Model-based Visual Contact Localization and Force Sensing for Compliant Robotic Grippers with Generalization to Unseen Objects

Kaiwen Zuo[1], Shuyuan Yang[1] and Zonghe Chua[1]

***Abstract*— Grasp force estimation can help prevent robots from damaging delicate objects during manipulation and improve learning-based robotic control. Integrating force sensing into deformable grippers requires negotiating trade-offs in cost, complexity, mechanical robustness, and performance. With the growing integration of RGB-D wrist cameras into robotic systems for control purposes, camera-based techniques are a promising solution for indirect visual force estimation. Current approaches mostly utilize end-to-end deep learning, which can be brittle when generalizing to new scenarios, while existing model-based approaches are unsuited to grasping and modern grasper geometries. To address these challenges, we developed a model-based visual force sensing approach integrating an iterative contact localization with generalization to unseen objects. The system extracts structural key points from wrist camera RGB-D images of deforming fin-ray-shaped soft grippers, and uses these key points to define parameters of an inverse finite element analysis simulation in Simulation Open Framework Architecture. A mesh reconstruction pipeline and visual iterative contact localization algorithm were developed to dynamically update the contact location in the simulation despite visual occlusion of the contact point for unseen objects. Our system demonstrated an average root mean square error of 0.23 N and normalized root mean square deviation of 2.11% during the load phase, and 0.48 N and 4.34% over the entire grasping process when interacting with different objects under various conditions, showcasing its potential for real-time model-based indirect force sensing of soft grippers.**

I. Introduction

Soft grippers are widely used in robotic manipulation due to their inherent safety and adaptability. These properties, together with their high mechanical robustness, make them widely used in learning-based dexterous manipulation. In this application, providing training demonstrators and the control policy access to grasp force information has been shown to enhance autonomous manipulation performance, enabling precise and reliable interactions with target objects [1], [2]. Integrating force sensing into compliant grippers requires negotiating trade-offs in cost, sensor size, sensing range, accuracy, reliability, and integration complexity. Metal strain-gauge sensors are susceptible to damage when exposed to impacts. Integrating a force sensor into a soft gripper without disrupting its natural deformation behavior during grasping [3] is also challenging. Novel force sensors, such as distributed soft tactile sensors and discrete deflection sensors, preserve the gripper's natural deformation. However, the distributed tactile sensors [4], [5] often suffer from the difficulty of decoupling external contact forces from the gripper's intrinsic deformation, while discrete sensors [6] typically have constrained sensing regions due to the physical limitation of their mechatronic design.

To address this challenge, indirect sensing approaches, such as camera-based force sensing, have been developed [7]–[12]. These approaches use RGB-D images to observe gripper deformation arising from external interaction, and are advantageous because of their relatively low cost, easy accessibility, and unobtrusiveness. Such an approach is further suited for modern learning-based dexterous manipulation, as most setups already require wrist cameras [13].

Data-driven approaches have been widely adopted for indirect visual force estimation because of their ability to learn complex representational mappings. These have typically used a neural network to observe the gripper deformation [7], [8], with some using finite-element simulation to create synthetic training data [9]. These studies reveal a shortcoming of data-driven methods, which is their dependency on the quality and diversity of training data, which can significantly limit their generalizability. Moreover, the performance of such approaches relies on the quality of the visual features observed from the deforming gripper. To ensure accuracy, past works have positioned the camera directly parallel to the gripper's main deformation plane, which can significantly limit the robot's workspace in real-world applications [8]–[10]. Alternatively, mounting the camera on the robot's wrist increases the feasible workspace but typically results in reduced estimation accuracy [7].

Model-based approaches, such as inverse finite element analysis (iFEA), can provide more robust out-of-distribution performance. Reddy et al. [11] estimated both the actuation and the grasp force on miniature compliant grippers by solving Cauchy's problem in elasticity. However, their method cannot be implemented in real time due to its stated computational inefficiency. Zhang et al. [14] addressed this limitation by using reduced-order approximations and a Quadratic Programming (QP)-based solver in Simulation Open Framework Architecture (SOFA) [15]. Their approach focused on estimating external forces acting on a soft robot with *visible* environmental contact locations. However, when grasps are viewed from an angled wrist camera, contact locations are often occluded.

In this letter, we present an approach to address the limitations of existing camera-based methods, including their high sensitivity to extracted features and the requirement for visibility of contact locations. We apply real-time iFEA simu-

This work is supported by the Office of Naval Research (ONR) via **********

[1]K. Zuo, S. Yang and Z. Chua are with the Department of Electrical, Computer, and Systems Engineering, Case Western Reserve University, Cleveland, OH 44106, USA kxz365@case.edu, sxy841@case.edu,zxc703@case.edu

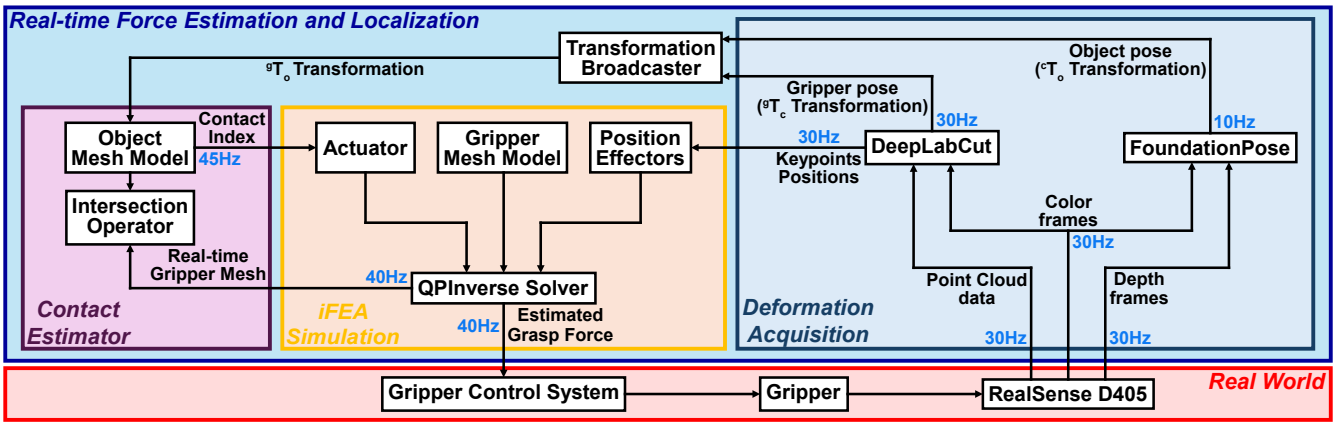

Fig. 1. Flow chart for the contact localization and force sensing system pipeline with the corresponding frame rate for the key components.

lations using SOFA to the problem of grasp force acquisition for fin-ray-shaped soft grippers, specifically dealing with non-visible contact interactions. Our approach introduces the following contributions:

1) A learning-based key point acquisition pipeline to reduce the system's sensitivity to the quality of extracted features.
2) A novel reconstructed mesh calibration pipeline and model-based iterative contact localization algorithm capable of handling occluded contact locations and pre-estimating the contact location for unseen object.
3) Experimentally validated accuracy through both static and on-robot experiments, supported by further demonstrations of the usefulness and effectiveness of the contact estimator.

II. Materials and Method

A. System Overview

Two fin-ray-shaped soft gripper jaws were mounted to an electromechanical base and observed by a wrist-mounted RGB-D camera. The structural key points of the grippers are segmented from the RGB-D image stream using Deeplabcut (DLC) [16], a deep neural-network-based skeletal pose estimation framework that leverages transfer learning. Their 3D positions are input to iFEA simulations developed using SOFA. The image stream is also used in conjunction with a reconstructed mesh of the grasped object to localize its pose relative to the gripper jaws via FoundationPose [17]. Using the object and gripper mesh intersection, a model-based iterative contact localization algorithm computes the contact position between the grasped object and the gripper. This contact estimate is provided to the simulation, and the grasp force is then computed based on the acquired constraints. The flow chart of the system pipeline and the corresponding frame rate of the key components are shown in Fig. 1. The whole system runs at 10 Hz, limited by the Foundation pose used for the contact estimator. However, because our algorithm can pre-estimate the contact location, the system runs at 30 Hz in scenarios without contact shifts during grasping, with the frame rate then constrained by the RGB-D camera. The limited frame rate of the contact estimator only affects the gripper's approaching speed toward the object, while the frame rate of the iFEA simulation remains sufficient for force control during dynamic grasping.

B. Gripper Geometry and Fabrication

Two fin-ray-shaped grippers were 3D-printed using thermoplastic polyurethane (TPU) 95A. TPU 95A was used because it is easily accessible to purchase online, and has a relatively large linear region in its stress-strain curve [18]. The latter property made it compatible with the QP Inverse Solver from the SofaRobots Inverse plugin [15], which was developed under the assumption of linear elasticity. The geometry and size of our gripper are similar to those of the Festo adaptive gripper (DHAS-GF-80-U-BU), but with thinner bands and ribs to achieve greater compliance. The effective compliance was evaluated by displacing a 20mm-diameter cylinder 10 mm at the gripper midpoint, and showed our gripper achieving two times greater compliance than the Festo gripper. We installed the gripper on an SSG-48 adaptive electric gripper base (Source Robotics, Zagreb, Croatia) as shown in Fig. 2A.

C. SOFA Simulation Setup

Left gripper (GL) refers to the gripper jaw located on the left side of the camera view, and right gripper (GR) refers to the one on the right, as shown in Fig. 2B. Both grippers use an identical mesh with 522 vertices, chosen to balance simulation accuracy and computational cost. The simulation for each gripper runs in parallel, enabling a simulation frame

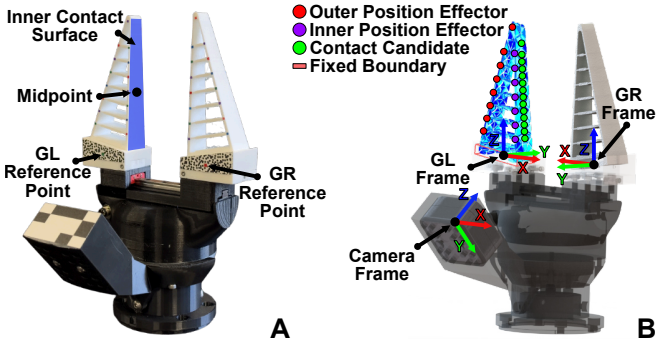

Fig. 2. Physical dual-jaw gripper and its digital twin. (A) Physical dual-jaw gripper with the RGB-D camera mounted on its wrist. (B) CAD model of the dual-jaw gripper, the left gripper is overlaid with the iFEA simulation mesh, and its relevant simulation boundary conditions. The outer and inner position effectors and contact candidates are indicated.

rate up to 40 Hz. The simulation takes the key point positions and the contact location as inputs and outputs the estimated grasp force, as shown in Fig. 1.

To build the simulation in SOFA, we define two constraints: position effector constraints and an actuator constraint. In each iteration of the simulation, the QP Inverse Solver will compute the force applied to the gripper at the actuator position that minimizes the distance between the position effectors and their corresponding targets. As shown in Fig. 2B, we chose the nodes that correspond to structural key points on the physical gripper to be the position effectors, and we further categorize them into outer and inner effectors based on their positions with respect to the inner contact surface. These key points are thus the targets of their respective position effectors and are captured by our key point acquisition pipeline as described in Sec. II-D.

For the actuator constraint, we chose to use the force point actuator rather than the force surface actuator because it provided more stable solutions during preliminary tests. The position of the force point actuator is crucial to the simulation, as an incorrect contact condition can lead to inaccurate force estimation. SOFA allows us to initialize a set of contact candidates, and the positions of the contact candidates are updated iteratively as the gripper mesh model deforms. The actuator constraint can be dynamically changed by mounting the force point actuator on a different contact candidate during simulation. Here, we selected 14 linearly arranged points along the central axis of the inner contact surface as contact candidates whose positions in the undeformed state are denoted as $\{\mathbf{a}_i\}$, and the actuator is initialized on the middle contact candidate.

Another key factor in the simulation is ε, which represents the weight of the deformation energy term in the QP inverse solver objective function [15], given as

$$J = \tfrac{1}{2}\,\lambda_a^\top \big(W_{ea}^\top W_{ea} + \varepsilon W_{aa}\big)\lambda_a + \lambda_a^\top W_{ea}\,\delta_e^{\text{free}}. \qquad (1)$$

The solver accounts for more internal strain and less shape transformation with higher ε. According to the datasheet of the Festo gripper, the effective stiffness of the fin-ray-shaped gripper varies at different contact positions, indicating that the ratio of internal strain and shape transformation is dependent on where contact occurs. Therefore, ε was empirically tuned for each contact candidate. At the beginning of each iteration, the actuator is mounted on the estimated contact candidate, and the corresponding ε value is updated accordingly.

D. Key Points Acquisition and Calibration

3D positions of structural key points were identified using DLC [16] from RGB-D data captured via an Intel RealSense D405 at 30 Hz. DLC requires only minimal labeled data (typically 50-200 frames) to adapt to a specific skeletal pose estimation task. To adopt DLC to our application, we collected and annotated 300 frames of the gripper grasping cylinder and cube, as in the on-robot evaluation. We labeled 15 structural key points and one reference point per gripper in each frame. When the key point is occluded, the corresponding label remains unannotated. Reference points as defined in Fig. 2A were used to localize the gripper jaw during the grasping process. These data were used to fine-tune a pretrained DLC model using the official API, which allows adjusting the output dimensionality without directly modifying the network architecture.

In each frame of the camera stream, DLC predicts the pixel indices of the key points and reference points in RGB images with their corresponding likelihood. Then, we obtain the position data of the key points $\{{}^{\mathrm{c}}\mathbf{q}_{\mathrm{kp},i}\}$ and reference point ${}^{\mathrm{c}}\mathbf{q}_{\mathrm{rp}}$ in the camera frame for each gripper by referring the pixel indices to the point cloud data. The key points are mapped from the camera to the gripper jaw frame using the transformation matrix ${}^{\mathrm{g}}\mathbf{T}_{\mathrm{c}}$, which is computed by chaining the camera-to-gripper-base transformation ${}^{\mathrm{r}}\mathbf{T}_{\mathrm{c}}$, and the gripper-base-to-jaw transformation ${}^{\mathrm{g}}\mathbf{T}_{\mathrm{r}}$. The gripper base frame shares the orientational alignment as the jaw frame, but its origin tracks the reference point, which translates as the gripper jaws open and close. ${}^{\mathrm{g}}\mathbf{T}_{\mathrm{r}}$ and ${}^{\mathrm{r}}R_{\mathrm{c}}$, the rotational component of ${}^{\mathrm{r}}\mathbf{T}_{\mathrm{c}}$, are specified from CAD models.

Capturing all the key points allows for more constraints to be specified in the simulation, which can improve the accuracy of force estimations. However, in most grasping scenarios, some key points are occluded. Furthermore, some key points may have incorrect position estimates due to misidentification in DLC or noisy point cloud reconstruction. These can lead to infeasible position effector target specifications that can contribute to inaccurate and unstable solutions. To prevent this, we applied keypoint detection confidence thresholding followed by spatial bounding box filtering for all key points. The threshold confidence is determined empirically based on the model output to retain most visible key points while filtering out all the occluded ones. We set the spatial bounding box to contain the workspace for key points to filter out the incorrectly inferred ones in the background.

When updating the key points at the beginning of each iteration of the simulation, the simulation controller receives an indexed list of incorrect key points and deactivates their corresponding position effector. We use a binary switch instead of continuous smoothing, since occluded keypoints can drift unpredictably onto the object surface or into the background, which makes the continuous smoothing unreliable.

E. Contact Estimator

The contact position serves as an essential constraint in the simulation. Unlike in existing approaches [14], our perspective camera view obscures the contact position in the point cloud data. Additionally, the simulated gripper deformation and force depend on accurate contact localization, which in turn depends on the gripper deformation. This phenomenon will be more pronounced with larger deformation and will reduce the performance of naive closest point approaches [14]. To address these, we developed a novel model-based iterative contact localization algorithm. Since pose estimation is not the primary contribution of this work, we use FoundationPose [17] to estimate the object pose, by providing it with a reconstructed mesh of the grasped object. This object mesh is also used as a digital twin, whose pose is described

Algorithm 1 Iterative Contact Localization Algorithm

Input: undeformed vertex matrix $\mathbf{V}$, object mesh $\mathbf{M}_\mathrm{o}$, candidate set $\{\mathbf{a}_i\}$, offset l_0, contact status flag $\mathbf{S}$ = False
for each iFEA simulation iteration **do**
  Compute $({}^{\mathrm{g}}R_\mathrm{o}, {}^{\mathrm{g}}t_\mathrm{o}) \leftarrow ({}^{\mathrm{g}}\mathbf{T}_\mathrm{c}, {}^{\mathrm{c}}\mathbf{T}_\mathrm{o})$
  Get $\overline{\mathbf{V}}$ and reconstruct $\overline{\mathbf{M}}_g \leftarrow \overline{\mathbf{V}}$
  $\hat{\mathbf{M}}_\mathrm{o} \leftarrow \mathbf{M}_\mathrm{o}$, and apply rotation ${}^{\mathrm{g}}R_\mathrm{o}$ to $\hat{\mathbf{M}}_\mathrm{o}$
  if S is True **then**
    Apply translation ${}^{\mathrm{g}}t_\mathrm{o}$ to $\hat{\mathbf{M}}_\mathrm{o}$
  else
    Update ${}^{\mathrm{g}}t_\mathrm{o}$ and apply to $\hat{\mathbf{M}}_\mathrm{o}$
  end if
  Compute intersection $\mathbf{M}_\mathrm{in} \leftarrow \overline{\mathbf{M}}_\mathrm{g} \cap \hat{\mathbf{M}}_\mathrm{o}$
  if $\mathbf{M}_\mathrm{in} \neq \emptyset$ **then**
    Find $\mathbf{a}_\mathrm{c}$ and mount to the c^{th} candidate
    $\mathbf{S} \leftarrow$ True
  else
    $\mathbf{S} \leftarrow$ False
  end if
  The QP Solver computes the simulated force.
end for

in the frame of the deformed gripper jaw.

1) Mesh Reconstruction: The object mesh is initially generated using the SAM-3D Objects model [19]. Given an RGB image and the object mask, the model reconstructs the object mesh in a canonical space and jointly estimates the object scale and pose. However, because the model doesn't take the camera intrinsics as input, there is a severe scale mismatch between the mesh and the physical object, leading to inaccurate pose estimation and contact localization. To address these, a novel scale calibration workflow is developed.

With an RGB-D camera, we can segment the object point cloud from the point cloud data using the object mask. The reconstructed mesh is first scaled and transformed into the camera frame based on the model's output. Since the mesh represents the complete object geometry, whereas the acquired point cloud only contains partial observations, directly applying the iterative closet point (ICP) algorithm yields poor alignment. To obtain the partial view of the reconstructed mesh, we projected it onto a pixel grid defined on the XY plane of the camera frame, with the origin aligned to the camera frame origin and the grid resolution determined by the camera intrinsics. We obtain a partial mesh by preserving the points closest to the origin for each pixel. Then we perform ICP twice between the partial mesh and the captured point cloud using Open3D, enabling scale adjustment only at the second ICP. We apply the resulting scale factor to the complete reconstructed mesh and generate a watertight triangular mesh via Poisson surface reconstruction.

2) Iterative contact localization algorithm: The deformed mesh of the gripper jaw is reconstructed using the real-time vertex matrix $\overline{\mathbf{V}} \in \mathbb{R}^{n_\mathrm{v} \times 3}$, which is updated at each simulation iteration. Here, n_v denotes the number of vertices. FoundationPose predicts the transformation matrix from the object frame to the camera frame, denoted as ${}^{\mathrm{c}}\mathbf{T}_\mathrm{o}$. As discussed in Sec. II-D, the transformation matrix from the camera frame to the gripper jaw frame, ${}^{\mathrm{g}}\mathbf{T}_\mathrm{c}$, is known. By chaining ${}^{\mathrm{g}}\mathbf{T}_\mathrm{c}$ and ${}^{\mathrm{c}}\mathbf{T}_\mathrm{o}$, the complete transformation from the object frame to the gripper jaw frame, ${}^{\mathrm{g}}\mathbf{T}_\mathrm{o}$, can be obtained, enabling the transformation of a copy of the localized object mesh $\mathbf{M}_\mathrm{o}$, denoted as $\hat{\mathbf{M}}_\mathrm{o}$, into the gripper jaw frame and consequent Boolean intersection computation with respect to the deformed gripper mesh $\overline{\mathbf{M}}_\mathrm{g}$.

When the result of the Boolean intersection, denoted as $\mathbf{M}_\mathrm{in}$, is empty, it indicates that there is no contact between the gripper and the object. Otherwise, we identify the vertex $\overline{\mathbf{v}}_\mathrm{j} \in \overline{\mathbf{V}}$ that is closest to the center of $\mathbf{M}_\mathrm{in}$. Then, the position of the j^{th} vertex in the undeformed state, $\mathbf{v}_\mathrm{j}$, can be retrieved through the undeformed vertex matrix of the gripper jaw, $\mathbf{V}$. Typically, $\mathbf{v}_\mathrm{j}$ lies inside the gripper, but contact can only occur on the inner contact surface. Thus, $\mathbf{v}_\mathrm{j}$ is projected onto the inner contact surface to obtain a point $\mathbf{q}_\mathrm{p}$. This point is used to find the closest contact candidate in the set of candidates $\{\mathbf{a}_i\}$, denoted as $\mathbf{a}_\mathrm{c}$. The force point actuator is then mounted onto the identified c^{th} contact candidate.

Since Foundation pose has a limited frame rate, which significantly affects the system's performance. To address this issue, we pre-estimate the contact position based on the relative pose between the gripper and the object. In cases where the gripper and the object are not yet in contact, their digital twins are brought into slight contact by adjusting ${}^{\mathrm{g}}t_\mathrm{o}$, the x-axis translation of ${}^{\mathrm{g}}\mathbf{T}_\mathrm{o}$, thereby enabling early estimation of the contact position. Since the solver computes the simulated force by minimizing the distance between the position effectors and their targets, the pre-estimation process does not influence the simulation results when no physical contact occurs. ${}^{\mathrm{g}}t_\mathrm{o}$ is updated accordingly as

$$ {}^{\mathrm{g}}t_{\mathrm{o},x} = l_0 + 0.5 l_\mathrm{bd}, \tag{2} $$

where ${}^{\mathrm{g}}t_{\mathrm{o},x}$ represents the translation along the x-axis. The term l_0 denotes the x-distance from the inner contact surface to the origin of the gripper jaw frame in the undeformed state, offset by 1 mm into the surface, and l_bd denotes the width of the object model along the x-axis after applying the predicted rotation, ${}^{\mathrm{g}}R_\mathrm{o}$. The pseudocode of the proposed iterative contact localization algorithm is provided in Algorithm 1.

F. Baseline Model Training

We train a neural network as a benchtop baseline for our system. Following Zhu et al. [10], we adopt a ResNet-50 backbone to encode the input images, and feed a sequence of 10 images into a transformer to predict the grasp force. The dataset consists of the dual-jaw gripper grasps on the cylinder and the cube, as in the on-robot evaluation, with three force levels, five motion speeds, and 14 distinct contact locations. In total, the dataset contains approximately 3.44 hours of video recorded at 30 Hz, which is comparable in scale to that used by Zhu et al. [10].

III. Experimental Evaluations, Results, and Discussion

A. Static Force Evaluation

To evaluate the accuracy of the proposed grasp force estimation system, we constructed a benchtop setup and

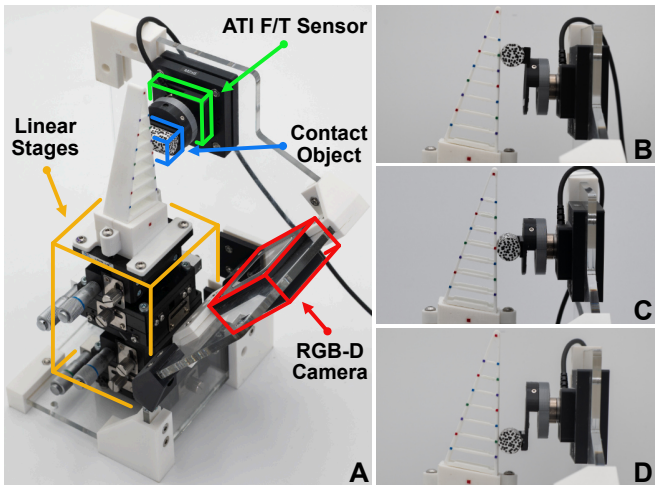

Fig. 3. Benchtop setup for static evaluation and configurations of contact positions. (A) Experimental setup. (B–D) The gripper is set to contact a 15 mm diameter cylinder at (B) upper, (C) middle, and (D) lower positions

conducted a static evaluation on a single gripper jaw, as shown in Fig. 3A. In this setup, the jaw is mounted on two 3-axis linear stages. The RGB-D camera is mounted at the same initial relative position and viewing angle with respect to the gripper as in the dual-jaw setup shown in Fig. 2A. By preserving the spatial relationship between the camera and the gripper, this benchtop configuration enables us to evaluate the system's on-robot performance under controlled conditions. As shown in Fig. 3A, an ATI Nano17 F/T sensor (ATI Industrial Automation, Apex, NC, USA) is placed beneath the grasped object to obtain the ground-truth grasp force, denoted as $\mathbf{F}_{\text{gt}}$. The grasped objects were connected to the force sensor via an adapter. The adapter was equipped with a roller bearing that allowed the object to translate along the contact surface. This reduced surface-parallel forces arising from sliding friction during gripper deformations.

The static evaluation was conducted by applying five different load levels to the gripper while interacting with three cylinders of varying diameters: 15, 25, and 35 mm. Each cylinder was tested at three distinct contact positions: upper, middle, and lower, as defined in Fig. 3B-D. The load levels were achieved by manually displacing the gripper base by 2, 4, 6, 8, and 10 mm, with 0 mm representing the initial state in which the gripper is positioned but not yet in contact with the object. Because the effective stiffness of the gripper varies across different contact positions, we chose to control the displacement of the linear platform rather than the force applied to the gripper. For each displacement measurement, the gripper is translated to its target displacement. The system is then allowed to settle for 10 seconds to reduce any transient impact and viscous effects. Then the ground-truth and simulated grasp forces, denoted as $\mathbf{F}_{\text{sim}}$, are recorded over a subsequent 10 seconds. One cycle is defined as loading the gripper base from its initial position to each displacement level, up to 10 mm, and then unloading to each displacement level, until the system returns to its initial configuration. For each contact condition, five cycles were conducted to ensure the repeatability of the results.

TABLE I

SIMULATED GRASP FORCE ACCURACY UNDER DIFFERENT CONTACT POSITIONS AND CYLINDER DIAMETERS.

Contact Position	**Ø** (mm)	**RMSE** (N)	**NRMSD** (%)	**Max Error** (N)
Upper	15	0.22 ± 0.03	1.97 ± 0.30	0.54
	25	0.52 ± 0.09	4.67 ± 0.79	1.02
	35	0.57 ± 0.05	5.20 ± 0.50	1.52
Middle	15	0.24 ± 0.02	2.16 ± 0.21	0.62
	25	0.28 ± 0.07	2.54 ± 0.59	0.98
	35	0.24 ± 0.07	2.18 ± 0.60	0.60
Lower	15	0.69 ± 0.05	6.26 ± 0.48	1.55
	25	0.62 ± 0.11	5.62 ± 0.98	1.47
	35	1.21 ± 0.15	10.93 ± 1.37	3.33

The accuracy of the system was evaluated by computing the root mean square error (RMSE) and normalized root mean square deviation (NRMSD) between the ground-truth and simulated forces. The NRMSD is calculated by normalizing the RMSE by the observed force range across all loading cycles. The normalizing force range was 11.04 N in both static force and on-robot evaluation. To quantify the system's error in the worst case, the maximum error of the simulated results during the evaluation was calculated.

The mean RMSE and NRMSD, along with their standard deviations across three test cycles, are presented in Table I. The corresponding maximum errors are also reported. We compared errors for both the load and unload phases under different contact and displacement conditions, as shown in Fig. 4A–C. It can be observed that when contact occurs at the upper or lower positions, the estimated grasp force becomes less accurate as the cylinder diameter increases, evidenced by the data points deviating further from the gray dashed unity line. In contrast, when contact occurs at the middle position, the estimation accuracy remains consistent as the cylinder diameter changes, with a mean RMSE ranging from 0.20 N to 0.26 N and mean NRMSD ranging from 1.81% to 2.35%, as shown in Table I.

Like other camera-based force estimation methods, the accuracy of our system depends on the quality of the extracted gripper features, which are the structural key points. In our scenarios, this is primarily affected by the size of the object being grasped, as larger objects tend to occlude more keypoints near the contact region. This effect becomes particularly significant at the upper contact position. As shown in Fig. 4C, our system tended to overestimate the ground-truth grasp force as the cylinder diameter increased. This is because the occlusion of the outer position effectors at the upper position consequently reduces the contribution of the corresponding constraints in the iFEA simulation. Therefore, the solver accounts for more of the lower-position constraints, where the gripper exhibits a higher effective stiffness, leading to an overestimation of the grasp force. This issue did not significantly affect the accuracy in the middle or lower contact positions, where only the inner position effectors are occluded. This phenomenon indicates that the

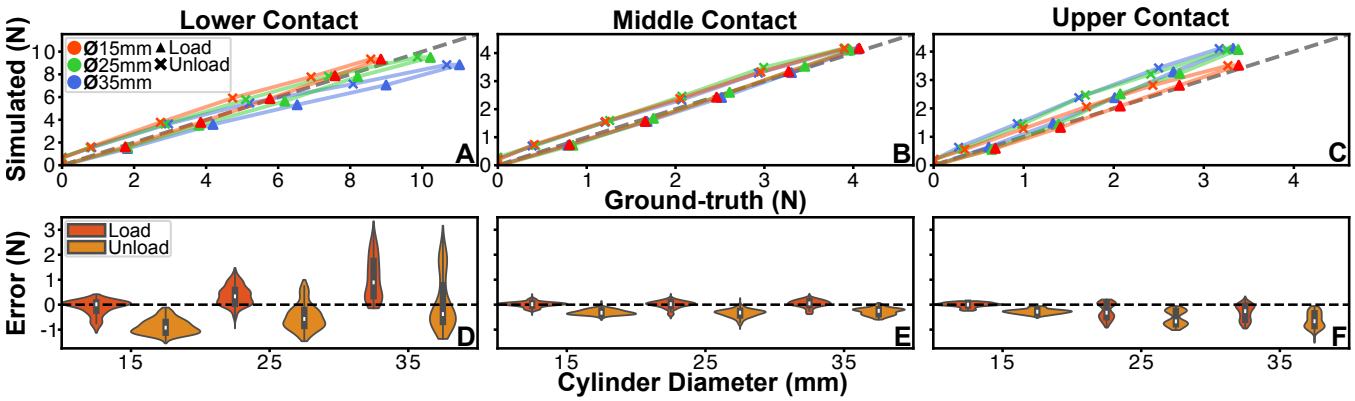

Fig. 4. Contributions of contact position and cylinder size to grasp force estimation error. Top row (A–C): Comparison between $\mathbf{F}_{\text{sim}}$ and $\mathbf{F}_{\text{gt}}$ during load and unload phases for cylinders with diameters of 15, 25, and 35 mm, under three contact positions: (A) lower, (B) middle, and (C) upper. Bottom row (D–F): Violin plots of estimation error across cylinder diameters for contact positions: (D) lower, (E) middle, and (F) upper. (D) and (F) indicate a bimodal error pattern, in which our system tends to underestimate the force at lower contact position and overestimate it at upper position. (A) and (C) further demonstrate that this misestimation occurs under large deformations.

visibility of the outer position effectors, especially for the ones on the tip, is critical to the accuracy of our system.

Our system performance is also dependent on the amount of internal strain within the gripper. During the deformation process, the gripper experienced different ratios of internal strain and shape transformation due to variations in effective stiffness across contact positions. As shown in Fig. 4A, at the lower contact position where the gripper exhibited higher effective stiffness compared to the other two positions, our system tended to underestimate the ground-truth grasp force as the displacement increased. This is because, in the regions with higher effective stiffness, the gripper has less observable shape deformation while accumulating increased internal strain. As a deformation-based system, our method has limited observability of high internal strain, leading to an underestimation of the grasping force.

We further analyzed the error distribution in different load phases, as shown in Fig. 4D–F. The results indicate that the estimation error was generally lower in the load phase than in the unload phase. This observed hysteresis suggests an unmodeled viscous effect. When contact occurred at the upper or lower positions, the error distribution exhibited a bimodal pattern. The diminished accuracy at these two contact positions both occur under high deformation, but are attributable to two distinct factors, which are the occlusion of outer key points in the upper contact scenario, and more unobservable internal strain in the lower contact scenario.

B. Contact Estimator Evaluation

In the static force evaluation, the contact position in the iFEA simulation was predefined, whereas in real-world applications, it is typically unknown. Using our iterative contact localization algorithm, the system can adapt to different objects and remain effective even when the contact position shifts during grasping. In this experiment, we aimed to evaluate the contribution of the contact estimator to the overall system performance. Due to the difficulty in directly quantifying the accuracy of the estimated occluded contact position in practice, we assess its accuracy indirectly by analyzing the improvement in grasp force estimation performance. To control for the error in object pose estimation attributable to FoundationPose, we hard-coded the pose of the cylinder by initializing its digital twin model to the known pre-contact state. The translation along the x-axis was then acquired by computing the displacement of the reference point (${}^{c}\mathbf{q}_{\text{rp}}$) from its initial position.

TABLE II

GRASP FORCE ESTIMATION ACCURACY (RMSE (N) ± STD) UNDER DIFFERENT ACTUATOR POSITIONS AND THE CONTACT ESTIMATOR.

Sim / Real	Contact Estimator	Upper	Middle	Lower
Upper	0.20 ± 0.04	0.22 ± 0.03	1.21 ± 0.10	5.91 ± 0.13
Middle	0.26 ± 0.02	0.93 ± 0.14	0.24 ± 0.02	4.93 ± 0.17
Lower	0.61 ± 0.12	4.40 ± 0.52	3.61 ± 0.53	0.69 ± 0.05

* Blue cells indicate cases where the preset contact position matched the actual contact position, while green cells indicate cases where the system was implemented with the contact estimator.

We repeated the same experimental procedure described in the static evaluation, with the gripper interacting with a 15 mm cylinder at three distinct contact positions. For each contact position, the simulation actuator was configured in four different ways: (1) fixed on the upper position, (2) fixed on the middle position, (3) fixed on the lower position, and (4) initialized at the middle position with the contact estimator activated, i.e., our method. The mean RMSE and its standard deviation are reported in Table II. The results indicate that the contact estimator is essential to the system. Incorrect fixed actuator configurations significantly degraded estimation accuracy, whereas the contact estimator achieved similar accuracy to fixed configurations where the actuator position is correctly predefined.

C. On-robot Evaluation

To evaluate the system's performance in an on-robot scenario, we conducted grasping tests on three commonly encountered object shapes. The dual-jaw gripper was mounted on a Kinova Gen3 7-DoF robotic arm (Kinova, Quebec, Canada). The test objects, a cylinder, a cube, and an asymmetric object, were 3D-printed as two halves, so that they could contain an embedded F/T sensor for $\mathbf{F}_{\text{gt}}$ measurement as shown in Fig. 5A. Using our mesh reconstruction

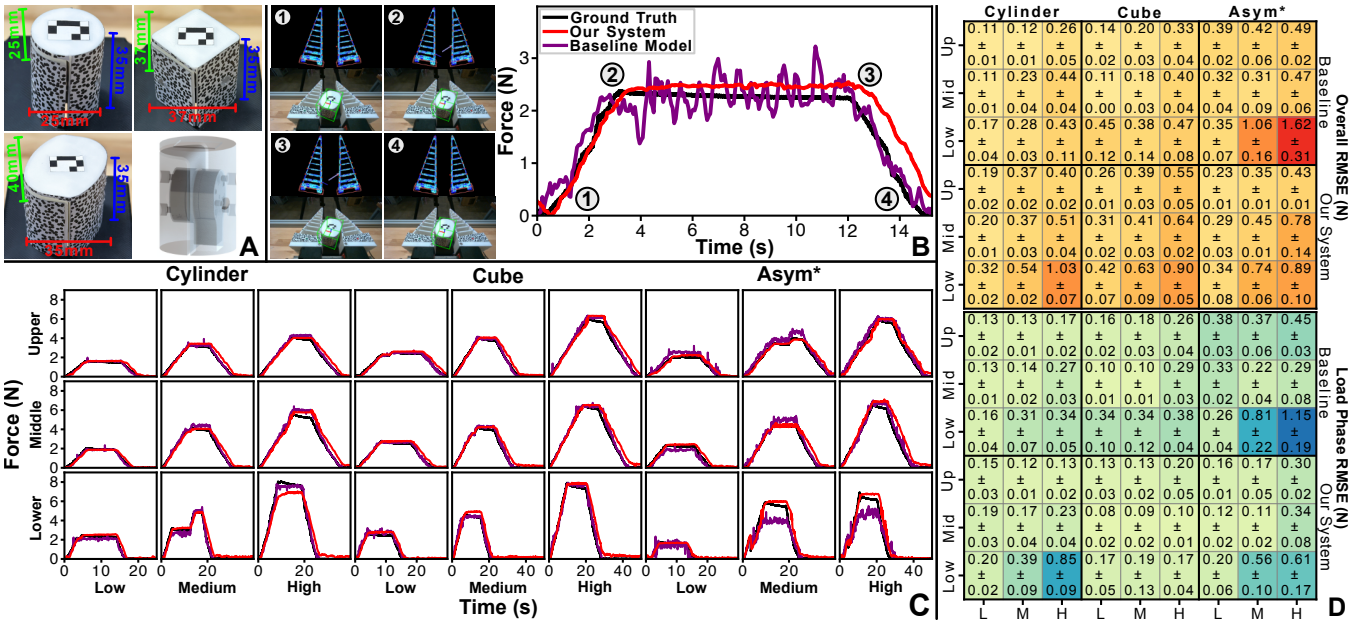

Fig. 5. Objects with built-in load cells and experimental results of the on-robot evaluation. (A) Cylinder, cube, asymmetric object with built-in load cells and CAD design under perspective view. (B) Estimation results when grasping the asymmetric object at lower position under low force level. Images of the physical grippers and their digital twins at four stages of the grasping process are shown. (C) Representative grasp force estimation results across different grasping conditions. **Asym*** indicates that the asymmetric object is unseen for both our system and baseline model. (D) Mean RMSEs along with STDs of the entire grasping process and the load phase for both our system and baseline model. L, M, and H represent low, medium, and high force level.

approach, we obtain reconstructed meshes with L2 chamfer distances of 2.36, 3.53, and 3.35 mm for the tested objects with respect to their CAD models. In contrast, the raw model-generated meshes have larger chamfer distances of 13.39, 14.69, and 15.69 mm, demonstrating the necessity of the proposed calibration process. The robot was controlled to grasp each object at the upper, middle, and lower contact positions with a grasping speed of 1 mm/s. To further investigate the system's performance under varying force levels (low, medium, and high), the gripper was actuated with incremental control currents of 300 mA, 500 mA, and 800 mA. To equilibrate the transient effects, the gripper stayed fully grasped for six seconds in each grasping process. Each force level was repeated five times. We recorded the videos of each grasping process and benchmarked our system by running the baseline model offline. We also evaluated the system's performance in scenarios where the contact position shifts during the grasping process by grasping the asymmetric object. Embedding the F/T sensor within the object allowed us to measure the ground-truth force without interfering with the natural grasping process. To ensure reliable ground-truth measurements while minimizing the impact on our system bandwidth, we used a binocular optical camera, Micron Tracker 4 (Claronav, Toronto, Canada), to establish the kinematic chain from the sensor frame to the gripper jaw frame, enabling the transformation of the sensor-measured force for validation purposes. Two optical tracking markers were attached to both the top surface of the object and the RGB-D camera mount. The estimated grasp and manipulation force, $\mathbf{F}^g_{\text{sim}}$ and $\mathbf{F}^m_{\text{sim}}$, were calculated using the method described by Yoshikawa and Nagai [20],

$$\begin{bmatrix} \mathbf{F}^g_{\text{sim}} \\ \mathbf{F}^m_{\text{sim}} \end{bmatrix} = \begin{bmatrix} \min(|\mathbf{F}_{\text{gl}}|, |\mathbf{F}_{\text{gr}}|) \\ \mathbf{F}_{\text{gl}} - \mathbf{F}_{\text{gr}} \end{bmatrix} \tag{3}$$

where $\mathbf{F}_{\text{gl}}$ and $\mathbf{F}_{\text{gr}}$ denote the estimated force on GL and GR, respectively. Due to the difficulty of obtaining reliable measurements of the ground-truth manipulation force, we only evaluated the grasp force in this experiment.

We categorized the grasping process into four stages, as defined in Fig. 5B. At Stage 1, the grippers begin grasping the object. At Stage 2, the full commanded grasp is achieved. At Stage 3, the gripper begins to release the object. At Stage 4, the object is fully released. The load phase is defined as the period between Stage 1 and 2. The mean RMSEs and standard deviations for the entire grasping process and the load phase were calculated separately for both the baseline model and our system, as shown in Fig. 5D.With an inspected ≈ 0.25s delay, except for the worst-case scenario, where the gripper was commanded to grasp the object with a high force level at the lower contact position, the overall RMSE of our system ranged from 0.19 to 0.78 N, and the load phase RMSE ranged from 0.08 to 0.56 N with high repeatability. The corresponding NRMSD ranges from 1.75 to 7.04 % over the entire grasp process and from 0.71 to 5.08 % during the load phase. The results reveal a clear pattern, in which the RMSE increased with higher force levels due to the combined effects of observing fewer outer key points, and the increasing internal strain.

Both the baseline model and our system achieve high accuracy when grasping the cylinder and cube, which are included in their training datasets (this corresponds to the DLC dataset for our system). However, when grasping the asymmetric object, the baseline model exhibits a substantial drop in accuracy, while our system maintains high performance, as shown in Fig. 5B-D. This result demonstrates the generalization capability of the data-driven method and the robustness of our keypoint acquisition pipeline. It also indicates that our

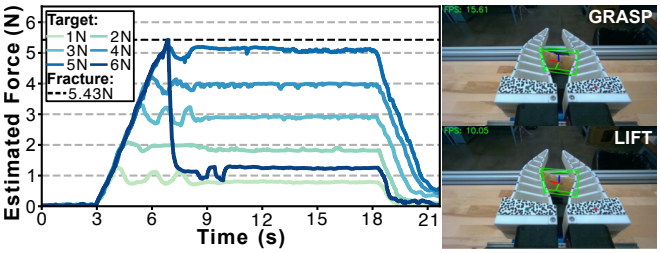

Fig. 6. Our System performance with incremental force control when grasping and lifting a potato chip. The images on the right show the gripper grasping and lifting the chip, along with the pose estimation using the reconstructed chip mesh.

system can accurately estimate the grasp force even when the contact positions shift during the grasping process. Our system's errors primarily occurred between Stages 2 and 4 of the grasping process and were likely due to the nonlinear behavior and viscosity of the soft material.

D. Grasping a Potato Chip

To demonstrate the capability of our system for force control during dynamic grasping of delicate objects, we commanded the gripper to repeatedly grasp and lift a potato chip at a closing speed of 4mm/s while incrementally increasing the force threshold in steps of 1N until fracture occurred. As shown in Fig. 6, the gripper successfully grasped and lifted the potato chip at a force level of 5N, whereas the chip fractured at 5.43 N, demonstrating the capability of our approach to perform force control during grasping of delicate objects.

IV. Conclusion and Future Work

In this letter, we presented the design and evaluation of a model-based visual grasp force sensing approach for deformable fin-ray-shaped grippers. This approach comprised real-time iFEA simulations and a learning-based key point acquisition pipeline using a RGB-D camera. A novel mesh reconstruction pipeline and model-based iterative contact localization algorithm were developed and integrated into the system, enabling reliable contact estimation when grasping unseen objects of varying shapes under occlusion. A baseline model was trained to benchmark our system. The system performance during dynamic grasping is further demonstrated by performing force control when grasping a potato chip.

Our system is compatible with other passive compliant grippers that exhibit distinct shape transformations under external contact forces by adjusting the acquisition pipeline and the parameters in the iFEA simulation. As the simulation models the material as linearly elastic, the gripper material should have an adequate linear region in its stress–strain curve over the target sensing range to maintain high accuracy due to the limited observability of internal strain. Moreover, material aging effects are not considered in the current framework. These can be addressed in the future by investigating alternative time-aware modeling approaches. While static force evaluation highlights the importance of outer key point visibility to system performance, a quantitative metric relating these two remains difficult to define due to the coupling of key point visibility and nonlinear material behavior.

References

[1] C. Cuan, A. Okamura, and M. Khansari, "Leveraging haptic feedback to improve data quality and quantity for deep imitation learning models," *IEEE Transactions on Haptics*, vol. 17, no. 4, pp. 984–991, 2024.

[2] A. E. Abdelaal, J. Fang, T. N. Reinhart, J. A. Mejia, T. Z. Zhao, J. Bohg, and A. M. Okamura, "Force-aware autonomous robotic surgery," *arXiv preprint arXiv:2501.11742*, 2025.

[3] H. Mun, D. S. Diaz Cortes, J.-H. Youn, and K.-U. Kyung, "Multi-degree-of-freedom force sensor incorporated into soft robotic gripper for improved grasping stability," *Soft Robotics*, vol. 11, no. 4, pp. 628–638, 2024.

[4] J. Qu, B. Mao, Z. Li, Y. Xu, K. Zhou, X. Cao, Q. Fan, M. Xu, B. Liang, H. Liu, X. Wang, and X. Wang, "Recent progress in advanced tactile sensing technologies for soft grippers," *Advanced Functional Materials*, vol. 33, no. 41, 2023.

[5] H. Li, Y. Lin, C. Lu, M. Yang, E. Psomopoulou, and N. F. Lepora, "Classification of vision-based tactile sensors: A review," *IEEE Sensors Journal*, vol. 25, no. 19, p. 35672–35686, Oct. 2025.

[6] G. Chen, S. Tang, S. Xu, T. Guan, Y. Xun, Z. Zhang, H. Wang, and Z. Lin, "Intrinsic contact sensing and object perception of an adaptive fin-ray gripper integrating compact deflection sensors," *IEEE Transactions on Robotics*, vol. 39, no. 6, 2023.

[7] J. A. Collins, C. Houff, P. Grady, and C. C. Kemp, "Visual contact pressure estimation for grippers in the wild," in *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2023, pp. 10 947–10 954.

[8] W. Xu, H. Zhang, H. Yuan, and B. Liang, "A compliant adaptive gripper and its intrinsic force sensing method," *IEEE Transactions on Robotics*, vol. 37, no. 5, 2021.

[9] D. De Barrie, M. Pandya, H. Pandya, M. Hanheide, and K. Elgeneidy, "A deep learning method for vision based force prediction of a soft fin ray gripper using simulation data," *Frontiers in Robotics and AI*, vol. 8, 2021.

[10] Y. Zhu, M. Hao, X. Zhu, Q. Bateux, A. Wong, and A. M. Dollar, "Forces for free: Vision-based contact force estimation with a compliant hand," *Science Robotics*, vol. 10, no. 103, p. eadq5046, 2025.

[11] A. N. Reddy, N. Maheshwari, D. K. Sahu, and G. K. Ananthasuresh, "Miniature compliant grippers with vision-based force sensing," *IEEE Transactions on Robotics*, vol. 26, no. 5, pp. 867–877, 2010.

[12] D.-K. Ko, K.-W. Lee, D. H. Lee, and S.-C. Lim, "Vision-based interaction force estimation for robot grip motion without tactile/force sensor," *Expert Systems with Applications*, vol. 211, p. 118441, 2023.

[13] P. Intelligence, K. Black, N. Brown, J. Darpinian, K. Dhabalia, D. Driess *et al.*, "$\pi_{0.5}$: A Vision-Language-Action Model with Open-World Generalization," *arXiv preprint arXiv:2504.16054*, 2025.

[14] Z. Zhang, A. Petit, J. Dequidt, and C. Duriez, "Calibration and external force sensing for soft robots using an rgb-d camera," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2356–2363, 2019.

[15] E. Coevoet, T. Morales-Bieze, F. Largilliere, Z. Zhang, M. Thieffry, M. Sanz-Lopez, B. Carrez, D. Marchal, O. Goury, J. Dequidt, and C. Duriez, "Software toolkit for modeling, simulation, and control of soft robots," *Advanced Robotics*, vol. 31, no. 22, pp. 1208–1224, 2017.

[16] T. Nath, A. Mathis, A. C. Chen, A. Patel, M. Bethge, and M. W. Mathis, "Using deeplabcut for 3d markerless pose estimation across species and behaviors," *Nature Protocols*, vol. 14, no. 7, pp. 2152–2176, 2019.

[17] B. Wen, W. Yang, J. Kautz, and S. Birchfield, "Foundationpose: Unified 6d pose estimation and tracking of novel objects," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 35, 2024, Conference Proceedings, pp. 17 868–17 879.

[18] B. Xie, M. Jin, Z. Yang, J. Duan, M. Qu, and J. Li, "Research on mechanical properties and model parameters of 3d printed tpu material," *Journal of Engineering Design*, vol. 30, no. 4, pp. 419–428, 2023.

[19] S. D. Team, X. Chen, F.-J. Chu, P. Gleize, K. J. Liang, A. Sax, H. Tang, W. Wang, M. Guo, T. Hardin, X. Li, A. Lin, J. Liu, Z. Ma, A. Sagar, B. Song, X. Wang, J. Yang, B. Zhang, P. Dollár, G. Gkioxari, M. Feiszli, and J. Malik, "Sam 3d: 3dfy anything in images," 2025.

[20] T. Yoshikawa and K. Nagai, "Manipulating and grasping forces in manipulation by multifingered robot hands," *IEEE Transactions on Robotics and Automation*, vol. 7, no. 1, pp. 67–77, 1991.