MOE-Touch More Deformation: Shape-Based Soft Robotic Contact Estimation for Manipulation

Anonymous Author(s)
Affiliation
Address
email

Abstract: Contact-rich interaction with the world is crucial for many challenging robot manipulation tasks, such as handling delicate objects or providing physical assistance to humans. Unlike commonly used rigid manipulators, soft robotic manipulators can interact safely and robustly with large distributed contact with the world. However, contact sensing for soft robots has been difficult because embedding sensors into soft bodies introduces rigidity, which undercuts the benefits of such compliant systems. In this paper, we present MOE-Touch, a method that reasons about contact conditions for soft robots by observing deformation. We introduce and test the idea that contact conditions and contact object geometry can be inferred by observing contact deformations in a compliant and soft robot manipulator. We propose Multi-finger Omnidirectional End-effector (MOE), a soft manipulator capable of safely interacting with delicate surfaces. We use a mesh energy optimization-based method for multi-shape estimation of MOE's deforming state. We then use a Graph Neural Network (GNN)-based contact estimation module to predict distributed contact locations from deformation. MOE-Touch can accurately estimate contact with 3.03 mm Chamfer distance error, which is a 50.65 % improvement on the baseline. We then demonstrate an application of MOE-Touch shape estimation and contact localization modules for the reconstruction of an occluded surface modeled as Gaussian Process Implicit Surfaces (GPIS) with averaged errors of 3.62 mm, and showcase the application of using MOE-Touch for grasping a piece of paper on a flat surface with an unknown orientation.

Keywords: Soft Robotics, Contact Estimation, Manipulation

1 Introduction

Humans often make large distributed contact with objects in our daily lives. Such distributed contact-rich interaction with the world can serve two purposes. First, it enables us to perceive occluded surfaces and understand the underlying object geometry. For example, a hairstylist can pat a customer's head to estimate the contour of the scalp underneath the voluminous hair and select feasible hairstyles. Second, manipulating certain objects unavoidably results in large contact. We can consider the example of picking up a piece of paper from a flat table, which we often accomplish by laying finger pads on top of the paper and bending the paper into the hand. In either case, our ability to perceive and reason about contact with the world is crucial [1].

Common rigid robotic manipulators often cannot safely make large distributed contact, without risking damage to the fragile hardware or the environment. Given safety concerns, most prior work relies on using costly or specialized sensors to avoid applying unsafe contact forces [2], and explicitly avoiding direct contact during human-robot interaction [3]. Contact avoidance is especially common in work on assistive robotics, to ensure a user's safety from rigid robots [4]. However, such constraints can produce overly conservative assistance that may be too slow and uncomfortable for human users [5].

Rather than avoiding contact, we target contact-rich manipulation scenarios for robots to embrace contacts safely to provide better assistance. To this end, soft robot manipulators offer unique advantages compared to rigid end effectors. The inherent compliance of soft robot manipulators [2] enables robust control and mechanically aids in safe real-world operation [6]. This is especially relevant for delicate manipulation [7] and human-robot interaction [8]. The ability to deform with contact also makes them safer than rigid manipulators, applying significantly less force on contacting objects during collision. However, embedding contact and tactile sensors into such soft manipulators is an open challenge. Most previously proposed tactile sensors are either at least partially rigid [9] or limit strain [10], undermining soft robots' advantages. The lack of effective and deployable contact estimation solutions for soft robots is a bottleneck to developing adaptable and intelligent soft robotic manipulators [11].

Toward addressing contact sensing for soft robotic manipulators, we present MOE-Touch, a method for reconstructing a deformed soft robot shape and estimating its contact conditions for contact-rich soft robotic manipulation. MOE-Touch tracks the movement of keypoints on a soft robot manipulator and reconstructs watertight surface meshes of the deforming soft robot manipulator using a mesh energy-minimization method based on As-Rigid-As-Possible (ARAP) principles [12]. We show that this keypoint mesh optimization-based shape estimation method produces robust, high-fidelity shape reconstructions, providing more 3D shape structure compared to end-to-end learning-based approaches [13]. MOE-Touch then uses the observed deformations of the soft robot manipulator to predict points over the mesh that are in contact with other object surfaces. We demonstrate practical applications of MOE-Touch with two contact-rich tasks. First, to reconstruct occluded surfaces during assistive-care manipulation tasks, we update a modified formulation of a Gaussian Process Implicit Surface (GPIS) [14, 15] model with the predicted contact conditions. We also show MOE-Touch in novel grasping tasks with 2D deformable objects such as paper on a flat surface, where we use MOE-Touch to predict the relative orientation of the surface to enable successful grasps.

With MOE-Touch, we introduce the idea of reasoning about contact conditions and contacting object geometry from observed deformations of a soft robotic manipulator, which is a unique advantage of soft robots. Our key insight is that the deformation of soft robots can be an effective signal for contact conditions and configurations for soft robots. In summary, we make the following contributions:

- MOE-Touch, a novel method for soft robotic contact estimation that reconstructs multi-finger manipulator shapes and accurately estimates contact conditions, by observing deformations with a GNN-based contact estimation model trained on simulated data to reason about contact conditions over the deformed shapes,
- Implementation of MOE, a dexterous, multi-fingered, tendon-driven soft robotic manipulator capable of interacting safely with delicate surfaces that can be reconfigured to have different numbers of fingers,
- Demonstration to estimate and reconstruct occluded surfaces, such as a human head under a wig or an arm under a hospital gown with relevance to assistive robotics applications where safe and accurate surface interaction is critical,
- Demonstration of MOE-Touch and MOE manipulator on a novel robotic task of paper grasping from a flat surface with distributed contact.

2 Related Work

2.1 Soft Robotic Manipulators

Soft robotic manipulators are typically characterized by their deformable and compliant constituent material [11]. They are becoming increasingly popular because of their ability to interact safely with delicate objects and environments [7]. A spectrum of soft robotic manipulators exists from partially rigid or functionally rigid-linked soft robotic manipulators [16, 17] to fully soft robotic manipulators that bend continuously [18]. Recent works have started to demonstrate the "mechanical intelligence" of fully soft robotic manipulators, where their continuous deformation behavior contributes to the robustness and dexterity [19]. We primarily focus on such fully soft robotic manipulators in this work and explore their unique advantages in the domain of perception for contact-rich manipulation.

2.2 Soft Robotic Sensing

The compliance and deformation of soft robot manipulators [2] pose a challenge for perception and sensing [11] to determine the manipulator's proprioceptive state. Soft robot proprioceptive sensing and shape representations must capture complex deformation patterns of the soft robot [13, 20]. Conventionally, the shape of soft robots has been represented by parameterized 2-dimensional curves, which reduces the state estimation problem by modeling more tractable, low degrees-of-freedom systems. The most compact state representation uses a single degree-of-freedom curve with constant curvature, defined by its bending radius [21, 22]. More expressive representations construct multiple geometric primitives such as piecewise constant curvature models [23], multiple rigid frames [24], or rigid links [25]. These primitive representations have been used for dynamic control of soft robot manipulators [26]. However, these representations fail to capture volumetric information, and more deformation behaviors such as distributed, contact-based deformations [13].

Some methods have been proposed to capture rich soft robot states using point clouds [27, 13], but they rely on learning a state estimation model to reconstruct shapes by training on large training datasets. Previous works have proposed both explicit representations such as meshes [20, 13] and implicit representations such as neural Signed Distance Functions (SDFs) for soft bodies [28, 29]. Explicit representations are particularly convenient for this work because we can directly leverage the reconstructed body's nodes and their correspondences for downstream tasks such as contact surface reconstruction. Recent work grounds shape reconstruction with mechanics-based priors, which yields more data efficiency and stable proprioceptive state estimation [20]. In this paper, we show how these methods can be extended beyond a single-finger proprioceptive state estimation without interaction to object interaction and robotic manipulation.

2.3 Robotic Tactile Sensing

We take inspiration from tactile sensors that use deformations on the surface membrane to infer contact points [30, 31, 9, 32]. Specialized tactile sensors such as GelSight [33] and Digit [9] can be attached to rigid end effectors to infer contacts [34]. Researchers have demonstrated promising applications of these tactile sensors in reconstructing surfaces through touch [35]. Although such sensors can provide high-fidelity tactile and texture information about the contacting surface, they require contacts to occur on the small sensorized contact region, which constrains sensor configuration when used in robot manipulators [34]. Furthermore, tactile sensors tend to be difficult to embed into soft robots without introducing undesired rigidity [36]. Prior work demonstrated that soft robot manipulators deform significantly with contact [37, 20]. In this paper, we show how contact deformations of a soft robot manipulator can be utilized to reconstruct 3D contact surfaces under occlusion during interaction.

Recent methods have also been proposed that do not use tactile sensors, but instead estimate contact of unactuated deformable objects. Wi et al. [38, 39] use continuous implicit surface representations to reconstruct the shape and contact points. Van der Merwe et al. [29] propose an implicit representation that uses unoccluded view of the scene and 6D wrench data to estimate the deforming geometry and contacts of a cube sponge mounted on a robot, while pressed against an object from the YCB Object Set [40]. A limitation of the prior implicit shape estimation approaches, however, is that sampling query points and reconstructing surfaces tend to take too long to be used in real-time [29]. In this work, we highlight that using explicit shape representation and learning with geometric structure with a GNN can enable real-time high-fidelity shape and contact estimation.

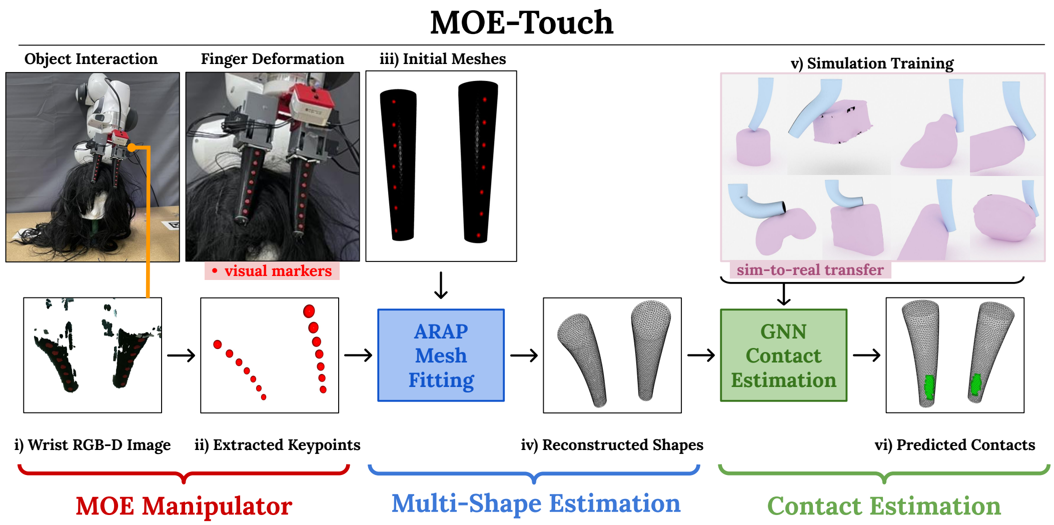

Figure 1: **Overview of MOE-Touch, which estimates contact conditions of soft robotic manipulators from deformation.** (i) We capture RGB-D images from the wrist camera during the interaction. (ii) We then extract keypoints on each finger of MOE. (iii) Using the initial mesh of MOE and extracted keypoints, we fit a mesh to the deforming state of MOE by ARAP principles. (iv) We reconstruct MOE's deforming shape surface geometry. (v) We train a GNN for contact estimation over the MOE surface mesh, using a large simulated dataset of MOE deformations and corresponding contact condition. (vi) From the reconstructed shapes, the GNN contact estimation model infers distributed, binary contact points for each MOE finger over the interaction.

3 Problem Statement

In this work, we aim to estimate the deformed shape of a continuum soft robotic manipulator and its contact regions based on the estimated deformation. To this end, we can make assumptions that are afforded to us because of the unique features of fully soft robotic manipulators. We assume that the material property is largely homogeneous and known. We also assume that the soft robot's material is soft enough to deform with contact, which we validated to be true in contact experiments.

The goal of the soft robot shape estimation in this work is to infer the overall mesh of the manipulator based on the sparse keypoint movements. We consider a soft robotic manipulator embodiment where the keypoints are tracked with visual markers attached to the soft fingers, although as with Yoo et al. [20], the keypoint movements could be indirectly tracked without external sensors or physical markers with a variety of sensors such as microphones. As such, the methods in this paper are relevant assuming the soft robot's sensors can lead to sufficiently reliable estimation of keypoints. In this work, we seek to use an optimization-based approach to infer deformation of a multi-finger soft robotic manipulator interacting with the environment. Based on these high-fidelity soft robot mesh shape reconstructions, we aim to use soft-body simulation to learn a model that infers contact points on the mesh.

4 Method

In this section, we describe the design of our soft robot manipulator MOE (Section 4.1). We then describe the components of MOE-Touch (Fig 1): proprioceptive sensing for MOE that reconstructs its deforming surface geometry (Section 4.2); contact estimation based on observed deformations in MOE (Section 4.3); and reconstruction of contacting surfaces using predicted contact conditions over an interaction trajectory (Section 6.1).

4.1 MOE Design

We design a soft tendon-driven manipulator which we call Multi-finger Omnidirectional End-effector (MOE), building on a single-finger tendon-driven soft robot [20]. The design is largely modular, where each of the fingers is an independent subsystem that can be detached and assembled to get multiple-finger configurations. In this work, we present results for a MOE with two fingers, as shown in Figure 2, and three fingers for the paper grasping task. Each of MOE's soft fingers is molded from silicone with low hardness. Each finger has four embedded tendons, which are actuated by two servo motors. Each pair of tendons actuated by a single servo motor controls MOE finger's range of motion in a bending plane. We include an RGB-D camera on the wrist of MOE to provide egocentric-view depth, as shown in Figure 1. Red markers are placed on the surfaces of the MOE fingers for the RGB-D camera to track MOE keypoints as the body deforms.

4.2 Multi-Shape Estimation

To guide the shape estimation of MOE, we track the 7 red keypoint markers placed on the surface of each MOE finger, as shown in Figure 1. We segment the markers using color thresholds and apply DBSCAN [41] to cluster the 3D points, localizing marker centers based on the point densities. In the initial frame, we find the nodes on the initial mesh closest to the keypoints and use them as handle points. From the initialization phase, we account for the movement of each of the keypoints frame-to-frame.

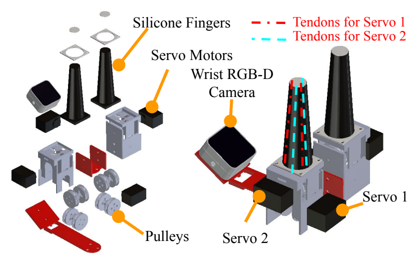

Figure 2: **Design of MOE.**

We consider the surface mesh $S_n = (E_n, V_n)$, representing the n^{th} individual finger of MOE and the deformed MOE finger mesh S'_n, where a surface mesh is defined by edges $e_{i,j} \in E$ composed from vertices $i, j \in V$. As previously proposed [42, 20], we include a penalty on the rotations of the neighboring edges $e_l \in N(e_k)$ to produce mesh updates that are physically admissible. The energy to minimize is

$$\begin{aligned} E_{\text{smoothed}}(\{S_n, S'_n\}) = \sum_{n=1}^{N} \min_{R_{n,1},\dots,R_{n,m}} \sum_{k=1}^{m} \Bigg(& \sum_{i,j \in e_k} c_{ijk} \| e^n_{ij} - R_{n,k} e^{n'}_{ij} \|^2 \\ & + \lambda \hat{A} \sum_{e_l \in N(e_k)} w_{kl} \| R_{n,k} - R_{n,l} \|^2 \Bigg), \end{aligned} \tag{1}$$

where c_{ijk} are the cotan weights [43], λ is the regularization weight, $R_1, ..., R_m \in SO(3)$ are the local rotations for each of the edges $e_k \in E$ where $m = |E|$, $\hat{A}$ is the triangle area and w_{kl} are the scalar weight terms defined by the cotan weights of the dual mesh of e_{kl} [43]. We iteratively minimize $E_{\text{smoothed}}(\{S_n, S'_n\})$ with local-global optimizer as outlined in Levi and Gotsman [42].

To reconstruct the full mesh shape of MOE, we treat vertices corresponding to the keypoints $p_{1,\dots,k}$ as being constrained to the new positions, based on the predicted keypoint positions. The rest of the mesh vertex positions are moved to minimize E_{smoothed}. Note that we jointly optimize the surface meshes of the fingers together for multi-shape estimation of the deformed state of the MOE manipulator.

We visually track keypoints observed by a wrist-mounted RGB-D camera, which simplifies tracking for multiple fingers without the need to embed sensors in each finger as in [20] where one finger needs 6 microphones. Our formulation for multi-shape estimation can be applied as long as keypoint positions can be observed (e.g., through modalities beyond vision) and the correspondence to the mesh vertices is known. Unlike previous contexts in which ARAP has been applied for soft bodies, we study deformations caused by interactions with the environment. To account for occlusion, which can occur during interaction, we remove the keypoints from consideration that are not observed. Based on our formulation, the vertex associated with the keypoint will be updated based on the observable keypoints. By approaching multi-shape estimation with this energy-optimization approach, we ground the predicted shapes on the undeformed finger mesh S_n to mitigate drift from accumulating errors or outlier shape estimation errors.

4.3 Contact Estimation

For MOE-Touch, we aim to use shape estimation of MOE's deforming state to infer its contact conditions. Graph neural network (GNN) architectures have been shown to learn and reason about complex physical interactions and spatial relationships [44, 45]. We present and train a GNN-based contact estimation model on the simulated dataset, where the inputs are MOE point clouds labeled with contact obtained from the simulation environment. We deploy the trained contact estimation model directly on real-world predictions of MOE mesh shapes to predict the contacting nodes as MOE deforms during an interaction trajectory, as visualized in Figure 1. By using observed deformation as a signal to predict contact conditions, we assume that the observed deformation is sufficiently expressive in disambiguating contact conditions. This may depend on the representation of the shape, the deformation behavior of the material, and the contact configuration representation.

4.3.1 Simulation Environment

Contact points are difficult to obtain directly from the real world due to occlusions from the contacting object. Previous works have demonstrated the capabilities of soft-body simulation to generate training datasets of deformed shapes and contact information [46, 13].

We model our soft-robot mamnipulator using the soft body simulator in SOFA [47] with its tools for solving Finite Element Method (FEM) problems. We follow previously recorded material properties for the silicone body of MOE, with Poisson's Ratio of 0.1 and Young's modulus of 100 kPa. For the integrator, we use the Rayleigh stiffness value of 0.1 and Rayleigh Mass of 0.1. We implement cable tensions for the tendons with displacement action input. The resulting simulator scenes are visualized in the appendix.

To sample from varying contact normals and surface orientations, we import objects from the YCB Object and Model Set [40] into a SOFA simulation environment. We also generate and import tendon-actuated meshes of MOE. We randomize the selected contacting object's orientation and position with respect to MOE's trajectory, to simulate various contact locations and orientations. We also apply different actuation forces to MOE's fingers, and the actuated manipulator towards the contacting object to observe further deformations. From these simulated trials, we generate a dataset of 174,590 meshes and corresponding contact points, which were recorded as the indices of the MOE mesh vertices in contact with an object.

4.3.2 Contact Estimation Model

For $i \in V'$, where V' is the set of vertices from the multi-body shape estimation in Section 4.2, we seek to predict its binary contact label. Adaptive graph construction allows the model to better capture the underlying structure of the point cloud as the features evolve through the network layers for estimating contact on MOE, especially compared to baseline approaches that do not encode geometric relationships as shown in the evaluation. Each layer applies an edge convolution operation introduced by Dynamic Graph CNN (DGCNN) [44], which updates the feature representation h_i of the point, by aggregating information from its neighboring points in the graphs constructed by k-nearest neighbors in the feature space. For a point $i \in V'$ and its neighbor $j \in \mathcal{N}(i)$ in the learned feature space h_i, the edge convolution operation is defined as:

$$h_i^{l+1} = \sum_{j \in \mathcal{N}(i)} \mathrm{ReLU}(\mathbf{\Theta} \cdot (h_i^l - h_j^l) + \mathbf{\Phi} \cdot h_i^l),$$

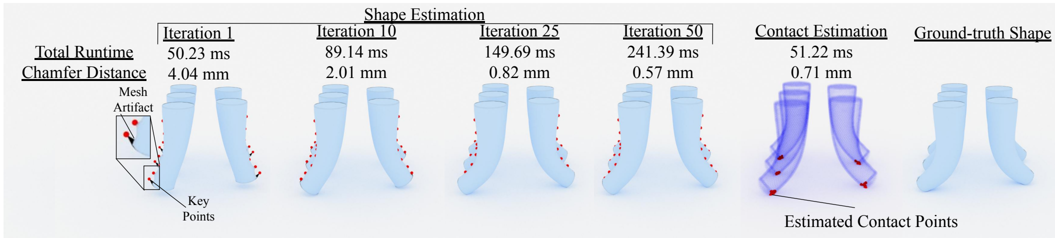

Figure 3: **Scaling MOE-Touch to More Complex Manipulators.** We demonstrate that MOE-Touch scales up from two fingers to a five-finger variant of MOE.

where h_i^l is the feature of point i at layer l, and $\mathbf{\Theta}$ and $\mathbf{\Phi}$ are learnable parameters. The network consists of three edge convolution layers, each followed by a max-pooling operation that aggregates features globally across all points, effectively capturing both local and global context to allow the model to reason about the deformation at multiple scales. The features extracted from all layers are concatenated to form a global feature vector, which is then processed by fully connected layers to predict MOE contact conditions. The model was trained on MOE point clouds sampled to 2048 points. To account for the imbalance in the dataset, where there are noticeably more points not in contact than the ones in contact, we use a weighted softmax cross entropy loss function. The training details are provided in the appendix.

5 Evaluation

5.1 Baselines

We evaluate the proposed MOE-Touch multi-shape and contact estimation against state-of-the-art baseline approaches with two metrics: Chamfer distance (CD) and runtime per observation. We first compare against a k-nearest neighbors (KNN) baseline, where we select the training example with the closest keypoint positions compared to the observed keypoint positions. We use the keypoints as input for KNN to reduce search complexity and make the method computationally tractable. We show results for both KNN[Sub.] where a random sampling of 10% of the training data is used to reduce runtime per observation and for KNN[All], where the entire training data is provided for prediction.

We also evaluate against Neural Deformation and Contact Field (NDCF) [29], which was presented as a method to jointly predict deformation and contact conditions. In the original work [29], NDCF uses unobstructed side-view point cloud observations of the soft end-effector and wrist wrench measurements as inputs to the model. As we do not have access to the wrench measurements in this work, we only provide an unobstructed side-view point cloud of MOE to the NDCF pipeline and pre-train on MOE's undeformed finger shape after normalizing and centering the mesh. Also compared to the original work, the testing scenarios for MOE-Touch are different. Notably, NDCF was only tested on a symmetric sponge (46 mm x 46 mm x 46 mm) that only undergoes surface-level local deformation and indentations. MOE undergoes global shape deformation through bending. To account for the domain differences, we also included results for the sponge interactions with YCB objects, which should provide the most optimistic results for NDCF.

Implicit surface representations generally suffer from longer runtimes due to the need to query and sample points densely. As an additional baseline, we provide results for MOE-NDCF, which uses MOE-Touch's mesh shape estimation module and queries contact points using NDCF's contact estimation model. For all of the methods, we evaluated on 10% of the dataset sampled from unseen contact trajectories.

	Model	Input	Soft Manipulator	Performance Metrics: BCD (*mm* ↓)	Performance Metrics: Runtime (*ms* ↓)
Shape	KNN [Sub.]	KP	MOE	1.719 ± 1.924	**37.61 ± 2.154**
	KNN [Full]	KP	MOE	0.978 ± 0.276	250.4 ± 8.741
	NDCF [29]	PC	Sponge	0.974 ± 0.305	2546 ± 473.4
	NDCF [29]	PC	MOE	3.455 ± 4.069	2139 ± 172.1
	MOE-NDCF	KP + PC	MOE	-	-
	MOE-Touch	KP	MOE	**0.617 ± 0.047**	**47.89 ± 2.980**
Contact	KNN [Sub.]	KP	MOE	8.318 ± 6.173	**$37.61^* \pm 2.154$**
	KNN [All]	KP	MOE	3.079 ± 3.042	$250.4^* \pm 8.741$
	NDCF [29]	PC	Sponge	4.891 ± 3.174	$2546\ ^* \pm 473.4$
	NDCF [29]	PC	MOE	19.31 ± 8.347	$2139\ ^* \pm 172.1$
	MOE-NDCF	KP + PC	MOE	9.189 ± 5.394	$112.3^* \pm 19.31^*$
	MOE-Touch	KP	MOE	**2.740 ± 2.827**	**$86.97\ ^* \pm 4.111$**

Table 1: Evaluation of Shape and Contact Estimation with Ground Truth from Simulation Environments with YCB objects. Runtime is evaluated on the same environment with the same computing hardware. Some methods use unobstructed partial point clouds (PC) from the side view as input [29] while others use mesh keypoint (KP) positions. * Runtime for contact estimation includes processing time for the shape estimation, which all of the methods require before or during contact estimation. ↓ indicates that lower is better.

5.2 Simulation Study

Simulation environments readily provide unoccluded ground-truth contact conditions, allowing us to use bidirectional Chamfer distance (BCD) as the metric for both shape and contact estimations of the methods. We can also obtain segmented point clouds of MOE, which the NDCF-based baselines require. As noted, contact estimation generally relies on accurate shape estimation since the contact points are registered onto the deforming body surface. As shown in Table 1, MOE-Touch produces lower shape estimation error with an average BCD of 0.617 mm across the test dataset. Additionally, the runtime of MOE-Touch's shape estimation module is faster than any of the baselines except KNN[Sub.].

For contact estimation, the proposed MOE-Touch contact estimation module outperformed all of the baseline methods on BCD with 2.740 mm. The total runtime of MOE-Touch was faster than the NDCF and KNN[All] baselines. KNN[Sub.] had the fastest total runtime but with degraded performance compared to the KNN[All].

We also demonstrate in the simulated environment that the proposed MOE-Touch pipeline performs well even with increased system complexity by testing a five-finger variant of the proposed MOE end-effector as shown in Figure 4. We note that the shape estimation module converges by iteration 50 with BCD of 0.57 mm for all of the five MOE fingers. We then note that the contact estimation step also scales well with an inference time of 51.22 ms.

5.3 Real-world Evaluation

We demonstrate that MOE-Touch can estimate contact conditions accurately in varying contacting conditions with controlled contact on a thin plate (see Figure 5). Table 2 (top rows) reports the quantitative contact estimation results with comparisons to the baseline. By using a known simple geometry such as a thin plate, we can evaluate the contact estimation performance on specific surface regions of MOE. We evaluate contact estimation performances for contact at the tip, in the middle, and close to the base of the robot for contact from the front and contact from the side. For each combination of contact conditions, we run 3 trials, resulting in 6 trials for each contact region.

We note that the contact estimation is accurate with <10 mm unidirectional Chamfer Distance (CD) with notably higher error at the base. The performance is likely worse near the base of the MOE

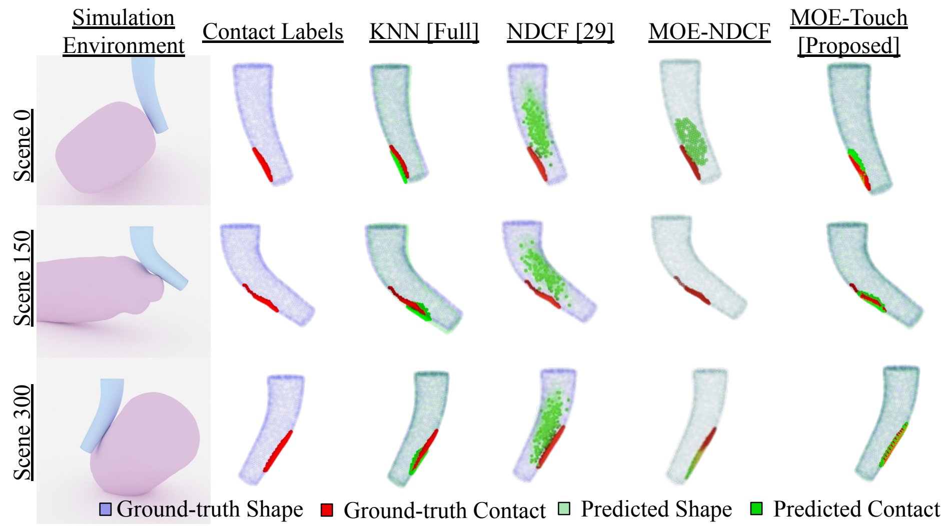

Figure 4: **Visualization of Sampled Shape and Contact Estimation Results compared to Baseline Approaches**. We show the results across three different trajectories with different YCB objects.

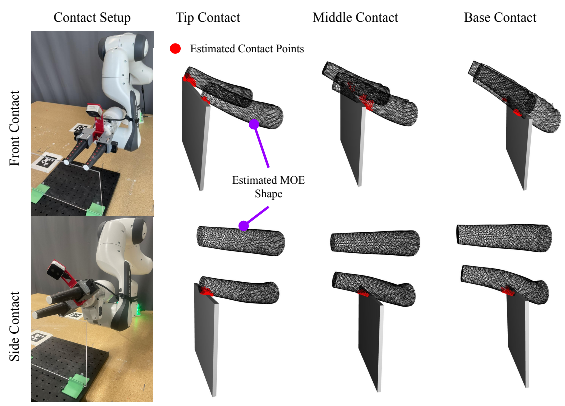

Figure 5: Controlled thin plate contact estimation experiment to demonstrate that MOE-Touch is sensitive in a large portion of the MOE soft robot.

Method	**Contact Object**	**UCD** (*mm* ↓)
KNN [Full]	Plate (Tip)	3.41± 0.318
MOE-Touch	Plate (Tip)	**3.03** ± 0.475
KNN [Full]	Plate (Middle)	13.5 ± 1.87
MOE-Touch	Plate (Middle)	**7.08**± 0.512
KNN [Full]	Plate (Base)	20.1± 1.89
MOE-Touch	Plate (Base)	**9.92** ± 1.28
KNN [Full]	Head (Bald)	7.76± 1.06
MOE-Touch	Head (Bald)	**6.58**± 0.827
KNN [Full]	Head (Wig)	13.7 ± 2.11
MOE-Touch	Head (Wig)	**12.2** ± 1.37
KNN [Full]	Arm (Gown)	6.88 ± 0.581
MOE-Touch	Arm (Gown)	**6.24** ± 0.419

Table 2: Quantitative comparisons of MOE-Touch to a KNN-based sparse contact estimation baseline [48], both for a controlled experiment setting and task-relevant settings.

because the robot is less compliant and deforms less, making it more difficult for the model to disambiguate possible contact conditions. In all contact conditions, MOE-Touch performs better than the baseline, most notably at the base with 50.65 % reduction in CD error.

We also test the contact estimation module on accurate models of the head and arm. Both environments are motivated by common contact-rich assistive robotic settings, where visual occlusion may be common and unavoidable, requiring the robot to safely interact with the human subject. For the head setup, we randomly selected a head mesh of an adult person from a craniofacial shape dataset [49], 3D-print the meshed model, and test with and without a voluminous wig.

We then test 30 distinct MOE contact conditions on the head to evaluate shape and contact estimation modules with and without a wig. We register the point clouds together from the wrist-mounted RGB-D camera to show the contact coverage across the head in Figure 6. We then evaluate the shape and contact estimation modules by registering the predicted contact points together, computing unidirectional average CD from the contact points to the head ground-truth mesh nodes. We then perform a similar series of 15 contact trials on a model of an adult human arm occluded by a hospi-

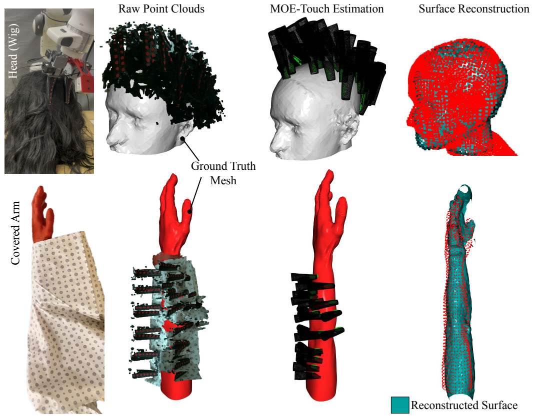

Figure 6: Surface reconstruction with MOE-Touch

Method	Contact Object	Uni. CD ($mm \downarrow$)
Non-Probabilistic	Head (Bald)	16.01
GP w/ Sphere	Head (Bald)	13.83
GP w/ Prior	Head (Bald)	**3.64**
Non-Probabilistic	Head (Wig)	16.05
GP w/ Sphere	Head (Wig)	14.41
GP w/ Prior	Head (Wig)	**3.62**
GP w/ Prior	Arm	**9.90 (4.82*)**

Table 3: Contacting surface reconstruction results compared to the ground truth. * denotes the result for the arm with the unsampled hand removed from evaluation.

tal gown (see Figure 6). Similar to the trials with the model head, we register the predicted contact points, compared to the ground-truth mesh, and compute the CD metrics.

We observe that in all three settings, the MOE-Touch pipeline performs functionally well and improves on the baseline method in all three cases with the lowest errors. The environments with the bald head and arm both result in an average MOE-Touch CD error of around 6.5 mm. We can notice a noticeably higher CD error of 12.22 mm in the environment with a head and a wig. A significant portion of the error may come from the thickness that the wig's inner hair net which is around 5mm thick. Because we do not have a separate ground-truth mesh for the head with a wig, we still evaluate the metrics with the bald head mesh.

On a consumer workstation with an RTX 4090 GPU, the MOE-Touch shape estimation module outputs a mesh with 2048 vertices from 50 iterations with a runtime of 49.55 ms, and the contact estimation inference time runs on average 43.62 ms for each deformed shape. For comparison, a neural implicit surface-based approach takes 2079 ms per scene to reconstruct the mesh and contact patch [29]. The efficiency of MOE-Touch is largely a result of the methods that we develop around our domain-specific assumptions for soft robotic perception, such as homogeneous material composition.

6 Applications

We demonstrate practical applications of our MOE-Touch approach for two real-world manipulation tasks that involve large distributed contact: contacting surface reconstruction (Section 6.1) and paper grasping (Section 6.2). In both tasks, the robotic manipulator must interact with the environment and make distributed contact. With contacting surface reconstruction, we demonstrate that MOE-Touch can be used to pat an occluded surface and reconstruct it. Then, with paper grasping, we demonstrate the advantages of MOE's softness to guide its perception with MOE-Touch and to robustly manipulate objects that are difficult to grasp.

6.1 Contacting Surface Reconstruction

MOE-Touch's shape estimation and contact estimation modules provide contact information. One useful application of a soft robotic manipulator is safely interacting with an occluded surface, such as the scalp under hair or arm under a hospital gown, and using the contact estimates to reconstruct them. In such tasks, we have useful priors on the occluded body part's geometry. We use a task-dependent prior mesh specific to the domain. For the initial task of reconstructing a human head, we use an open-sourced canonical head 3D mesh [49] and trained a Gaussian Process (GP) to learn a prior over the SDF of the mesh. Given a set of dense grid points $\mathbf{X}$ and corresponding SDF values

$\mathbf{Y}$, the GP model is

$$f(\mathbf{x}) \sim \mathcal{GP}\left(c, \sigma^2 \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2l^2}\right)\right), \tag{2}$$

where σ^2 is the variance and l is the length scale of the Radial Basis Function (RBF) kernel, with the observation model

$$y = f(\mathbf{x}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2) \tag{3}$$

where n is the number of training points. The training objective maximizes the marginal log-likelihood loss

$$\log p(\mathbf{Y}|\mathbf{X}) = -\frac{1}{2}\mathbf{Y}^\top(\mathbf{K}+\sigma_n^2\mathbf{I})^{-1}\mathbf{Y} - \frac{1}{2}\log|\mathbf{K}+\sigma_n^2\mathbf{I}| - \frac{n}{2}\log 2\pi, \tag{4}$$

where $\mathbf{K}$ denotes the covariance matrix constructed using an RBF kernel over the training inputs in $\mathbf{X}$. Once we have a trained prior, we fine-tune the GP with the contact-point information to obtain the posterior SDF. Finally, we reconstructed the head mesh using Poisson Surface Reconstruction (PSR) [50] on a point cloud obtained by running the Marching Cubes Algorithm (MCA) [51] over the zero-level set of the SDF.

For training, the GP takes grid points as input and generates a multivariate normal distribution ($\mathcal{N}$) for the output. An SDF is sampled from $\mathcal{N}$ for each point in the dense grid and compared with the ground truth using an exact marginal log likelihood loss. The gradients of the loss value with respect to the kernel parameters are computed and updated with gradient descent. The pretrained GP takes grid points as input and outputs $\mathcal{N}$. The output SDF is formed using only the mean of $\mathcal{N}$.

Results in Table 3 show that our method reconstructs the mesh from real-world contact points accurately with an average CD of 3.64 mm for the bald head, 3.62 mm for the head with a wig, and 9.90 mm for an arm dressed in a hospital gown. The capability of the task-dependent prior method to generate a watertight mesh after accommodating real-world data is shown in Figure 6 for the arm and the head. For the evaluation of the arm, we reported errors for the entire arm and for the arm with the hand removed since we did not sample from it during experiments. The prior mesh used to pretrain the head GP has a CD of 5.46 mm with the 3D printed head mesh, and the prior mesh for the arm GP has a CD of 5.677 mm with the 3D printed arm mesh.

We also present a baseline method based on some previous works that assume a primitive geometric shape as initialization for interactive perception and mesh reconstruction [35, 15]. We use a spherical prior as a naive method to obtain the posterior distribution over the real-world contact points. The main point of failure in this method can be attributed to the fitted sphere mesh, with points that are significantly out of distribution from an average human head.

We also compare the GP-based implicit surface methods to using a non-probabilistic approach, where we use the subset of vertices on the prior mesh and deform them towards the nearest neighbor contact points. We then apply Laplacian smoothing to interpolate a smooth mesh between the contact points. The non-probabilistic method results in a qualitatively worse formed surface compared to the GP method with a spherical prior. This method performs the worst for surface reconstruction with average CD values of 16.01 mm for bald head data, and 16.05 mm for head with a wig. The proposed task-dependent prior-based surface reconstruction module performs better than the two baseline methods, resulting in 73.68 % and 77.26 % reduced average CD metric error for the head without a wig compared to using a spherical prior and the non-probabilistic method, respectively.

6.2 Paper Grasping

Grasping a 2D deformable object such as a piece of fabric or paper from a flat surface is a difficult because it requires the robot to make large distributed contact with the object and fold the object into a grasp while maintaining contact. Previous works aim to address this challenging task by searching for a sufficiently wrinkled area on the object to pinch [52] or using

Incline Angle	Success w/o MOE-Touch	Success w/ MOE-Touch
0 deg	**5/5**	**5/5**
30 deg	1/5	**5/5**
45 deg	0/5	**4/5**

Table 4: **Paper Grasping Results.**

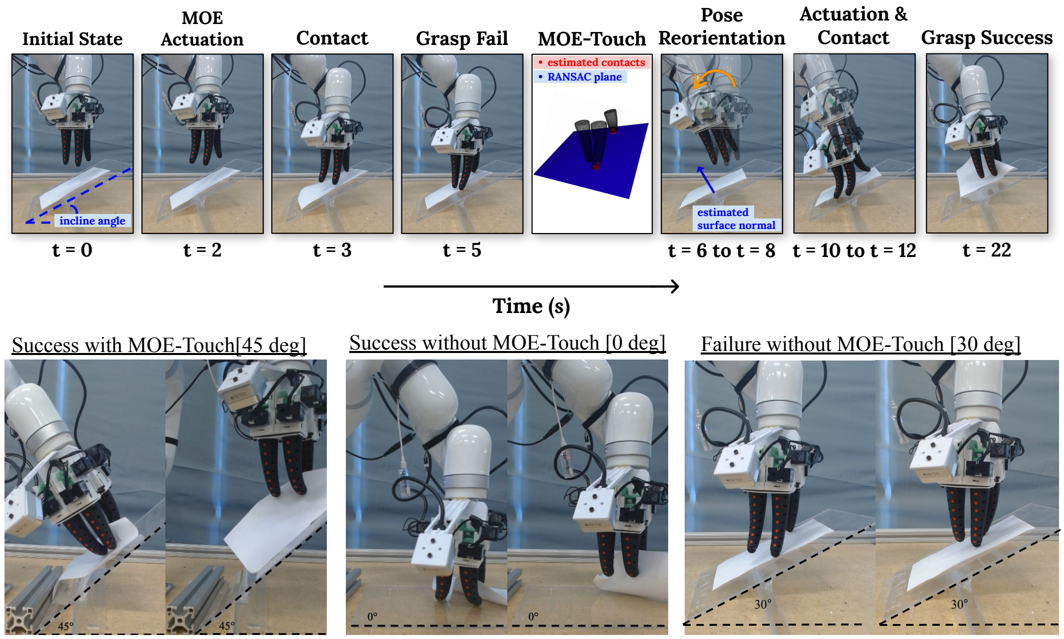

Figure 7: **Application of MOE-Touch in flat deformable object grasping task on inclined surfaces.**

a specialized mechanism such as suction gripper to pick up the flat object [53]. For a multi-finger manipulator to perform this task, each finger must be contacting the flat object and be aligned to fold the object into a secure grasp. To this end, we evaluate MOE-Touch on the task of grasping paper on a surface with an initially unknown incline angle. Additionally, to evaluate the modularity of MOE-Touch, we tested a variant of MOE with three fingers. We prepared an inclined clear acrylic flat surface with a 190×130 mm common printer paper on top. MOE made contact with the paper initially misaligned from the acrylic surface. We used MOE-Touch to estimate contact points with the surface and fitted a plane to the points with Random Sample Consensus (RANSAC). We then reoriented MOE to be normal to the surface and grasped the paper (see Figure 7). We tested with 0, 30, and 45-degree incline of the surface. We compared the success rates of the paper grasping task out of 5 trials for each setting against not using MOE-Touch. Surprisingly, MOE could still grasp the paper at 30-degree incline once without MOE-Touch, showing robustness of its compliance and mechanical intelligence. However, with 45-degree inclines, MOE needed MOE-Touch to succeed in the task.

7 Conclusion

In this work, we introduce methods for contact estimation in contact-rich soft robotic manipulation. We develop MOE, a modular Multi-finger Omnidirectional End-effector that can safely and robustly interact with the world for contact-rich manipulation. We use a mesh energy optimization-based method to estimate the shape of MOE in interaction with the environment. The proposed MOE-Touch method takes an explicit mesh optimization-based approach to reconstruct the deformed shape of the soft robot and reason about contact conditions with a GNN over the mesh. We show that MOE-Touch can estimate occluded surface contact with an average distance error of 6.25 mm, improving on the baseline by 17.53%. We show that the MOE-Touch can be deployed to reconstruct an occluded surface with averaged errors of 3.62 mm. We then show the use case of MOE-Touch for a manipulation task of grasping paper on arbitrarily inclined surfaces, where contact estimation guides re-orientation of MOE to be normal to the contacting surface.

8 Limitations

One limitation of this work is that we train the contact estimation module with binary contact labels. Extending MOE-Touch to estimate contact pressure may present advantages in downstream manipulation tasks that require more complex interactions. Just as human skins have four different mechanoreceptors responsible for different tactile stimuli [54], robotic tactile modalities offer different advantages and multi-tactile modality sensor fusion may be a promising direction to augment MOE-Touch. Currently, our approach relies on visually tracking finger keypoint markers on the backs of the fingers using a wrist-mounted camera, which in some cases may be occluded in real-world deployment. A potential solution is to track a large number of keypoints so that failure is less likely. Additionally, embedding sensors within the fingers [9] or incorporating acoustic sensing [20] pose promising directions to overcome occlusion with other modalities to estimate mesh keypoint positions.

Although the MOE-Touch grounds the multi-shape estimation on the undeformed meshes of the fingers to prevent drifting and accumulating errors, there is no mechanism implemented to ensure frame-to-frame prediction consistency in MOE-Touch, which may be important for long-horizon real-world deployment with noise. This limitation may be addressed using approaches such as Kalman filters [55] or by incorporating the history of previous observations. In this work, we also assume that the contacting object causes observable deformation in the soft robot and therefore must be more rigid than the material used to construct MOE's fingers.

References

[1] Y. C. Nakamura, D. M. Troniak, A. Rodriguez, M. T. Mason, and N. S. Pollard. The complexities of grasping in the wild. In *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, pages 233–240. IEEE, 2017.

[2] J. Hughes, U. Culha, F. Giardina, F. Guenther, A. Rosendo, and F. Iida. Soft manipulators and grippers: A review. *Frontiers in Robotics and AI*, 3:69, 2016.

[3] Y. Wang, Z. Sun, Z. Erickson, and D. Held. One policy to dress them all: Learning to dress people with diverse poses and garments. In *Robotics: Science and Systems XIX*, 2023.

[4] S. Li, N. Figueroa, A. Shah, and J. Shah. Provably safe and efficient motion planning with uncertain human dynamics. In *Robotics: Science and Systems XVII*, 2021.

[5] G. Canal, C. Torras, and G. Alenyà. Are preferences useful for better assistance? a physically assistive robotics user study. *ACM Transactions on Human-Robot Interaction (THRI)*, 10(4): 1–19, 2021.

[6] N. Kuppuswamy, A. Alspach, A. Uttamchandani, S. Creasey, T. Ikeda, and R. 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.

[7] N. R. Sinatra, C. B. Teeple, D. M. Vogt, K. K. Parker, D. F. Gruber, and R. J. Wood. Ultragentle manipulation of delicate structures using a soft robotic gripper. *Science Robotics*, 4(33): eaax5425, 2019.

[8] P. Polygerinos, N. Correll, S. A. Morin, B. Mosadegh, C. D. Onal, K. Petersen, M. Cianchetti, M. T. Tolley, and R. F. Shepherd. Soft robotics: Review of fluid-driven intrinsically soft devices; manufacturing, sensing, control, and applications in human-robot interaction. *Advanced Engineering Materials*, 19(12):1700016, 2017.

[9] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020.

[10] R. Bhirangi, T. Hellebrekers, C. Majidi, and A. Gupta. Reskin: versatile, replaceable, lasting tactile skins. In *5th Annual Conference on Robot Learning*, 2021.

[11] H. Wang, M. Totaro, and L. Beccai. Toward perceptive soft robots: Progress and challenges. *Advanced Science*, 5(9):1800541, 2018.

[12] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Symposium on Geometry processing*, volume 4, pages 109–116. Citeseer, 2007.

[13] U. Yoo, H. Zhao, A. Altamirano, W. Yuan, and C. Feng. Toward Zero-Shot Sim-to-Real Transfer Learning for Pneumatic Soft Robot 3D Proprioceptive Sensing. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 544–551. IEEE, 2023.

[14] O. Williams and A. Fitzgibbon. Gaussian process implicit surfaces. In *Gaussian Processes in Practice*, 2006.

[15] S. Dragiev, M. Toussaint, and M. Gienger. Gaussian process implicit surfaces for shape estimation and grasping. In *2011 IEEE International Conference on Robotics and Automation*, pages 2845–2850. IEEE, 2011.

[16] W. Zhu, C. Lu, Q. Zheng, Z. Fang, H. Che, K. Tang, M. Zhu, S. Liu, and Z. Wang. A soft-rigid hybrid gripper with lateral compliance and dexterous in-hand manipulation. *IEEE/ASME Transactions on Mechatronics*, 28(1):104–115, 2022.

[17] P. Mannam, K. Shaw, D. Bauer, J. Oh, D. Pathak, and N. Pollard. Designing anthropomorphic soft hands through interaction. In *2023 IEEE-RAS 22nd International Conference on Humanoid Robots (Humanoids)*, pages 1–8. IEEE, 2023.

[18] S. Puhlmann, J. Harris, and O. Brock. RBO hand 3: A platform for soft dexterous manipulation. *IEEE Transactions on Robotics*, 38(6):3434–3449, 2022.

[19] A. Bhatt, A. Sieler, S. Puhlmann, and O. Brock. Surprisingly robust in-hand manipulation: An empirical study. In *Robotics: Science and Systems XVIII*, 2022.

[20] U. Yoo, Z. Lopez, J. Ichnowski, and J. Oh. POE: Acoustic soft robotic proprioception for omnidirectional end-effectors. *arXiv preprint arXiv:2401.09382*, 2024.

[21] C. Della Santina, R. K. Katzschmann, A. Biechi, and D. Rus. Dynamic control of soft robots interacting with the environment. In *2018 IEEE International Conference on Soft Robotics (RoboSoft)*, pages 46–53. IEEE, 2018.

[22] U. Yoo, Y. Liu, A. D. Deshpande, and F. Alamabeigi. Analytical design of a pneumatic elastomer robot with deterministically adjusted stiffness. *IEEE Robotics and Automation Letters*, 6(4):7773–7780, 2021.

[23] C. Della Santina, A. Bicchi, and D. Rus. On an improved state parametrization for soft robots with piecewise constant curvature and its use in model based control. *IEEE Robotics and Automation Letters*, 5(2):1001–1008, 2020.

[24] Y. Liu, U. Yoo, S. Ha, S. F. Atashzar, and F. Alambeigi. Influence of antagonistic tensions on distributed friction forces of multisegment tendon-driven continuum manipulators with irregular geometry. *IEEE/ASME Transactions on Mechatronics*, 27(5):2418–2428, 2021.

[25] Y. She, S. Q. Liu, P. Yu, and E. Adelson. Exoskeleton-covered soft finger with vision-based proprioception and tactile sensing. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 10075–10081. IEEE, 2020.

[26] D. A. Haggerty, M. J. Banks, E. Kamenar, A. B. Cao, P. C. Curtis, I. Mezić, and E. W. Hawkes. Control of soft robots with inertial dynamics. *Science Robotics*, 8(81):eadd6864, 2023.

[27] R. Wang, S. Wang, S. Du, E. Xiao, W. Yuan, and C. Feng. Real-time soft body 3d proprioception via deep vision-based sensing. *IEEE Robotics and Automation Letters*, 5(2):3382–3389, 2020.

[28] C. Higuera, S. Dong, B. Boots, and M. Mukadam. neural contact fields: Tracking extrinsic contact with tactile sensing. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12576–12582. IEEE, 2023.

[29] M. J. Van der Merwe, Y. Wi, D. Berenson, and N. Fazeli. Integrated object deformation and contact patch estimation from visuo-tactile feedback. In *Robotics: Science and Systems XIX*, 2023.

[30] A. Yamaguchi and C. G. Atkeson. Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pages 1045–1051. IEEE, 2016.

[31] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora. The tactip family: Soft optical tactile sensors with 3d-printed biomimetic morphologies. *Soft Robotics*, 5(2):216–227, 2018.

[32] J. A. Collins, C. Houff, P. Grady, and C. C. Kemp. Visual contact pressure estimation for grippers in the wild. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10947–10954. IEEE, 2023.

[33] W. Yuan, S. Dong, and E. H. Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

[34] S. Suresh, H. Qi, T. Wu, T. Fan, L. Pineda, M. Lambeta, J. Malik, M. Kalakrishnan, R. Calandra, M. Kaess, et al. Neural feels with neural fields: Visuo-tactile perception for in-hand manipulation. *arXiv preprint arXiv:2312.13469*, 2023.

[35] S. Suresh, Z. Si, J. G. Mangelson, W. Yuan, and M. Kaess. ShapeMap 3-D: Efficient shape mapping through dense touch and vision. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 7073–7080. IEEE, 2022.

[36] H. Kim, O. C. Kara, and F. Alambeigi. A Soft and Inflatable Vision-Based Tactile Sensor for Inspection of Constrained and Confined Spaces. *IEEE Sensors Journal*, 2023.

[37] A. Sieler and O. Brock. Dexterous soft hands linearize feedback-control for in-hand manipulation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8757–8764. IEEE, 2023.

[38] Y. Wi, P. Florence, A. Zeng, and N. Fazeli. Virdo: Visio-tactile implicit representations of deformable objects. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 3583–3590. IEEE, 2022.

[39] Y. Wi, A. Zeng, P. Florence, and N. Fazeli. VIRDO++: Real-world, visuo-tactile dynamics and perception of deformable objects. In *6th Annual Conference on Robot Learning*, 2022.

[40] B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, and A. M. Dollar. Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set. *IEEE Robotics & Automation Magazine*, 22(3):36–52, 2015.

[41] E. Schubert, J. Sander, M. Ester, H. P. Kriegel, and X. Xu. DBSCAN revisited, revisited: why and how you should (still) use DBSCAN. *ACM Transactions on Database Systems (TODS)*, 42(3):1–21, 2017.

[42] Z. Levi and C. Gotsman. Smooth rotation enhanced as-rigid-as-possible mesh animation. *IEEE transactions on visualization and computer graphics*, 21(2):264–277, 2014.

[43] K. Crane, F. de Goes, M. Desbrun, and P. Schröder. Digital geometry processing with discrete exterior calculus. In *ACM SIGGRAPH 2013 courses*, SIGGRAPH '13, 2013.

[44] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon. Dynamic graph cnn for learning on point clouds. *ACM Transactions on Graphics*, 38(5):1–12, 2019.

[45] T. H. E. Tse, Z. Zhang, K. I. Kim, A. Leonardis, F. Zheng, and H. J. Chang. S^2Contact: Graph-based network for 3d hand-object contact estimation with semi-supervised learning. In *European Conference on Computer Vision*, pages 568–584. Springer, 2022.

[46] A. Sipos and N. 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.

[47] J. Allard, S. Cotin, F. Faure, P.-J. Bensoussan, F. Poyer, C. Duriez, H. Delingette, and L. Grisoni. Sofa-an open source framework for medical simulation. In *MMVR 15-Medicine Meets Virtual Reality*, volume 125, pages 13–18. IOP Press, 2007.

[48] G. Zöller, V. Wall, and O. Brock. Active acoustic contact sensing for soft pneumatic actuators. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7966–7972. IEEE, 2020.

[49] H. Dai, N. Pears, W. Smith, and C. Duncan. Statistical modeling of craniofacial shape and texture. *International Journal of Computer Vision*, 128(2):547–571, 2020.

[50] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006.

[51] W. E. Lorensen and H. E. Cline. Marching cubes: A high resolution 3D surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998.

[52] J. Qian, T. Weng, L. Zhang, B. Okorn, and D. Held. cloth region segmentation for robust grasp selection. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 9553–9560. IEEE, 2020.

[53] J. Chapman, G. Gorjup, A. Dwivedi, S. Matsunaga, T. Mariyama, B. MacDonald, and M. Liarokapis. a locally-adaptive, parallel-jaw gripper with clamping and rolling capable, soft fingertips for fine manipulation of flexible flat cables. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 6941–6947. IEEE, 2021.

[54] J. M. Loomis and S. J. Lederman. Tactual perception. *Handbook of perception and human performances*, 2(2):2, 1986.

[55] R. Choudhury, K. M. Kitani, and L. A. Jeni. tempo: Efficient multi-view pose estimation, tracking, and forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14750–14760, 2023.

[56] D. C. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

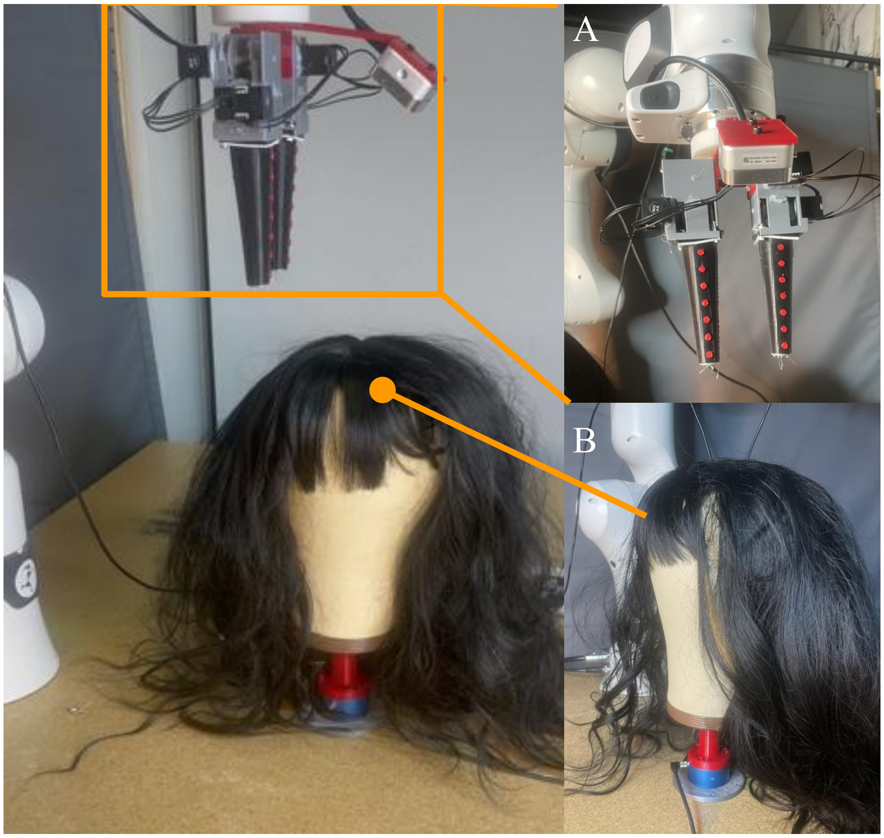

Figure 8: Experimental setup for evaluating the proposed MOE in interaction with a force-sensorized mannequin head. A: Two-fingered MOE soft manipulator with an RGBD camera. B: Mannequin head with a wig and 6-axis force sensor at its base.

A MOE Interaction Forces

We hypothesized that soft robotic manipulators would be safer and more comfortable for the human subject in hair manipulation and close-contact tasks.

Toward evaluating the hypothesis, we compared the forces experienced by the force-sensorized mannequin head with open-loop experiments, where a rigid parallel jaw gripper (FE Gripper, Franka Robotics) and the proposed MOE moved to a specified depth (2.0 mm, 4.0 mm, 6.0 mm) into the hair to grasp. The depths are measured with respect to the position where the robot is barely making contact with the hair to account for different lengths of the end-effectors. As the robot followed specified trajectories, we measured forces at the mannequin head base. After the grippers grasped the hair, the robot hand moved up to lift the grasped bundle of hair. We then measured the minimum packing perimeter of the bundle of hair. Figure 9 shows a sample result and the experimental procedure. Figure 11 shows the forces and torques experienced by the force-sensorized mannequin head.

Lower forces and torques experienced by the mannequin head could indicate reduced discomfort if applied to a human subject. Concurrently, a hair-care robot will need to be able to grasp hair that may be close to the scalp, which will likely result in higher forces experienced by the mannequin head. Then, we note that an ideal hair-care robot must be able to grasp hair effectively while also applying minimal force on the head. Table 5 reports the maximum force experienced by the head at varying depths and the amount of hair grasped.

We note that at 6.0 mm depth, the rigid end-effector exerts 7.67 N of force on the mannequin head. At the same depth, MOE applied 1.98 N of force. This constitutes a 74.1 % reduction in the maximum force applied to the head. Meanwhile, on the grasped hair metric, MOE grasped approximately 10 % less hair. A potential explanation of this marginal decrease in the amount of hair grasped is

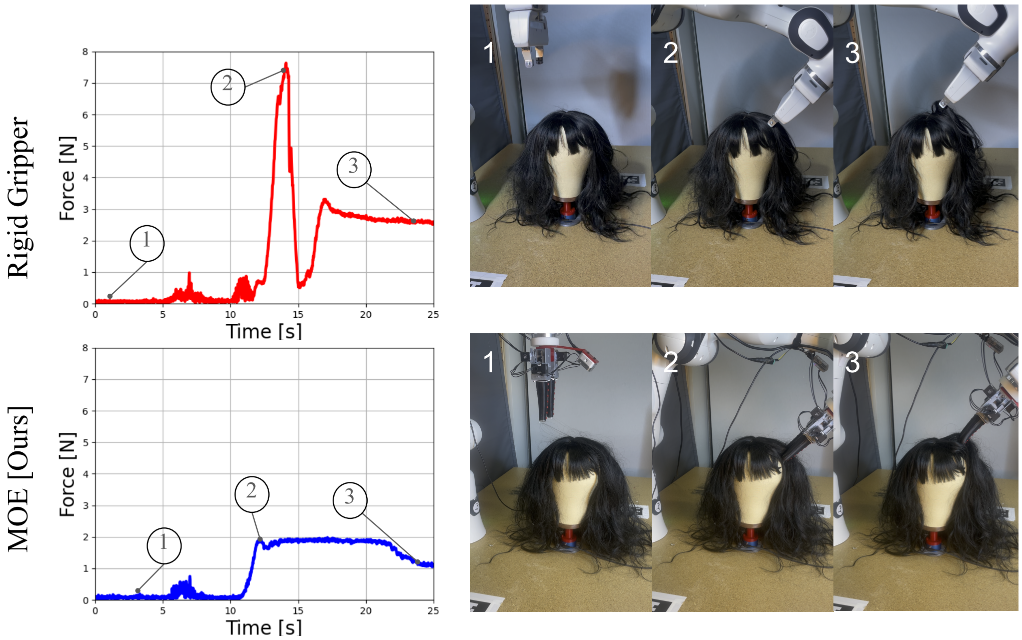

Figure 9: Hair grasping evaluation task experimental procedure and sample result at 6.0 mm depth. Top: experienced net forces and key frame images of the experiment with a baseline rigid gripper. Bottom: experienced net forces and key frame images of the experiment with the proposed MOE.

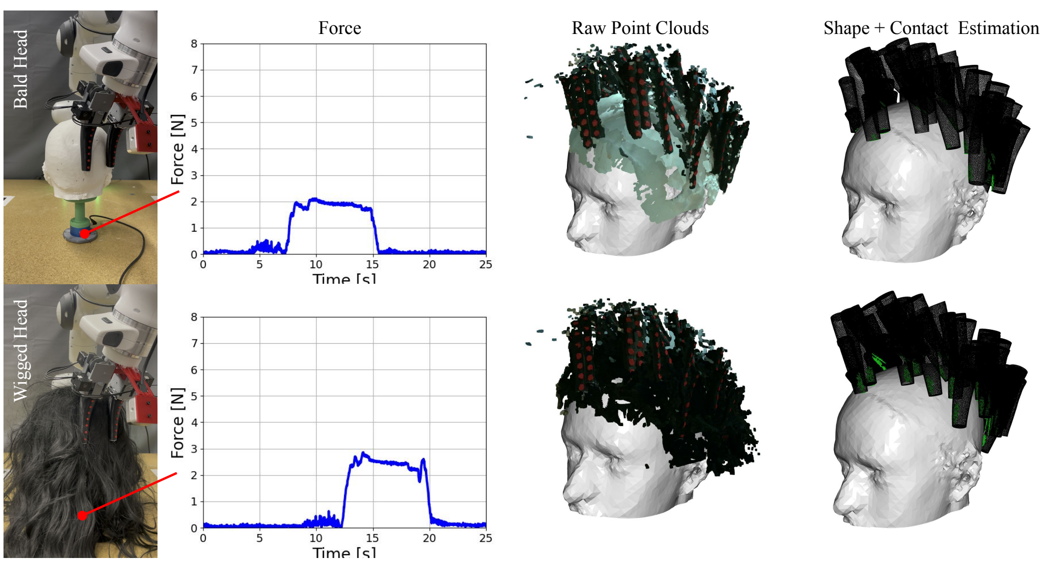

Figure 10: **Contact estimation experiment results.** The experimental setup for the head contact estimation experiments where a 3D-printed head is mounted on a force-torque sensor. The net force readings are plotted, showing the interaction forces experienced by the head. We visualize the registered wrist-mounted RGB-D camera point clouds from the 30 contact conditions, as well as the predicted MOE shape and contact points on the head. We show results for the head with and without a wig.

that the compliance of MOE allowed some of the grasped hair to be pried away as the end-effector moved away. This is partially supported by the fact that as the rigid gripper moved away from the head, the mannequin head experienced large changes in the forces applied, indicating possible hair-

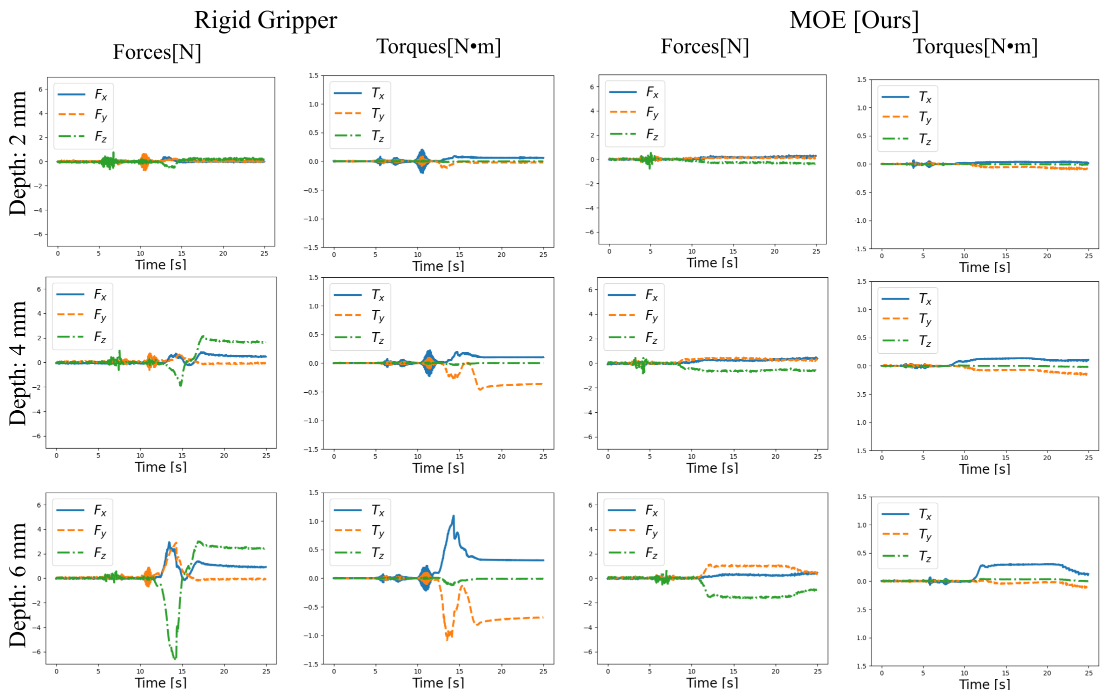

Figure 11: Sample set of YCB and a single headspace meshes simulated in contact with MOE. Relative poses were randomized to diversify the dataset.

End-effector	Depth (*mm*)	Performance Metrics	
		Max Force (*N*, ↓)	Grasped Hair (*mm*, ↑)
Rigid	2.0	1.11	4.0
	4.0	3.38	20.0
	6.0	7.67	25.0
MOE	2.0	**1.09**	**5.0**
	4.0	**1.38**	**18.7**
	6.0	**1.98**	**22.5**

Table 5: Hair Grasping Evaluation.

pulling by the end-effector. This change in forces as the robot hand moves away is not as evident in experiments with MOE.

B MOE Shape Estimation

As-Rigid-As-Possible (ARAP) involves minimizing the energy function E_{ARAP}, which is defined as the following:

$$E_{\text{ARAP}}(S, S') = \sum_{k=1}^{|E|} \min_{R \in SO(3)} \sum_{e_{i,j} \in E} w_{i,j} \| e'_{i,j} - R e_{i,j} \|.$$

We can then find the solution mesh that minimizes E_{ARAP} with an iterative local-global optimizer. Minimizing E_{ARAP} as is with sparse handle points on surface meshes can result in undesirable surface artifacts such as folds.

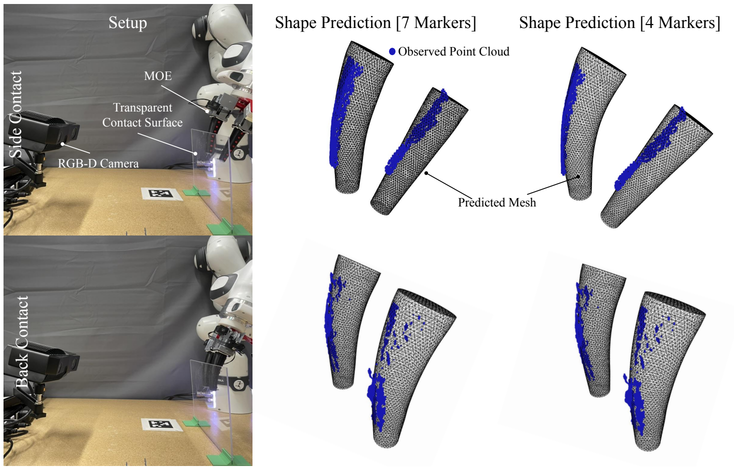

Figure 12: Shape reconstruction with 7 and 4 markers. For the 4 markers case, every other points were removed

Minimizing the E_{ARAP} over a tetrahedral mesh can prevent these artifacts by implicitly applying soft volumetric constraints that prevent such artifacts from forming. However, operating over tetrahedral meshes is more computationally expensive which is especially undesirable in the context of real-time robotic tools.

Instead, a modification of ARAP to include a penalty on the rotations of the neighboring edges produces more intuitively physically admissible results. The new energy to minimize is formulated as

$$E_{\text{smoothed}}(S, S') = \min_{R_1,\dots,R_m} \sum_{k=1}^{m} \Big(\sum_{i,j \in e_k} c_{ijk} \| e_{ij} - R_k e_{ij} \|^2 + \lambda \hat{A} \sum_{e_l \in N(e_k)} w_{kl} \| R_k - R_l \|^2 \Big).$$

We note that minimization of E_{smoothed} with $\lambda = 0$ results in the minimization of E_{ARAP}. We consider the vertices corresponding to the keypoints $p_{1,\dots,|p_k|}$ are constrained to the new positions based on the predicted key-point positions, and the rest of the mesh vertex positions are moved to minimize E_{smoothed}.

C Shape Estimation Evaluation

Prior work has shown that the rigidity and rotation regularization of the ARAP formulation as presented in Section 4.2 generally produces more physically admissible deformed soft bodies, compared to end-to-end learning-based methods [20]. A key difference in our implementation of the ARAP-based soft robot reconstruction is that the wrist-mounted RGB-D camera can only observe one side of MOE's soft surface. The underlying assumption with such an implementation choice is that the observation of one side of MOE can directly inform us about the changes to the state of the other side. As a consequence, we also assume that the cross-section of MOE's fingers remains largely the same, to allow us to infer the opposing surface's transformation. This assumption is supported by previous works in mechanics-based modeling and validation of tendon-driven soft robotic manipulators [24].

We validate shape fidelity and consistency on the side of MOE that is normally occluded from the wrist-mounted RGB-D camera, as shown in Figure 12. We place a high-resolution RGB-D camera (Zivid, One Plus) in a third-person view facing MOE, from either its side or back, to capture the side that is normally unobserved in our pipeline. We also place a clear acrylic sheet facing the third-person view RGB-D camera. This setup allows us to deform MOE against the clear sheet with a

		Performance Metrics	
Contact Condition	# of Keypoints	Mean Uni. CD (*mm*, $\downarrow$)	Max Uni. CD (*mm*, $\downarrow$)
Side	7	1.16	3.19
Side	4	1.23	3.47
Back	7	1.17	3.18
Back	4	1.19	3.35

Table 6: MOE Shape Estimation Evaluation.

large contact surface, while remaining fully observable to the third-person view RGB-D camera. We present the average and maximum unidirectional Chamfer Distance (CD) results from third-person RGB-D point cloud to the complete estimated shape, for both side and back contact conditions, in Table 6. We can observe that the shape estimation average CD error is small at 1.16-1.17 mm for the two contact conditions. Notably, the error is smaller than the 4.89 mm best average CD error reported in [20]. Such results highlight a potential advantage of directly observing keypoint movements with wrist-mounted cameras compared to indirectly inferring keypoint movements.

We also experiment with testing the robustness of the MOE shape estimation module by removing markers from being considered during ARAP mesh optimization. With 4 markers, we note a marginal increase in both average and maximum CD errors from when the shape estimation module considered the full set of 7 markers for each finger. The relatively small change in performance highlights the robustness of the shape estimation module, which can be partially attributed to the well-tuned smoothing penalty to produce meshes that conform well to soft body mechanics.

D GNN Training Details

We trained the GNN-based contact estimation model to label the vertices in the deformed mesh $i \in V'$ with the weighted cross-entropy loss:

$\mathcal{L} = -\sum_{i=1}^{N} \left[w_C \cdot y_i \cdot \log(p_i) + \cdot (1 - y_i) \cdot \log(1 - p_i) \right]$

where y_i denotes the label for the vertex i and p_i is the output probability and w_C denotes the weight for the contact points. We trained with the following parameters:

- Learning Rate: 0.001
- Batch Size: 32
- Number of Neighbors (k): 30
- Epochs: 400
- Weight Decay: 1e-4
- Momentum: 0.3
- Learning Rate Decay:
  - *Rate*: 0.5
  - *Decay Step*: 20 epochs
- Dropout Rate: 0.5

We implemented edge convolution MLP layers have the following hidden layers: [64, 64], [64, 128], [128, 256]. After the edge convolution layers, the concatenated features are processed by fully connected MLP [512, 256].

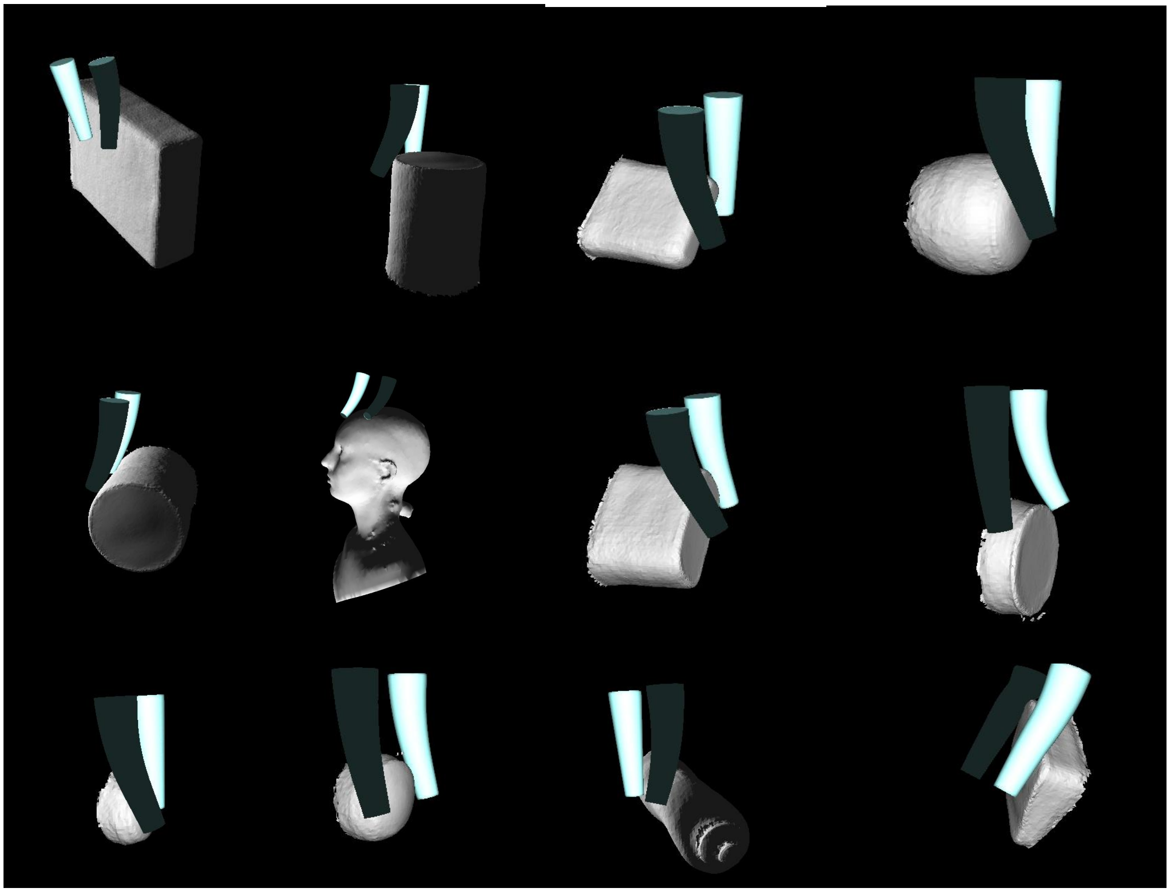

Figure 13: Sample of the simulated scenes for generating the training data for MOE contact estimation.

E Surface Reconstruction

We use *GPyTorch* to train ExactGPs with Radial Basis Function (RBF) Kernels on a single GPU with a sparse grid to fit within the memory of a single RTX 4090. We have shown effective extrapolation capabilities of GPs by generating SDFs at twice the density during inference using CPU.

The task-dependent GP-based surface reconstruction pipeline follows the following steps:

1. Pretrain a GP on a prior mesh that is dependent on the task to be done. The objective of the GP is to take a grid of points $(50 \times 50 \times 50)$ and compute the SDF with respect to the surface of the mesh. Since the GP is trained for 5000 epochs, this one-time process is slow and takes about 30 mins on a single GPU. Due to sparse discretization, the reconstructed prior is not watertight and results in holes in the mesh.
2. Next, given a set of real-world contact points, fine-tune the GP on the new points. This process is much faster and takes about 100 epochs to train.
3. Finally, we create a dense grid and query the GP to obtain the SDF values of individual points. Then we implement a Marching Cubes Algorithm to find the zero-level set of the SDF. To reconstruct the final mesh, we use Poisson Surface Reconstruction from Open3D and show the posterior reconstruction as wireframes overlayed on top of the prior reconstruction in Figure 10 of the paper.

However, grid formulation (1) is limiting for surfaces that are not uniformly distributed. This is crucial because reconstructing the head is relatively easy due to the approximately 1:1:1 aspect ratio.

Since the hand reconstruction grid is non-uniform with a 2:20:1 aspect ratio, a uniformly distributed grid ($50 \times 50 \times 50$) can not be directly used. To address this issue, we sample an extremely dense grid of shape ($200 \times 200 \times 200$), and randomly sample $30,000$ points and follow the same pipeline as above. This significantly improves reconstructions and enables us to leverage the expressivity of Gaussian Processes on non-linear surfaces with higher fidelity, compared to uniformly generated dense grids. We obtained the prior mesh of the arm with a language-conditioned mesh generator (GENIE, Luma Labs) while the prior head shape was obtained from randomly sampling the cranio-facial shape dataset.

E.0.1 No Prior

A naive approach for surface reconstruction is a non-probabilistic method by fitting a sphere to the contact points collected in the simulation. Given a set of points $\{\mathbf{P_i}\}_{i=1}^N$, where $\mathbf{P_i}$ is the i^{th} point in 3D space, the objective function for fitting a sphere to the points is defined as

$$f(\mathbf{c}, r) = \sum_{i=1}^{N} \left(\sqrt{\sum_{j=1}^{3} (P_{ij} - c_j)^2} - r \right)^2, \tag{5}$$

where $\mathbf{c} \in \mathcal{R}^3$ represents the center of the sphere, and r is the estimated sphere's radius. The initial guess for the optimization is

$$\mathbf{c_0} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{P_i}, \tag{6}$$

$$r_0 = \frac{1}{N} \sum_{i=1}^{N} \sqrt{\sum_{j=1}^{3} (P_{ij} - c_{0j})^2}. \tag{7}$$

We solve the optimization problem $\min_{\mathbf{c},r} f(\mathbf{c}, r)$ using L-BFGS [56] to find the $\mathbf{c}$ and r that minimize $f(\mathbf{c}, r)$.

We sample a point cloud for the sphere and implement a k-d tree-based nearest neighbor search to average the residuals between the contact points and the spherical mesh. Finally, we smooth out the abrupt changes to the mesh using a smoothing Laplacian filter.

E.0.2 Spherical Prior

Gaussian Processes have been extensively studied for implicit surface reconstruction in the literature [14, 15, 35]. We implement a modified version of GPIS that runs on GPU, to represent the signed distance functions (SDFs) of the head without needing surface normals. Generally, active exploration algorithms assume an initial condition of uniformly distributed points in a grid. Every measurement reduces the uncertainty until the final shape of the object is represented by the GP mean.

We fit a spherical mesh to the these points, and use this sphere as a prior to train the GP over a dense 3D array of grid points encompassing the mesh. The SDF values for each point P_i are computed as

$$SDF(P_i) = r - \|P_i - \mathbf{c}\|. \tag{8}$$

Given a set of dense grid points $\mathbf{X}$ and corresponding SDF values $\mathbf{Y}$, the GP model is defined as

$$f(\mathbf{x}) \sim \mathcal{GP}\left(c, \sigma^2 \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2l^2}\right)\right), \tag{9}$$

where σ^2 is the variance l is the length scale of the Radial Basis Function (RBF) kernel, with the observation model

$$y = f(\mathbf{x}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_n^2) \tag{10}$$

where n is the number of training points. The training objective is to maximize the marginal log-likelihood loss

$$\begin{aligned}\log p(\mathbf{Y}|\mathbf{X}) = &-\frac{1}{2}\mathbf{Y}^\top(\mathbf{K}+\sigma_n^2\mathbf{I})^{-1}\mathbf{Y} \\ &-\frac{1}{2}\log|\mathbf{K}+\sigma_n^2\mathbf{I}| - \frac{n}{2}\log 2\pi,\end{aligned} \tag{11}$$

where $\mathbf{K}$ denotes the covariance matrix constructed using an RBF kernel over the training inputs in $\mathbf{X}$. Once we have a spherical prior, the GP is updated with the contact point information to obtain the posterior SDF. Finally, the head mesh is reconstructed using Poisson Surface Reconstruction (PSR) on a point cloud obtained by running the Marching Cubes Algorithm (MCA) over the zero-level set of the SDF.