)$. Rendered tactile images with different calibrated sensor backgrounds and corresponding height maps are shown in Fig. 4. We sample 4000 labeled tactile images on each of 25 objects' mesh from the YCB dataset.

Network Evaluation: We adapted a modified fully convolutional residual network (FCRN) [33] as our depth estimator, which uses a ResNet-50 as the backbone network and up-sampling blocks as the regression head. The depth estimator is responsible for $\mathbf{I} \to \mathbf{M}$ translation. After 15 epochs of training with a learning rate of $2\mathrm{e}^{-5}$, the prediction result in the real world and the network structure is shown in Fig. 5.

B. Tactile Feature Representation and Extraction

We extract the tactile feature t_f from the estimated height map. Consider a height map $\mathbf{M} \in \mathbb{R}^{h \times w}$ as demonstrated in Fig. 6, we aim at extracting the center of intensity and the size of contact area as our tactile feature. The center of tactile intensity $(\overline{x}, \overline{y})$ is related to the contact pressure and the shape of the contact area which can be obtained as

$$(\overline{x}, \overline{y}) = \frac{\sum_{i=1}^{h} \sum_{j=1}^{w} (x_i, y_j) m_{i,j}}{\sum_{i=1}^{h} \sum_{j=1}^{w} m_{i,j}} \quad (5)$$

in which x and y denote the pixel index for columns and rows. The size of the contact area c is computed by the total number of pixels with a value larger than 10^{-4}. Finally, the tactile feature vector is concatenated in 3 dimensions as
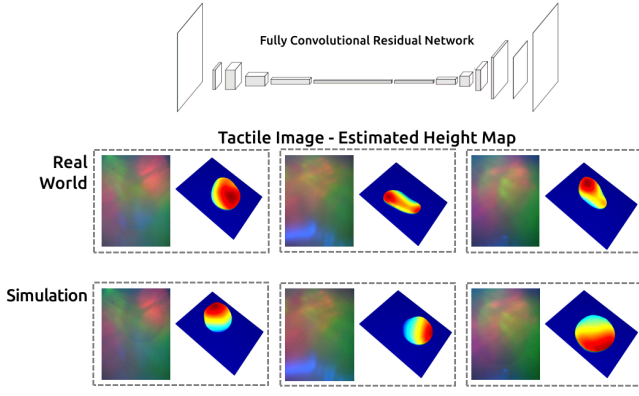
Fig. 5. A fully convolutional residual network is adopted as our depth estimator. The upper line image pairs show the height map prediction results on real-world tactile images from different DIGIT sensors, and the bottom line image pairs show prediction results from simulated tactile images.

follows:

$$t_f = \begin{bmatrix} \overline{x} \\ \overline{y} \\ c \end{bmatrix} = \begin{bmatrix} \text{X center of contact intensity} \\ \text{Y center of contact intensity} \\ \text{contact area} \end{bmatrix} \quad (6)$$

The tactile features extracted from fingers with different roles will be used to perform different tasks as follows:

- **From the "leader" fingertip**: Tactile feature will be used for hand palm twists and sliding direction calculation.
- **From the "follower" fingertips**: Tactile feature will be used for adapting the fingertips' pose for a larger contact field for tactile sensing.

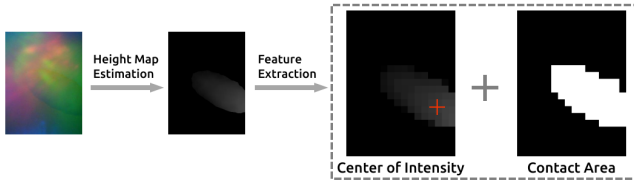The detailed mapping will be introduced in Section IV-B.

Fig. 6. After estimating the height map from the tactile image, we extract the center of tactile intensity and contact area as our tactile feature for further servoing. The corresponding height map is resized into 32×24 for better visualization. The red cross in the left image from the dotted box denotes the extracted center of intensity and the right image from the dotted box demonstrates the binary contact area.

IV. SLIDING TOUCH FOR SHAPE RECONSTRUCTION

The pipeline of our proposed framework is illustrated in Fig. 7. By fusing P_{ti}^w with observed $P^w = \{P_d^w, P_{t1}^w, P_{t2}^w, ..., P_{t(i-1)}^w\}$ after each sliding touch, we are able to observe the target object's implicit function as introduced in Section II and update our GPIS. The complete exploration process consists of several sliding touches $\{T_1, T_2, ..., T_k\}$, and one sliding touch consists of dozens of touch frames as previously defined in (3).
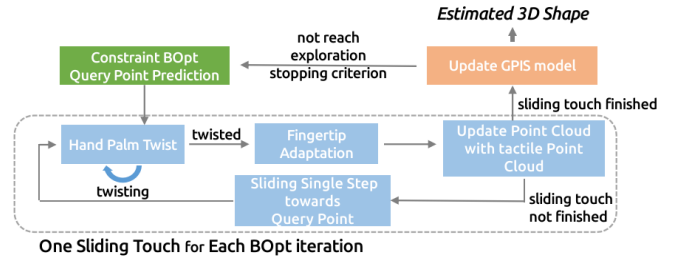

Fig. 7. The online perception pipeline of our multi-finger hand sliding touch strategy. Given the query point from each constrained BOpt iteration, first, the hand palm will twist based on the tactile feature from the "leader" fingertip, and the "follower" fingertips will adapt their pose given the current hand pose for a larger sensing area. After the visuo-tactile point cloud is updated, the hand will slide under the direction of obtained query point with a pre-defined step size.

A. GPIS Shape Representation

Based on the values Y observed from points $X = P^w$, our goal is to reconstruct an implicit surface approximation of the object's shape while also providing uncertainty information for exploration guidance. The value of our GPIS $g^{IS}(x)$ is defined as follows:

$$g^{IS}(x) \begin{cases} < \ 0 & \text{if } x \text{ is } \textbf{below} \text{ the surface} \\ = \ 0 & \text{if } x \text{ is } \textbf{on} \text{ the surface} \\ > \ 0 & \text{if } x \text{ is } \textbf{above} \text{ the surface} \end{cases} \quad (7)$$

which represents a mapping from point cloud to a scalar value $\mathbb{R}^3 \rightarrow \mathbb{R}$. We adopt the RBF[1] (squared-exponential) kernel $k_{\text{RBF}}(\mathbf{x_1}, \mathbf{x_2}) = \exp\left(-\frac{1}{2}(\mathbf{x_1} - \mathbf{x_2})^\top \Theta^{-2}(\mathbf{x_1} - \mathbf{x_2})\right)$ to approximate the target implicit function f, which leads to good reconstruction results for the experiment objects. Given a query point $x_* \in \mathbb{R}^3$, the g^{IS} computes its posterior mean and variance as:

$$\Sigma = (k_{\text{RBF}}(X, X) + \sigma^2 I)^{-1} \quad (8)$$

$$\overline{g}(\boldsymbol{x_*}) = k_{\text{RBF}}(X, \boldsymbol{x_*})^{\text{T}} \Sigma Y \quad (9)$$

$$\mathbb{V}(\boldsymbol{x_*}) = k_{\text{RBF}}(\boldsymbol{x_*}, x)^{\text{T}} \Sigma k_{\text{RBF}}(\boldsymbol{x_*}, x)^{\text{T}} \quad (10)$$

B. Tactile Servoing and Perception using Multi-finger Hand

Our one-leader-multi-follower multi-finger tactile sensing strategy mainly consists of two parts: (1) Hand Palm twist for tactile servoing, which is based on the tactile feature provided by the **"leader"** fingertip, and (2) **"follower"** fingertips adapt their poses for larger sensing area based on their tactile features.

Palm Twist for Tactile Servoing: Based on the tactile feature $t_f = [\overline{x}, \overline{y}, c]^T$ extracted from the "leader" fingertip, the goal of this phase is to minimize the distance $\Delta t_f = [\Delta x, \Delta y, \Delta c]^T$ between current t_f and reference $t_{fr} = [\frac{h}{2}, \frac{w}{2}, c_r]^T$ only by hand twisting. *The joints of each finger will remain unchanged.* The hand twist motion

---

[1]For rectangular objects the kernel can be replaced by the thin-plate kernel.
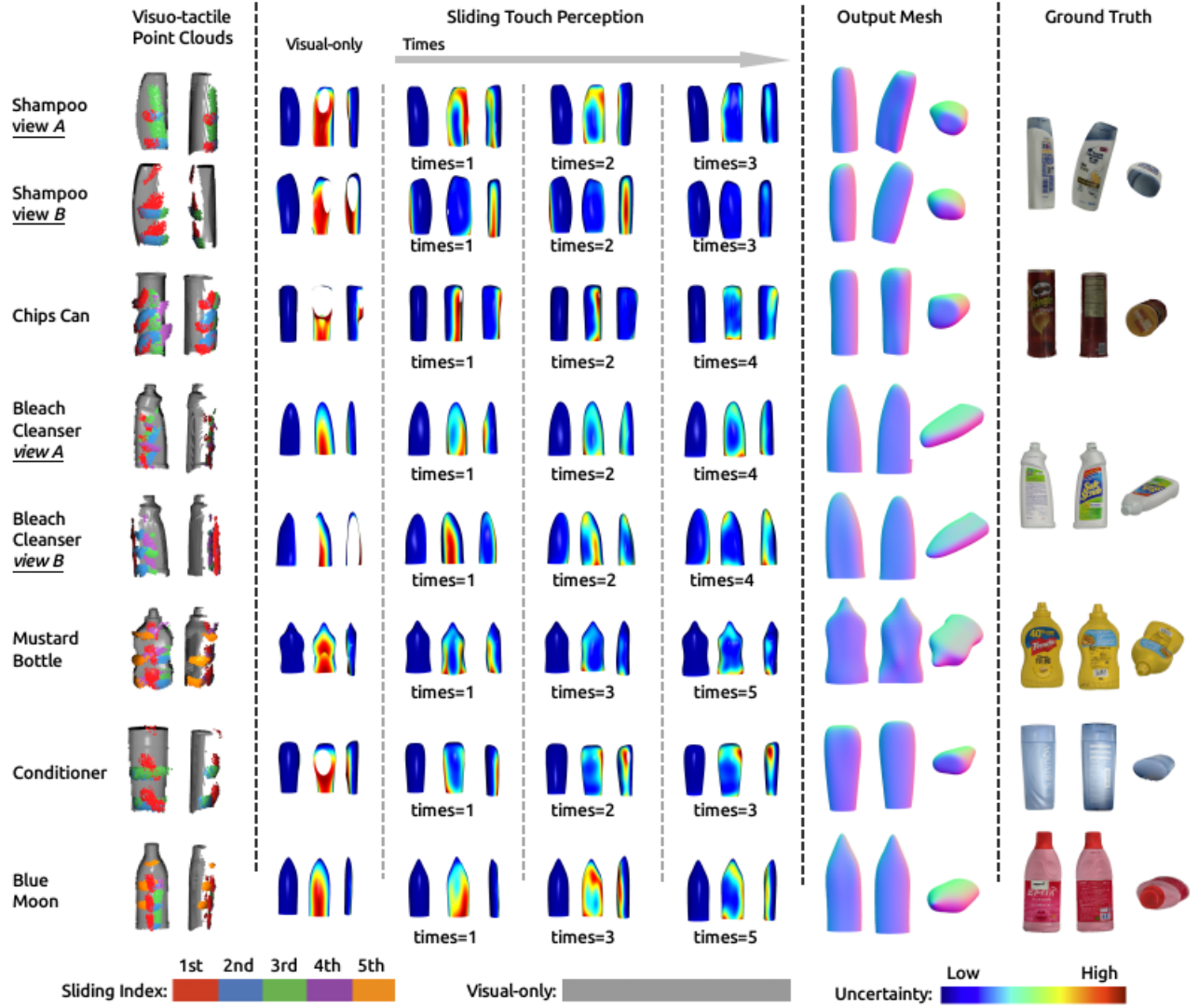
Fig. 8. The shape exploration and reconstruction process of 6 objects in the simulation. From left to right respectively: (I) the final combined visuo-tactile point cloud, where grey colored points represent the visual point cloud, and tactile point cloud generated by each sliding touch are colored with corresponding color map, (II) how the uncertainty map evolves during the sliding process (III) the final reconstructions and (IV) the ground truth models.

$V_{palm} = [\dot{\theta}_y, \dot{\theta}_z, \dot{d}_x]^T$ is mapped from Δt_f by a *task-dependent* tactile Jacobian J_s and *task-dependent* selection matrix P as introduced by [24]. In which R_y and R_z denote hand palm rotation angle along the Y axis and Z axis respectively and translation distance along the X axis w.r.t. the "leader" fingertip frame.

$$J_s^{palm} = \begin{bmatrix} \frac{\partial \overline{x}}{\partial \theta_z} & 0 & 0 \\ 0 & \frac{\partial \overline{y}}{\partial \theta_y} & 0 \\ 0 & 0 & \frac{\partial c}{\partial d_x} \end{bmatrix} \quad P = \begin{bmatrix} i & 0 & 0 \\ 0 & i & 0 \\ 0 & 0 & j \end{bmatrix} \quad (11)$$

We set $i = 0, j = 1$ for translation mapping and $i = 1, j = 0$ for rotation mapping. The core mapping is formulated as

$$V_{palm} = P J_s^\dagger \Delta t_f \quad (12)$$

where † denotes the pseudo inverse. A PID controller is adopted for decreasing the Δt_f.

Fingertip Adaptation for Tactile Perception: Based on the tactile feature extracted from the "follower" fingertips, the goal of this phase is to sense a larger contact area by adjusting the "follower" fingertips' pose only by moving the corresponding finger joint position. *The hand palm's pose will remain fixed at this moment.* Different from the above tactile servoing phase, due to the joint control accuracy and manipulability of the Allegro hand, this fingertip adaptation motion will directly change the fingertips' pose without feedback control and only be triggered if the fingertip is in contact. A similar task-dependent Jacobian and selection matrix as (11) is pre-defined for fingertip pose adaptation.

C. Constrained Bayesian Optimization for Sliding Touch

Bayesian Optimization (BOpt) is used to optimize black-box functions that are expensive to evaluate. In the context of this paper, the target (black-box) function is based on the implicit function that represents the object surface. BOpt builds a (probabilistic) surrogate model, GPIS, of the target function and uses an acquisition function to determine where

to evaluate next. We approximate the target function by optimizing our GPIS through sequential tactile perception.

Surrogate Model: The surrogate model is a statistical model to describe the target function based on observations, and also provides predictive uncertainty. In our case, the GPIS is the surrogate model for the target function approximation and description. Given a query point x_q from the world frame, our surrogate model provides a mean estimate $\bar{g}(x_q)$ and the corresponding variance $\mathbb{V}(x_q)$. The mean value defines whether or not the given query point is on the object's surface based on (7) and the variance value denotes the prediction uncertainty.

Acquisition Function: An acquisition function is designed for balancing the exploration of new parameters vs the exploitation of current knowledge based on previous experiments. We choose Expected-Improvement (EI) as our acquisition function:

$$\text{EI}(x) = \mathbb{E}(\max(y - g_{best}, 0)), \quad y \sim g(x), \quad (13)$$

where g_{best} is the best-seen value of the black-box function. The next query point is computed by

$$x_{n+1} = \arg\max_x \text{EI}_n(x) \quad (14)$$

EI calculates how much its function value can be expected to improve over our current optimum for every point from our constrained search space. Then we will choose the point with maximal improvement as our query point for the next sliding touch guidance.

Constrained Search Space: To improve the efficiency of sampling query points and well integrate with the hand sliding motion, the search space is updated with the GPIS for each iteration. The search space S^x for each sampling is constrained by the implicit surface

$$S^x_{n+1} = \{x \mid \arg g^{\text{IS}}_n(x) = 0 \wedge \|x - x^{ob}_n\| > d\} \quad (15)$$

where $d = 0.005$ denotes the minimized length between points from search space and observed points.

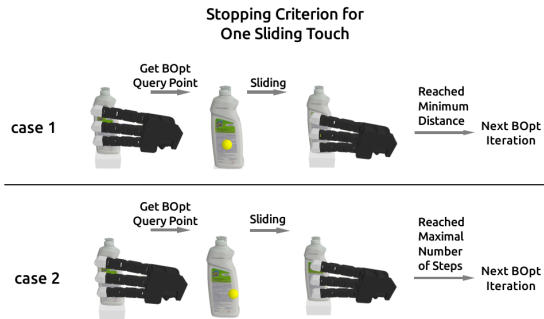### D. Surface Exploration and Exploitation

Fig. 9. Two different examples of how stopping criterion is triggered.

BOpt is designed to output a query point $x_{BO} \in \mathbb{R}^3$ with respect to the world frame in each iteration. Combined with the current "leader" fingertip's tactile feature, the output x_{BO} will guide the robotic hand's sliding servoing direction

to explore the target object's surface. Given the estimated coordinate of the center of intensity $x_c = [i, j, v]^T \in \mathbb{R}^3$ and its corresponding surface normal $n \in \mathbb{R}^3$ from the "leader" fingertip w.r.t. the world frame, the sliding direction t_{BO} is the projection of $x_{BO} - x_c$ on the estimated surface (defined by normal n), which is calculated as

$$t_{BO} = (x_{BO} - x_c) - \frac{(x_{BO} - x_c) \times n}{n \times n} n. \quad (16)$$

The translation for the next frame is computed as

$$x_{next} = x_c + \frac{t_{BO}}{\|t_{BO}\|} \Delta step, \quad (17)$$

where the $\Delta step = 0.005$ is the step size between each sliding frame and needs to be set small enough to keep the "leader" fingertip in contact with the target object.

Sliding Exploration for each BOpt iteration: One sliding touch consists of sequential touch frames as introduced in (II-B). Each frame is obtained from every single step, and the total number of frames for the ongoing sliding touch is determined by (1) the distance between the current center of contact intensity from the "leader" fingertip and the query point smaller than the minimal value $D_{min} = 0.01$, (2) the number of frames reached the maximal boundary $N_{max} = 15$. If one of the two conditions is met, the ongoing sliding touch will terminate and enter the next BOpt iteration as shown in Fig. 9.

Stopping Criterion for Surface Exploration: The goal of surface exploration is to estimate the target object's shape with relatively low global uncertainty. This judgment is based on the uncertainty distance reduction ratio (UDRR) of our GPIS model. Since the GPIS yields low uncertainty in the areas with the visual point cloud, the global estimation is evaluated based on the uncertainty distance compared with areas with a partial view. Given the initial GPIS learned from only the visual point cloud, the initial uncertainty distance is calculated as:

$$dist_{max} = \max(\mathbb{V}_{init}(x_m) - \mathbb{V}_{init}(x_n)), \quad x_m, x_n \in \mathbf{P^w_d} \quad (18)$$

After the sliding touch of each BOpt iteration is over, the outcome UDRR is calculated as follows

$$\text{UDRR} = \frac{\max(\mathbb{V}(x_m) - \mathbb{V}(x_n))}{dist_{max}}, \quad x_m, x_n \in \mathbf{P^w} \quad (19)$$

If UDRR $< 30\%$, we consider the target object has been explored sufficiently and the sliding perception is over.

V. EXPERIMENTAL VALIDATIONS

We evaluate our proposed sliding strategy in simulation (Section V-A) and the whole pipeline in the real world (Section V-B). We set up our robotic platform with a Kuka iiwa 7 with Allegro Hand as shown in Fig. 2. An Azure Kinect DK depth camera is used as a scene camera for visual perception. All test objects are from the YCB [34] and the OCRTOC benchmark datasets [35]. The shape estimation performance is quantified using a shape similarity metric, the Chamfer distance (CD) [36].

Implementation: The whole pipeline is executed on an Intel Core i7-11800H CPU, 16 RAM with an NVIDIA GeForce RTX 3070 GPU. To balance the point cloud density from visual and tactile perception, the tactile point cloud is randomly down-sampled with a ratio of 0.02. The input point cloud is finally down-sampled to 4500 points. During the GPIS evaluation phase, we set the grid size with a value of 10 for surface reconstruction.
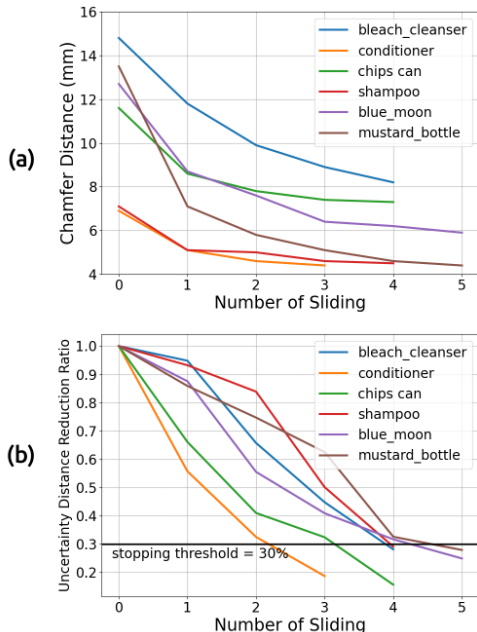

Fig. 10. The figure illustrates shape estimation accuracy (a) and the UDRR (b) of 6 objects in simulation with respect to the number of sliding touches.

A. Sliding Simulation

We tested our sliding touch strategy in simulation with 6 different objects, which are of varying shapes, curvature, and sizes to verify the generalization of the proposed method. The performance of our method is evaluated by (1) the accuracy of estimated meshes and (2) the efficiency of object exploration. The sliding touch exploration is visualized in Fig. 8, which shows the tactile point cloud obtained from every single sliding touch colored differently for clarity in the first column. The evolution of the global uncertainty and estimated mesh during the perception process can be seen in the middle columns. For objects with non-symmetric shapes, we tested our method under two different camera views (the rotation angle between view A and view B is larger than $60°$) for a comprehensive evaluation. Our method output the estimated shape with considerably reduced uncertainty, with respect to the initial model built based on only visual data, only using a few continuous touches. The change in the uncertainty map clearly demonstrates that each sliding touch increases the information gain and helps to efficiently capture the surface's underlying geometry. The accuracy of the estimated shape calculated using CD is demonstrated in Fig. 10 (a) and the evolution of UDRR is shown in Fig. 10 (b). As the number of sliding touches increases, CD w.r.t

Objects	Avg., Std. Times[1]		Avg., Std. CD[2]	
	BOpt (Ours)	Random*	BOpt (Ours)	Random*
Shampoo[b]	**3.3, 0.48**	4.4, 1.43	4.33, **0.11**	**4.29**, 0.14
Chips Can[a]	**4.5, 0.67**	6.2, 1.54	**7.52**, 0.23	7.57, **0.21**
Bleach Cleanser[a]	**4.7, 0.64**	5.7, 1.10	8.25, 0.31	**8.21, 0.24**
Conditioner[b]	**3.2, 0.40**	5.3, 1.50	4.45, **0.17**	**4.34**, 0.20
Blue Moon[b]	**5.0, 0.45**	7.2, 0.98	6.03, 0.34	**5.98, 0.27**
Mustard Bottle[a]	**4.8, 0.75**	6.7, 0.90	**4.72, 0.45**	4.77, 0.51

[1] Sliding touch times to achieve UDRR $< 30\%$ from 10 experiments.
[2] Output shape's Chamfer Distance (mm) w.r.t GT from 10 experiments.
[a] and [b] respectively denotes from the YCB and OCRTOC benchmark.
* The constrained random policy shares the same constrained search space.

the ground truth model decreases for each object. For those objects with irregular or non-symmetric shapes in particular, e.g. mustard bottle and bleach cleanser, in which GPIS has a relatively bad initial estimation due to partial and vision-only observation points, it is seen from a large amount of decrease in CD that our proposed method can efficiently explore the areas (not visible to the camera) and provides sufficient estimated shape correction.

We compare the efficiency of our method with the constrained random policy. To fairly evaluate both policies, *the search space for the constrained random policy is set the same as our BOpt*. The exploration is ended by the same stopping criterion with a value of UDRR lower than 30%. And the comparison result is listed in Table I. Even though both policies under the constrained search space can perform full exploration on the target object's region, our proposed BOpt approach achieves similar accuracy with a significantly reduced required number of sliding touches. The smaller standard deviation of BOpt iteration times also demonstrates that our method is more stable compared to the baseline.

B. Real-world Experiment

We tested our proposed method with real-world objects from the YCB dataset. We selected three objects with different materials, including symmetric and non-symmetric shapes with different sizes. The initial approach position is set on the left side of the target object w.r.t camera frame, and the example of one complete sliding sensing is shown in Fig. 11 (a). Fig. 11 (b) shows the qualitative result of the shape reconstruction on three objects. Due to safety concerns for avoiding collision with the table, the coordinate of the query point will be limited with a Z value between 0.05 and 0.15 w.r.t the world frame. The proposed method yields reconstructed shapes for unknown objects with low uncertainty and detailed curvature, capturing the main geometric details in explored areas. However, there are still some areas that are not well explored due to arm reachability, which are left with relatively high uncertainty.

VI. CONCLUSIONS

We present a novel approach for unknown object modeling based on multi-finger hand sliding touches. To efficiently address this problem, an essential aspect is to fully exploit the local shape sensing ability of each fingertip and integrate
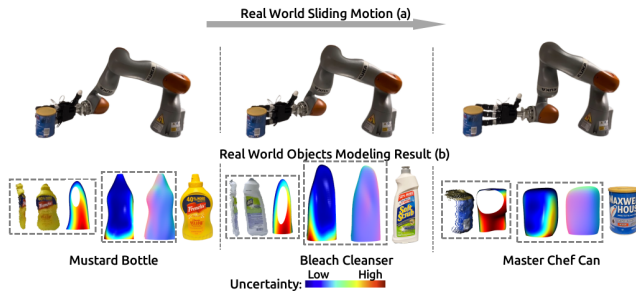
Fig. 11. Figures from the top line show the sliding motion of one complete experiment on Master Chef Can. Figures from the bottom line show the visual-only point cloud with the initially estimated shape, the final estimation result with an uncertainty map, and the corresponding real-world object separately for each object. For objects from left to right, the exploration process takes respectively 5, 4, and 4 sliding touches for the output mesh.

them in a continuous and smooth manner. Our method yields successful reconstruction results capturing the underlying shape in visually unseen areas with only a few continuous sliding touches. However, our work relies on several assumptions due to the limitation of the real setup. First, the objects are fixed on the table to be able to apply a larger contact force to generate sufficient deformation on the gel of tactile sensors. Second, the explored region on the object's surface is considered low curvature to meet the limited reachability of the robotic arm. In future works, we will combine the proposed approach with a mobile manipulator which provides a larger workspace and sensors with higher sensitivity and precision to address the above issues.

REFERENCES

[1] K. Fu, J. Peng, Q. He, and et al., "Single image 3d object reconstruction based on deep learning: A review," *Multimedia Tools and Applications*, vol. 80, pp. 463–498, 2021.

[2] H. B. Helbig and M. O. Ernst, "Optimal integration of shape information from vision and touch," *Experimental brain research*, vol. 179, no. 4, pp. 595–606, 2007.

[3] R. S. Dahiya, G. Metta, M. Valle, and et al., "Tactile sensing—from humans to humanoids," *IEEE T-RO*, vol. 26, no. 1, pp. 1–20, 2009.

[4] C. Yang, S. Luo, N. Lepora, and et al., "Biomimetic perception, cognition, and control: From nature to robots [from the guest editors]," *IEEE Robotics & Automation Magazine*, vol. 29, no. 4, pp. 8–8, 2022.

[5] G. Cheng, E. Dean-Leon, F. Bergner, and et al., "A comprehensive realization of robot skin: Sensors, sensing, control, and applications," *Proceedings of the IEEE*, vol. 107, no. 10, pp. 2034–2051, 2019.

[6] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[7] B. Ward-Cherrier, N. Pestell, L. Cramphorn, and et al., "The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies," *Soft robotics*, vol. 5, no. 2, pp. 216–227, 2018.

[8] A. Alspach, K. Hashimoto, N. Kuppuswamy, and et al., "Soft-bubble: A highly compliant dense geometry tactile sensor for robot manipulation," in *IEEE RoboSoft*, 2019, pp. 597–604.

[9] E. Donlon, S. Dong, M. Liu, and et al., "Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger," in *IEEE/RSJ IROS*, 2018, pp. 1927–1934.

[10] M. Lambeta, P.-W. Chou, S. Tian, and et al., "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE RA-L*, vol. 5, no. 3, pp. 3838–3845, 2020.

[11] S. Wang, M. Lambeta, P.-W. Chou, and et al., "Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors," *IEEE RA-L*, vol. 7, no. 2, pp. 3930–3937, 2022.

[12] Z. Chen, S. Zhang, S. Luo, and et al., "Tacchi: A pluggable and low computational cost elastomer deformation simulator for optical tactile sensors," *IEEE RA-L*, 2023.

[13] Z. Si and W. Yuan, "Taxim: An example-based simulation model for gelsight tactile sensors," *IEEE RA-L*, vol. 7, no. 2, pp. 2361–2368, 2022.

[14] D. F. Gomes, A. Wilson, and S. Luo, "Gelsight simulation for sim2real learning," in *ICRA ViTac Workshop*, 2019.

[15] A. Agarwal, T. Man, and W. Yuan, "Simulation of vision-based tactile sensors using physics based rendering," in *IEEE ICRA*, 2021.

[16] M. Björkman, Y. Bekiroglu, V. Högman, and et al., "Enhancing visual perception of shape through tactile glances," in *IEEE/RSJ IROS*, 2013, pp. 3180–3186.

[17] G. Z. Gandler, C. H. Ek, M. Björkman, and et al., "Object shape estimation and modeling, based on sparse gaussian process implicit surfaces, combining visual data and tactile exploration," *Robotics and Autonomous Systems*, vol. 126, p. 103433, 2020.

[18] J. Ilonen, J. Bohg, and V. Kyrki, "Three-dimensional object reconstruction of symmetric objects by fusing visual and tactile sensing," *IJRR*, vol. 33, no. 2, pp. 321–341, 2014.

[19] J. Mahler, S. Patil, B. Kehoe, and et al., "Gp-gpis-opt: Grasp planning with shape uncertainty using gaussian process implicit surfaces and sequential convex programming," in *IEEE ICRA*, 2015, pp. 4919–4926.

[20] C. de Farias, N. Marturi, R. Stolkin, and Y. Bekiroglu, "Simultaneous tactile exploration and grasp refinement for unknown objects," *IEEE RA-L*, vol. 6, no. 2, pp. 3349–3356, 2021.

[21] S. Suresh, Z. Si, J. G. Mangelson, and et al., "Shapemap 3-d: Efficient shape mapping through dense touch and vision," in *IEEE ICRA*. IEEE, 2022, pp. 7073–7080.

[22] S. Wang, J. Wu, X. Sun, and et al., "3d shape perception from monocular vision, touch, and shape priors," in *IEEE/RSJ IROS*. IEEE, 2018, pp. 1606–1613.

[23] Q. Li, O. Kroemer, Z. Su, and et al., "A review of tactile information: Perception and action through touch," *IEEE T-RO*, vol. 36, no. 6, pp. 1619–1634, 2020.

[24] Q. Li, C. Schürmann, R. Haschke, and et al., "A control framework for tactile servoing." in *RSS*, 2013.

[25] J. Kerr, H. Huang, A. Wilcox, and et al., "Learning self-supervised representations from vision and touch for active sliding perception of deformable surfaces," *arXiv preprint arXiv:2209.13042*, 2022.

[26] N. F. Lepora, K. Aquilina, and L. Cramphorn, "Exploratory tactile servoing with active touch," *IEEE RA-L*, vol. 2, no. 2, pp. 1156–1163, 2017.

[27] Q. Li, R. Haschke, and H. Ritter, "A visuo-tactile control framework for manipulation and exploration of unknown objects," in *IEEE-RAS Humanoids*, 2015, pp. 610–615.

[28] D. Driess, P. Englert, and M. Toussaint, "Active learning with query paths for tactile object shape exploration," in *IEEE/RSJ IROS*, 2017, pp. 65–72.

[29] H. Liu, B. Huang, Q. Li, and et al., "Multi-fingered tactile servoing for grasping adjustment under partial observation," in *IEEE/RSJ IROS*, 2022, pp. 7781–7788.

[30] Z. Murvanidze, M. P. Deisenroth, and Y. Bekiroglu, "Enhanced gpis learning based on local and global focus areas," *IEEE RA-L*, vol. 7, no. 4, pp. 11 759–11 766, 2022.

[31] C. Pan, M. Lepert, S. Yuan, and et al., "Task-driven in-hand manipulation of unknown objects with tactile sensing," *arXiv:2210.13403*, 2022.

[32] M. Li, Y. Bekiroglu, D. Kragic, and et al., "Learning of grasp adaptation through experience and tactile sensing," in *IEEE/RSJ IROS*, 2014, pp. 3339–3346.

[33] I. Laina, C. Rupprecht, V. Belagiannis, and et al., "Deeper depth prediction with fully convolutional residual networks," in *3DV*, 2016, pp. 239–248.

[34] B. Calli, A. Singh, A. Walsman, and et al., "The ycb object and model set: Towards common benchmarks for manipulation research," in *ICAR*, 2015, pp. 510–517.

[35] Z. Liu, W. Liu, Y. Qin, and et al., "Ocrtoc: A cloud-based competition and benchmark for robotic grasping and manipulation," *IEEE RA-L*, vol. 7, no. 1, pp. 486–493, 2021.

[36] H. G. Barrow, J. M. Tenenbaum, and e. a. Bolles, "Parametric correspondence and chamfer matching: Two new techniques for image matching," in *Image Understanding Workshop*, 1977, pp. 21–27.