MANIPULATION

Forces for free: Vision-based contact force estimation with a compliant hand

Yifan Zhu[1]*†, Mei Hao[1]†, Xupeng Zhu[2]†, Quentin Bateux[1], Alex Wong[3], Aaron M. Dollar[1]

Force-sensing capabilities are essential for robot manipulation systems. However, commonly used wrist-mounted force/torque sensors are heavy, fragile, and expensive, and tactile sensors require adding fragile circuitry to the robot fingers while only providing force information local to the contact. Here, we present a vision-based contact force estimator that serves as a more cost-effective and easier-to-implement alternative to existing force sensors by leveraging deformations of compliant hands upon contacts when compliant hands are in use. Our approach uses an estimator that visually observes a specialized compliant robot hand (available open source with easy fabrication through 3D printing) and predicts the contact force on the basis of its elastic deformation upon external forces. Because using wrist-mounted cameras to observe the gripper is common for robot manipulation systems, our method can obtain additional force information provided that the gripper is compliant. We optimized our compliant hand to minimize friction and avoid singularities in finger configurations, and we introduced memory to the estimator to combat the partial observability of the contact forces from the remaining friction and hysteresis. In addition, the estimator was made robust to background distractions and finger occlusions using vision foundation models to segment out the fingers. Although it is less accurate and slower than commercial force/torque sensors, we experimentally demonstrated the accuracy and robustness of our estimator (achieving between 0.2 newton and 0.4 newton error) and its utility during a variety of manipulation tasks using the gripper in the presence of noisy backgrounds and occlusions.

INTRODUCTION

The ability to sense force is essential for robot manipulation systems because it allows robots to understand the interaction between the robot and the external world (*1*, *2*) and apply desired contact forces (*3*–*5*). However, the community is still searching for robust and cost-effective force sensors. Commercial wrist-mounted force/torque sensors, despite being highly accurate, are expensive, heavy, and fragile upon impacts (*6*, *7*). Another option is finger-mounted tactile sensors, which require adding fragile circuitry to the robot fingers and often necessitate hardware modifications to accommodate them; even then, these sensors only provide information local to the contacts. External RGB (red, green, blue) cameras, on the other hand, are light, cheap, and reliable and do not interfere with the gripper. In this work, we aimed to make initial steps toward cost-effective and reliable force sensing for robots with a wrist-mounted RGB camera by leveraging an open-source compliant gripper whose deformation upon external contact provides cues to contact forces.

Leveraging a compliant gripper for force sensing is essential to our approach, given that visual cues for contact forces are subtle or nonexistent for traditional rigid hands manipulating rigid objects. For example, it is almost impossible to estimate the contact force of a rigid parallel-jaw gripper grasping a rigid object from an image alone given that the visual appearances remain unchanged over a large range of external contact forces. On the other hand, compliant end effectors consist of soft materials or compliant mechanisms, and once an object is grasped, they deform in response to external contact forces, behaving like springs. This deformation gives rich visual information about the magnitude and direction of those forces (*8*, *9*). Compared with traditional end effectors that are generally very stiff, compliant grippers, which can also often be underactuated (*10*, *11*), react passively to external forces and disturbances (*12*–*15*). When designed properly, they adapt to and passively accommodate large variations in object geometry and physical properties, making highly effective grasping possible with simple open-loop control and minimum sensing. Hirose and Umetani (*10*) started the concept of compliant hands consisting of multilinks and series of pulleys, which, upon closing, could adapt to object outlines passively. Since then, great progress has been made in compliant grippers to make them more capable and robust. Designed mainly for power grasping, the shape deposition manufacturing hand (*16*) was a four-fingered gripper with viscoelastic flexure joints actuated by cables and a single motor. The iRobot-Harvard-Yale hand (*17*) expanded the manipulation capability beyond power grasping by using three fingers with four actuators that could change its configuration to execute a variety of common grasping types. However, these multifingered grippers had large self-occlusions, making it challenging to estimate the contact force accurately using a single camera. One compliant gripper design that is particularly suitable for force sensing with a camera is the T42 gripper (*18*), upon which we based our design. It is a planar, cable-drive two-fingered gripper with elastic flexure joints made of polyurethane, originally developed for fingertip precision manipulation. Simple planar grippers such as this are heavily used in robotics and can grasp a broad range of common objects. Because of its simple kinematic structure, the T42 hand allows observing deformation of the fingers using a single RGB camera mounted next to the hand (Fig. 1A and Movie 1) and provides a clear view of the fingers without self-occlusion.

In this work, we aimed to exploit compliance for contact force estimation by using a wrist-mounted camera observing a compliant hand. Given that it is common to use wrist-mounted cameras in robot manipulation tasks for general perception needs (*19*–*21*), this enables a robot equipped with compliant grippers to essentially

[1]Department of Mechanical Engineering and Materials Science, Yale University, New Haven, CT, USA. [2]Khoury College of Computer Sciences, Northeastern University, Boston, MA, USA. [3]Department of Computer Science, Yale University, New Haven, CT, USA.
*Corresponding author. Email: yifan.zhu@yale.edu
†These authors contributed equally to this work.

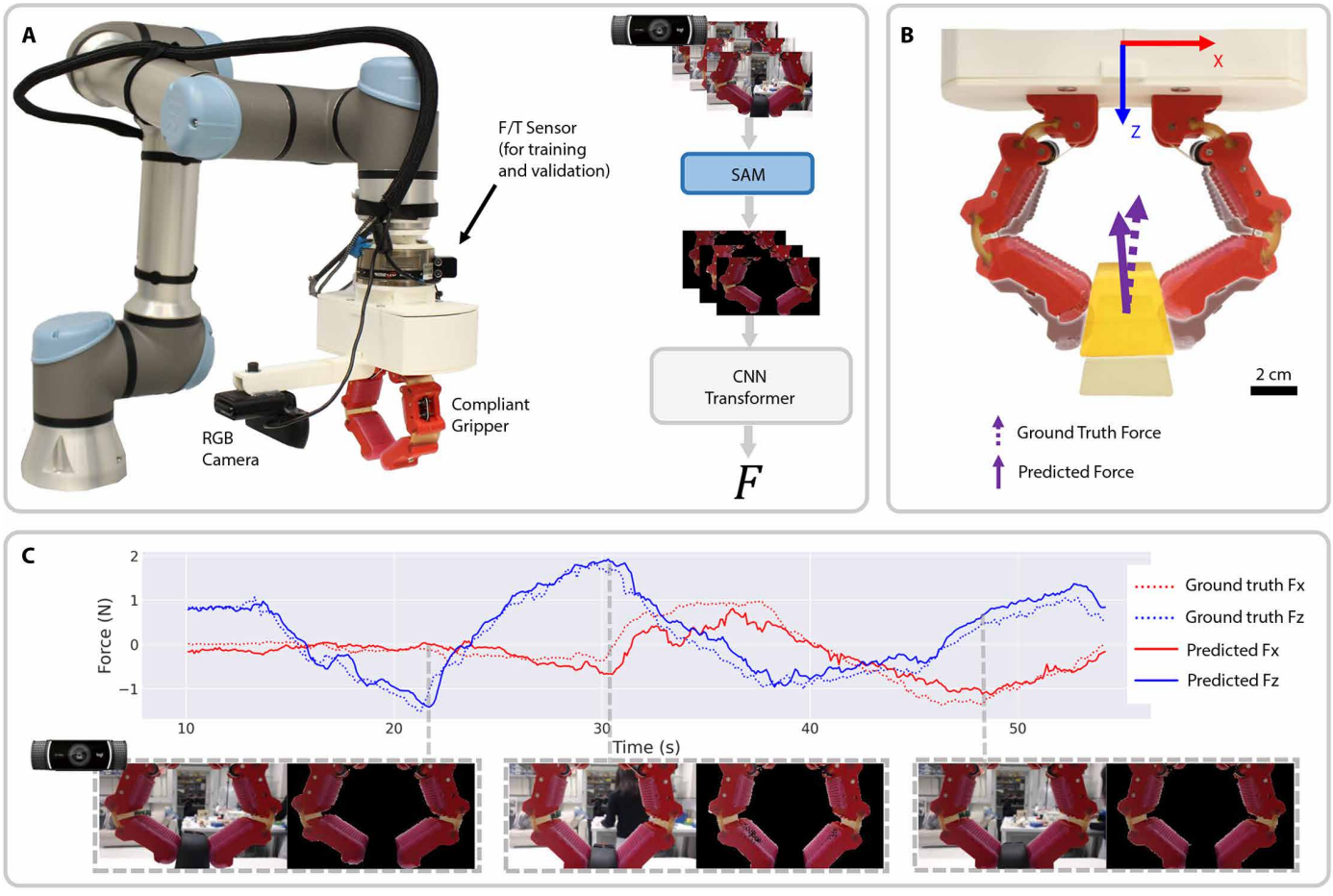

Fig. 1. Overview of our system. (**A**) Our gripper-estimator system is mounted on a UR5e robot arm. A wrist-mounted ATI F/T sensor is used only for training and validation purposes. Our estimator takes in image sequences, segments them using the SAM model, and processes them using the same CNN encoder and a transformer to predict the 2D contact forces. (**B**) The soft flexure joints of the compliant gripper act as springs and deform under external forces. The deformed finger positions can be used to estimate the total forces acting on the gripper when interacting with objects in the environment. The coordinate frame where forces will be expressed in our paper is denoted as well. The 2-cm scale bar on the bottom right provides spatial reference. (**C**) The ground-truth and predicted forces for a clamp that is grasped by the gripper while random external forces are applied to the grasped object with a human hand. Both the raw RGB images and the segmented images with SAM are shown for three time steps.

Movie 1. Motivation and overview of our system.

obtain additional contact force information "for free" after a calibration procedure. Other existing open-source robot manipulation hardware such as the universal manipulation interface (*20*) already uses fin-ray fingers that are deformable in certain directions and might also be leveraged for similar purposes, but they need to be modified to optimize their utility in force sensing. Although one might estimate forces experienced by spring-like fingers from joint and actuator encoders or other configuration sensors, cameras can provide all of the necessary information about the hand configuration when there is no occlusion, and they do not interfere with the hand. Furthermore, many compliant fingers deform in a continuum-like manner, making it nearly impossible to sense their configuration with traditional sensors.

The concept of inferring compliant robot end-effector contact forces from visual information has begun to be explored in recent studies. Elangovan *et al.* (*22*) predicted contact forces of a single adaptive finger through sensing the finger configuration changes after contact. The reconfiguration was captured by the poses of low-dimensional ArUco markers using a camera and finger motor state information. A random forest algorithm was then used to estimate force. However, this approach required the addition of more markers onto the fingers, and the markers could be corrupted completely because of visual occlusions. De Barrie *et al.* (*9*) estimated the normal contact force of a single soft fin-ray finger and used

high-dimensional camera images to capture the deformation instead of ArUco markers. An image of the gripper was first segmented and then processed using a convolutional neural network (CNN) to predict both the internal stress map of the finger and the normal contact force. This method relied on a large amount of accurate simulated data for finite element analysis simulations to train the neural network, which were challenging to obtain for cable-driven compliant grippers. Both studies evaluated force-sensing capabilities of a single finger fixed in carefully controlled setups with uniform color backgrounds and did not demonstrate the approach with an entire gripper grasping objects, contact-rich manipulation tasks, or natural noisy backgrounds. More closely related to our work are recent papers by Collins *et al.* (*23*, *24*), where camera-based force estimation was conducted in the real world for a simple elastic gripper. Each of the two fingers of the gripper featured a soft rubber fingertip supported by spring steel flexures. A CNN was also adopted to predict the contact forces given an image observation. However, the finger structure exhibited minimal deformation upon contact, providing limited visual cues for inferring contact forces. As a result, the estimators in these two works were mostly used as a binary contact sensor for various manipulation tasks because of the inaccuracies in the predictions. In addition, for a vision-based force estimator to be used in manipulation tasks in the real world under a variety of scenarios, it needs to be robust to visual distractors such as the object being grasped and the background. However, training such an estimator could require a prohibitively large amount of data. Another important challenge not explicitly addressed in these studies is that many compliant grippers cannot be characterized as perfectly elastic springs and suffer from partial observability because of friction and hysteresis (*8*). For example, the cable transmissions in the T42 gripper exhibit substantial friction, and the same finger configuration could correspond to different finger forces because of static friction. One potential approach to alleviate this challenge is to use memory (*25*–*27*) and leverage the information given by a history of finger motions to better estimate the force. In this direction, Feng *et al.* proposed a memory-based machine learning model for the contact force sensing of a single tendon-driven continuum finger that exhibited hysteresis on the basis of tendon tension readings (*8*). Although the hysteresis was mitigated with the use of memory, the estimation required additional tension sensors for the finger tendons.

Here, we demonstrate a vision-based end-effector force estimator using a compliant gripper and a low-cost wrist-mounted camera capable of completing a range of manipulation tasks that require fine modulation of contact forces. Unlike prior works that adopt existing grippers (*23*, *24*), we modified an existing open-source T42 gripper to create the forces-for-free (F3) gripper, optimized for force sensing and made publicly available through the Yale OpenHand Project website (*25*). The gripper can be manufactured easily through three-dimensional (3D) printing. To address the challenge of partial observability introduced by friction and hysteresis in the gripper, our estimator uses memory and processes a sequence of recent images capturing recent motions instead of a single image. In addition, we used the segment anything model (SAM) (*28*), a recent visual foundation model trained on web-scale data, to segment out the gripper and achieve robustness against background and object variations. Last, we optionally augmented the estimator training process with random finger occlusions to make the estimator robust to finger occlusion, which is a common situation in real-world use.

The effectiveness of our estimator was demonstrated through several experiments, including force predictions for previously unseen objects being grasped and undergoing external contacts and feedback control tasks for peg-in-hole, wiping, and calligraphy writing, all with our estimator in the feedback control loop. Across these tasks, the estimator achieved force prediction errors ranging from 0.15 to 0.38 N within a 2.5-N force range. With the optional occlusion augmentation, the estimator was robust to moderate occlusion of the fingers with a slightly worse prediction accuracy. In contrast with the most-related recent works proposed by Collins *et al.* (*23*, *24*) that do not leverage memory, an optimized gripper for force estimation, or visual foundation models, our method achieved much better force prediction accuracies [compared with about 1.6 N of root-mean-square error (*24*)]. The source of the error for our system mainly stems from the remaining friction and hysteresis present in the system, which is alleviated using memory in our estimator. Although our estimator is slower (updates at 10 Hz) and less accurate than commercial force torque sensors, we demonstrated that our estimator is sufficiently accurate, robust, and fast for several simple closed-loop manipulation tasks in the real world. Note that our method is subject to an inherent trade-off between force-sensing range and error, and a stiffer gripper is needed for tasks that require much larger force magnitudes but less force resolution. It is also possible to explore the use of nonlinear compliance that provides both fine resolution at small force magnitudes and capacity for large ranges. Although the accuracy reported here is for the F3 gripper, we expect similar relative accuracy for grippers of different sizes and compliances. Last, our code and the hardware design of the F3 gripper are available online to enable others to easily implement our approach and encourage further research in this area.

RESULTS

Our proposed force estimator takes in a sequence of 20 recent images from an RGB camera (Logitech C920) pointed at the fingers, segments out visual distractors using a fine-tuned SAM visual foundation model, processes the images with a shared CNN encoder, and passes the sequence of latent image features through a Transformer network to predict the total 2D forces (F_x and F_z relative to the gripper frame). The time interval between each image frame in the memory was 1 s in these experiments. We assumed that the gripper had already grasped an object with the fingertips when the estimator predicted the external contact force acted on the grasped object. The force estimator was trained on a dataset of about 10.5 hours (375,000 frames) of automatically collected data of various grasped objects interacting with the external environment while recording ground-truth forces with an industrial-grade ATI Gamma force-torque (F/T) sensor.

Data collection

A high-quality dataset that covered a diverse set of finger positions was essential for the performance of the force estimator. There were a few important considerations for the data collection strategy. First, the contact force was invariant to the geometry of the grasped object within the gripper, and we aimed to train the estimator on differently sized objects. Second, diversity was needed for the changes in the contact, such as increasing and decreasing contact forces at different speeds, to generalize well to various manipulation tasks. To this end, we collected data through two setups with objects of varying

sizes, shown in Fig. 2 (A and B). In the random setup, the gripper grabbed a peg with slots constrained by a single rod and translated in the 2D plane between random commanded positions with varying translating speeds. In this setup, the contact force between the peg and the rod varied substantially, and it covered the case where the grasped object made and broke contact, which was common for contact-rich manipulation tasks. In the wiping setup, the gripper grabbed a peg and translated back and forth on a slope with varying speeds and downward force. Here, the grasped object made persistent contact with the environment. The slope provided a diverse combination of the F_x and F_z forces. Forces from the F/T sensor and webcam images were recorded at 10 Hz. Data on both setups were collected on five pegs whose widths ranged from 25 to 35 mm with a 2.5-mm interval (a discussion on this choice is provided in the Discussion). The distribution of the collected forces in the entire dataset is shown in Fig. 2C.

Static force prediction for novel objects

We first demonstrated the estimator's prediction accuracy on a test dataset of four novel objects, shown in Fig. 2D. The widths of the grasped parts on these objects fell between 25 and 35 mm. In this test, the gripper remained static, whereas the grasped object was perturbed by a human hand, in both a clean controlled environment with white backdrops (white) and a messy dynamic lab environment with humans walking by (lab). Two examples of the force predictions made by our estimator on the yellow pear object in the white

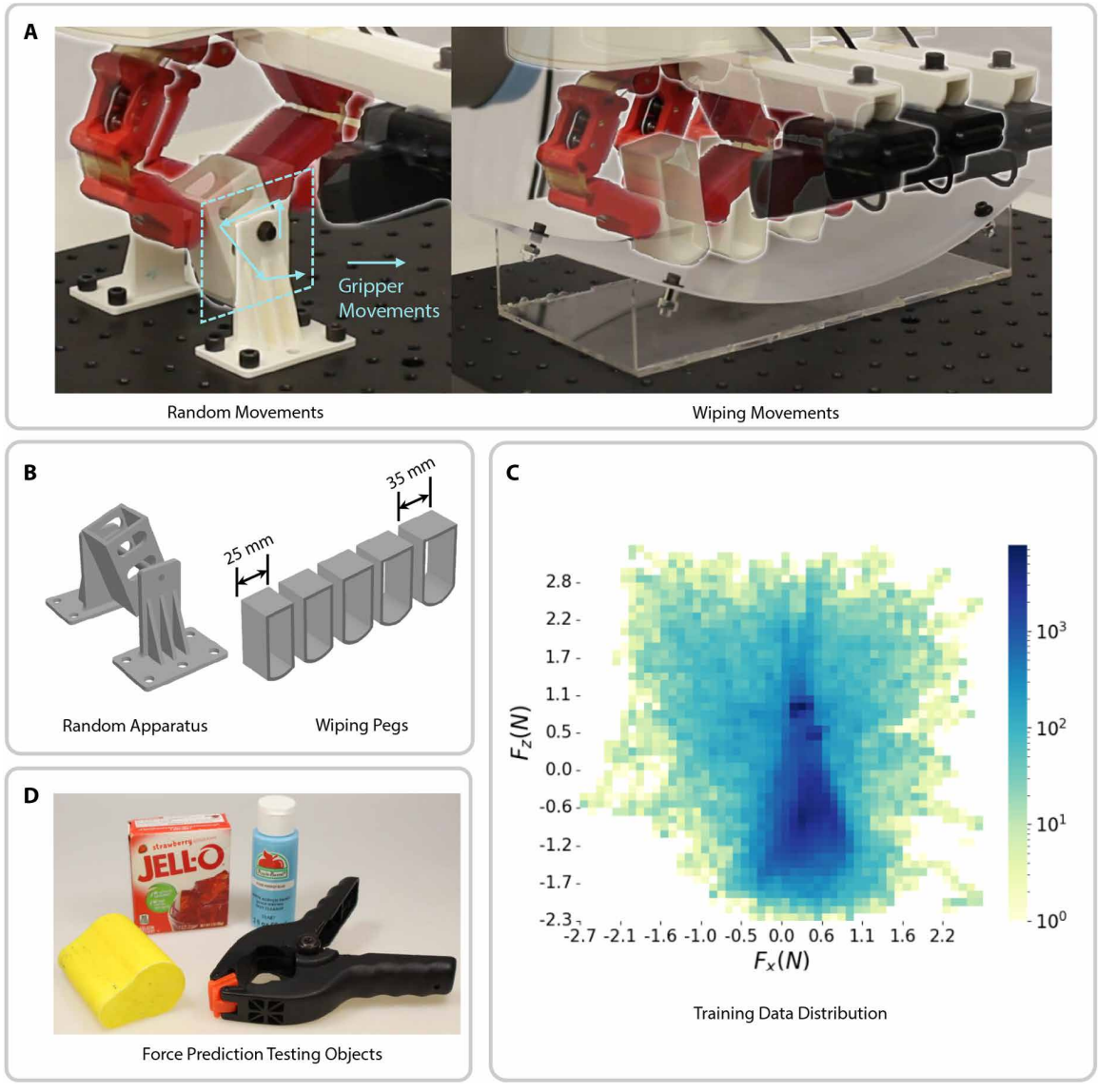

Fig. 2. Data collection and testing setups. (**A**) The random and wiping setups for data collection. For random, the gripper grabs pegs of varying sizes with slots in them that are constrained to a rod and moves between random points in a 2D plane with varying speeds. In wiping, the gripper grabs pegs of varying sizes and moves back and forth on a slope while maintaining varying downward forces with varying translation speeds. (**B**) The setup used to constrain the pegs for random and the different sizes of the pegs for wiping (random also has pegs of the same sizes). (**C**) The training data distribution where the color indicates the number of force data points that fall into each grid. (**D**) Novel objects used for force prediction testing: Jello box, paint bottle, clamp, and yellow pear.

environment and the clamp object in the lab environment are visualized in Figs. 1C and 3A, respectively. The SAM model could robustly segment out the fingers in novel backgrounds, allowing the approach to be effective even in cluttered novel backgrounds, while also reducing the amount of training data required. As reported in Table 1, our estimator was able to make predictions with a force error of about 0.35 N when the force magnitude range was 0 to 2.5 N. In addition, the estimator accuracy was mostly independent of the

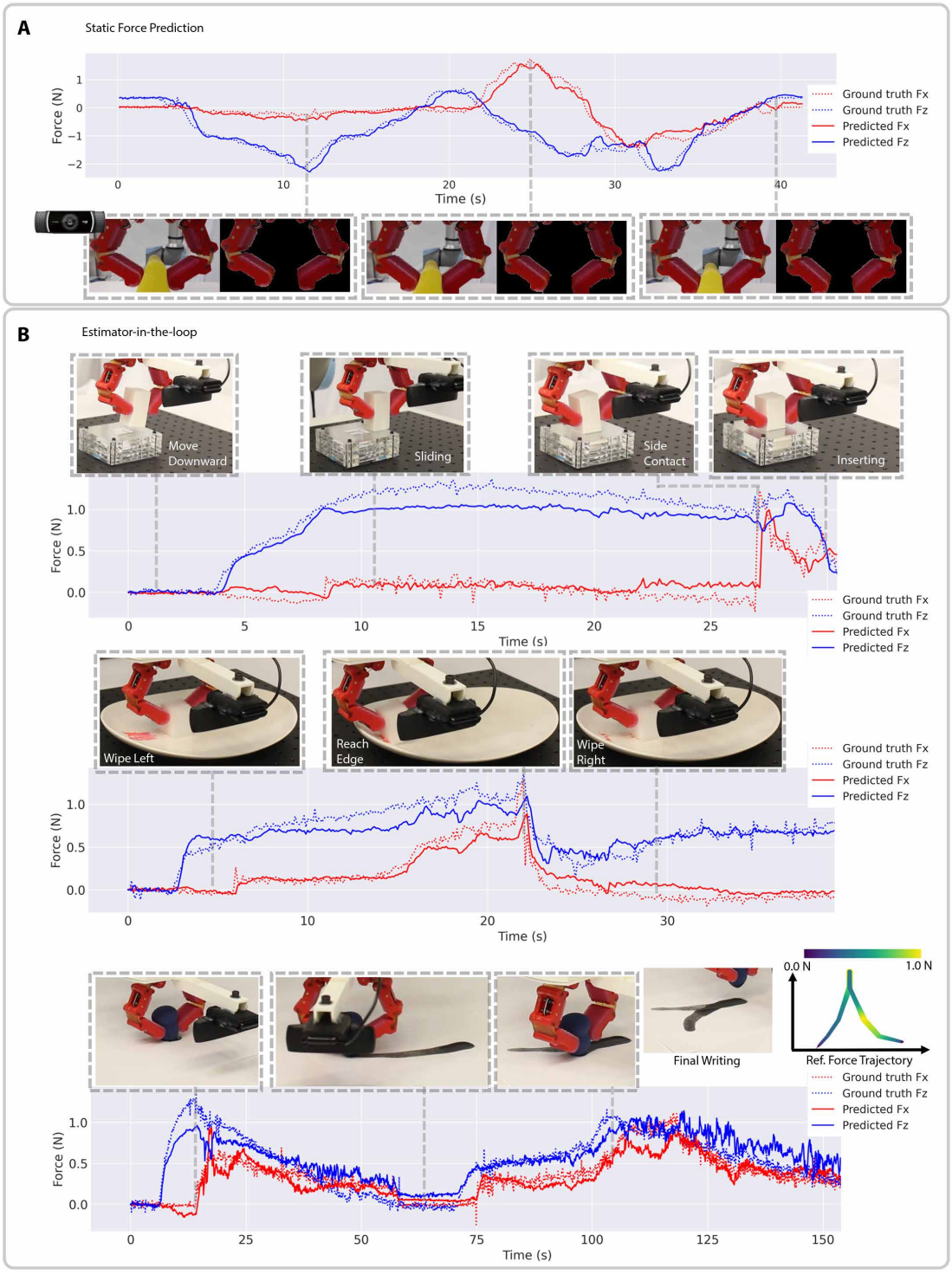

Fig. 3. Force prediction made by our estimator in different tasks. (**A**) Static force prediction. (**B**) Tasks that use the estimator in the loop for control: peg-in-hole, wiping, and calligraphy writing. For the calligraphy task, the desired character with the target forces along the trajectories is specified in the figure in the top right corner. Note that the character consists of two separate strokes.

Table 1. Average force errors for different tasks. Average errors for different tasks accumulated over models trained with three random seeds are reported. Average errors for different ground-truth force magnitude bins (0 to1 N, 1 to 2 N, and >2 N) and the overall average force errors for models trained with occlusion augmentation are also reported. NA, not applicable.

Task name	Avg. error (N)	0 to 1 N	1 to 2 N	>2 N	Error w/ occlusion aug.
Static force prediction					
White	0.342	0.296	0.366	0.350	0.433
Lab	0.371	0.308	0.429	0.404	0.421
Estimator in-the-loop					
Peg-in-hole	0.145	0.104	0.248	NA	NA
Wiping	0.150	0.138	0.204	NA	NA
Writing	0.171	0.155	0.226	NA	NA

force magnitude, and the error only increased slightly as the ground-truth force magnitude increased. We think that the remaining errors were mainly due to the presence of friction and hysteresis in the gripper, which the use of memory improved but did not eliminate.

Estimator-in-the-loop for dynamic manipulation tasks

We also evaluated the force estimator for realistic contact-rich manipulation tasks with the proposed estimator in feedback control. We completed a plate-wiping task, a 2D peg-in-hole task, and a Chinese calligraphy task using our estimator as force feedback, shown in Fig. 3B. In the plate wiping task, we used a hybrid force-motion control scheme, where a proportional-integral-derivative (PID) controller modulated the target position given to the robot end effector in the vertical direction on the basis of the desired force of 1 N in the vertical direction. The velocity in the horizontal direction was tracked by the end effector directly. To ensure that there were large force variations to evaluate the estimator, instead of a steady force, we intentionally limited the maximum velocity in the vertical direction to make the controller lag behind the force target on the curved plate. The peg-in-hole task followed the control strategy proposed by Morgan *et al.* (*29*), where a hybrid force-motion controller followed different force targets in stages to robustly complete contact-rich tasks in noisy environments with no vision. The robot first moved straight downward while seeking a force target of 1 N. Once reached, it moved horizontally to search for the hole and aimed to reach a force target of 0.8 N in the horizontal direction. This was because once the hole was reached, the downward pressure on the peg would cause the peg to tilt (because of finger compliance) and hit the edge of the hole with its corner. Next, the controller pushed down the peg while aiming to maintain zero force in the horizontal direction (clearance with the hole to avoid jamming) to complete the insertion. For the calligraphy task, the gripper grabbed a foam makeup "brush" and wrote with varying force to control the thickness of the stroke. The hybrid force-motion controller commanded the end effector to follow the position on the *xy* plane and yaw angle, specified by the reference trajectory generated from the shape of the Chinese character. The controller then tracked the force target in the vertical direction by setting the vertical position commands with a PID controller.

For each of these tasks, three models trained with three random seeds were executed once each, and the averages across the three trials were reported. As shown in Table 1, our estimator was able to track force accurately with an average force of 0.145 N for the peg-in-hole task, 0.150 N for the wiping task, and 0.171 N for the calligraphy task.

Robustness to occlusions

Gripper occlusion is common during manipulation tasks in the real world, and deploying the estimator robustly requires learning a visual representation that is invariant to occlusions, which we can achieve via data augmentation during training. As detailed in the Materials and Methods, we augmented data during training to improve robustness to occlusions by randomly cropping out a convex polygon in the images during training. As shown in Fig. 4, we occluded the links and flexure joints of the finger with strips that were 1, 2, and 4 cm in width while the gripper stayed still. Our estimator was robust to partial occlusions with the estimations remaining stable. With increasingly large occlusions that occluded the entire proximal finger link, the estimator performance naturally became worse. We noted that the use of such a data augmentation technique would slightly lower the accuracy of the estimator in the ideal case of no occlusion, as shown in the last column in Table 1. This was because the augmentation forced the model to pay more attention to parts of the image such as the proximal link of the gripper, which had a lower signal-to-noise ratio compared with the fingertips.

Ablation studies

We investigated the effect of memory length and different mechanisms for processing memory in our proposed force estimator. We evaluated our estimator trained with different memory lengths on the static force prediction task. In addition, we compared using attention in a transformer architecture to handle the memory versus a multilayer perceptron (MLP). With MLP, each image was first processed by a CNN, after which the extracted features were concatenated and passed through the MLP, which contained three layers. Shown in Fig. 5, both architectures were quite similar in their performance, with attention being slightly better and achieving the lowest average error with a memory of 20 frames. Although the attention mechanism was powerful at handling complex relationships between tokens in long sequences, the relationship between different frames in the fixed-length sequence was relatively static, and the transformer did not show advantages over a simpler MLP architecture. In addition, the benefit of using memory was demonstrated by lower errors with a larger memory. However, using memory beyond 20 frames did not further reduce the error given that image frames beyond 20 were no longer informative of the current contact force.

DISCUSSION

The results demonstrate that force sensing useful for various manipulation tasks can be achieved with only an inexpensive webcam and our proposed force estimator along with a compliant gripper. Although performance is lower than what can be achieved with top-of-the-line F/T sensors, they are sufficiently accurate to be used in a number of real-world tasks. The advantages of our approach include the extremely low cost (can be "free" if a camera observing the hand is already being used and the hand is compliant), lighter weight (F/T sensors can weigh hundreds of grams), and increased mechanical

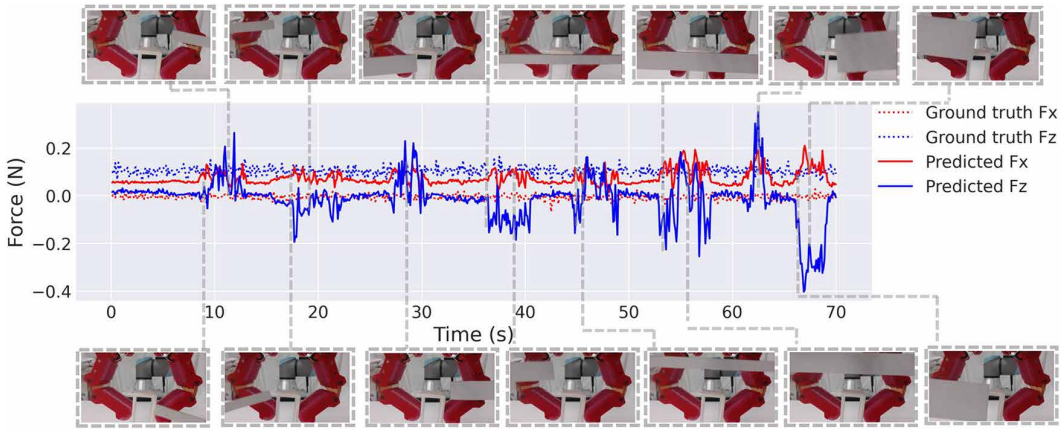

Fig. 4. Our estimator is robust to finger occlusions. Strips with widths of 1, 2, and 4 cm are placed in front of the fingers with different amounts of occlusion, whereas the gripper and object remain still. The predicted forces remain stable with partial occlusion of the distal links.

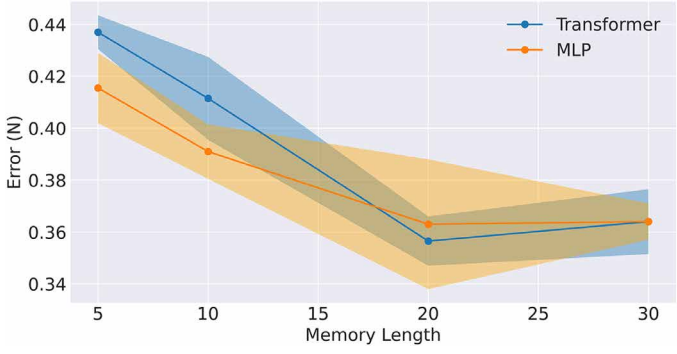

Fig. 5. Ablation studies for memory length and neural network architecture. Testing error on the force prediction tasks (both lab and white backgrounds) with different lengths of memory for the MLP and transformer architectures. The mean and SD of the errors (sample size $N = 3$) over three models trained with different random seeds are plotted. The mean is shown as solid dots, and ± SD is represented as the shaded region.

robustness (traditional tactile sensors are fragile). For the estimator, the results suggest that the combination of a diverse calibration dataset and a machine learning system that is designed to handle friction, hysteresis, visual distractors, and occlusions leads to an accurate force estimator based only on RGB images. However, compared with commercial F/T sensors, the proposed estimator runs at a slower frequency because of both the camera's sampling rate and the bottleneck on computation. This indicates that our estimator is best used on tasks where very quick reaction to force changes (such as collision detection) is not required. However, for many manipulation tasks where the motions are not fast, we believe our estimator has the potential to replace fragile and expensive F/T sensors. Another benefit of our estimator is that the measurements do not drift or creep, whereas this is a known issue for commercial F/T sensors (*30*). In addition, our method is subject to an inherent trade-off between force-sensing range and error, depending on the gripper stiffness. For tasks that involve much larger contact force magnitudes, a stiffer gripper might be required, at the cost of worse force resolution.

Perhaps the greatest limitation to the current proof-of-concept study is that we limited our focus to planar/2D forces. Although there are no fundamental hurdles to extending to 3D, including for multifinger grippers, substantial occlusions and data efficiency will make the problem more challenging. However, the issue of occlusions can be alleviated by adding additional cameras. In addition, instead of predicting the total force on the end effector, one potential approach is to predict forces on each finger to improve data efficiency when scaling up in terms of number of fingers. We also assume that there is only one contact between each fingertip and the grasped object, which can be violated in practice during manipulation. The same finger position could correspond to different finger forces when the number of contacts is different. In this case, this would require the estimation of the contact locations, potentially under heavy occlusion. We aim to explore the sensing of contact locations from vision in the future, which has been explored for contacts between rigid bodies using vision only (*31*).

Another limitation is the relatively small range of object sizes used in this study. Although the gripper can grasp a much larger set of objects and sizes, we wanted to remove variations in the actuation of the gripper from the sets of variables we considered. A direct extension of the current study will look at adding the actuation of the gripper as another dimension to consider. This will require a zeroing procedure (for example, closing the gripper to mechanical limits) to ensure that actuator positions map appropriately to finger positions as well as a procedure to vary actuation amount (for example, motor angles) and object sizes during training. We expect that this will greatly increase the amount of data required and, to a certain degree, reduce the overall accuracy of our approach. In all of the experiments, we fixed the roll and pitch of the gripper to be zero for simplicity. However, our method can accommodate the case when the gripper tilts by compensating for the gravity of the fingers, whereas mature techniques can calibrate the gravity well (*32*). A potential future direction is to leverage the visual cues about contact force implicitly in robot manipulation systems given recent trends in end-to-end vision-based manipulation systems. For example, using the visual cues of contact forces given by a compliant gripper as an extra observation modality has the potential to enhance the capabilities of robots without wrist-mounted F/T sensors.

Overall, our findings highlight the potential of leveraging low-cost, reliable cameras for force estimation in compliant grippers during contact-rich manipulation tasks. This opens the door for the

creation of more advanced vision-based force estimators for grippers and general manipulation platforms. Given the increasing use of low-cost, teleoperated systems for collecting human demonstration data, our system also offers a promising solution of empowering these systems with much-needed force-sensing capabilities, which they currently lack.

MATERIALS AND METHODS

Methodological overview

Our vision-based end-effector force estimator used a compliant and underactuated gripper and a low-cost wrist-mounted camera. The proposed F3 gripper was based on the T42 gripper (*18*) and offered two key improvements over the T42 gripper. First, the kinematic structure of each finger was optimized to exhibit larger finger deformations upon external contact for a better signal-to-noise ratio. In addition, friction in the tendon transmission was greatly reduced to mitigate the partial observability. These modifications greatly improved the accuracy of our method without introducing noticeable performance degradation for manipulation. We did not modify the stiffness of the flexure joints and merely improved the kinematic structure to obtain large finger deformation upon external contact force. To address the challenge of partial observability introduced by friction and hysteresis in the gripper, our estimator used the recent motion history of the fingers and substantially improved prediction accuracy over single images. In addition, we fine-tuned the recently proposed SAM (*28*), a visual foundation model trained on web-scale data, on a small custom dataset of segmentation masks to segment out the gripper and achieve robustness to background and object variations. For enhanced robustness to occlusions, we optionally augmented training with random finger occlusions, which improved robustness to real-world occlusions at the cost of a slight reduction in prediction accuracy. Although our method was applied to a simple two-fingered gripper for a clear view of the fingers, our method could potentially be extended to multifingered hands that suffer from self-occlusions by leveraging multiple cameras.

Finger optimization for better force estimation

The gripper used for this project was based on the Yale OpenHand model T42 (*18*). To achieve better force predictions based on the deformation of the hand, two major changes were introduced: First, modifying the distal finger link length and angle to maximize the deformation of the finger upon external forces given fixed stiffness value of the finger, and second, changing the routing of the tendon and structure of the fingers to minimize the friction that the tendon experienced when actuating the finger.

The optimization parameters of the finger kinematic structure included the link length ratio of the distal and proximal links and the distal link angle at rest. This optimization relied on a kinematic manipulability measure that determined how much the finger deformed upon external force.

Treating each of the fingers in T42 as a two-link arm with torsion springs, then the external force $\mathbf{F} \in \mathbb{R}^2$ applied at the fingertip and the joint torques $\boldsymbol{\tau} \in \mathbb{R}^2$ satisfies

$$\mathbf{J}_{\boldsymbol{\theta}}{}^{\mathrm{T}}\mathbf{F} = \boldsymbol{\tau}$$

Here, $\mathbf{J}_{\boldsymbol{\theta}} \in \mathbb{R}^{2\times 2}$ is the kinematic Jacobian at joint positions $\boldsymbol{\theta} \in \mathbb{R}^2$. Using Hook's law for springs, we obtain

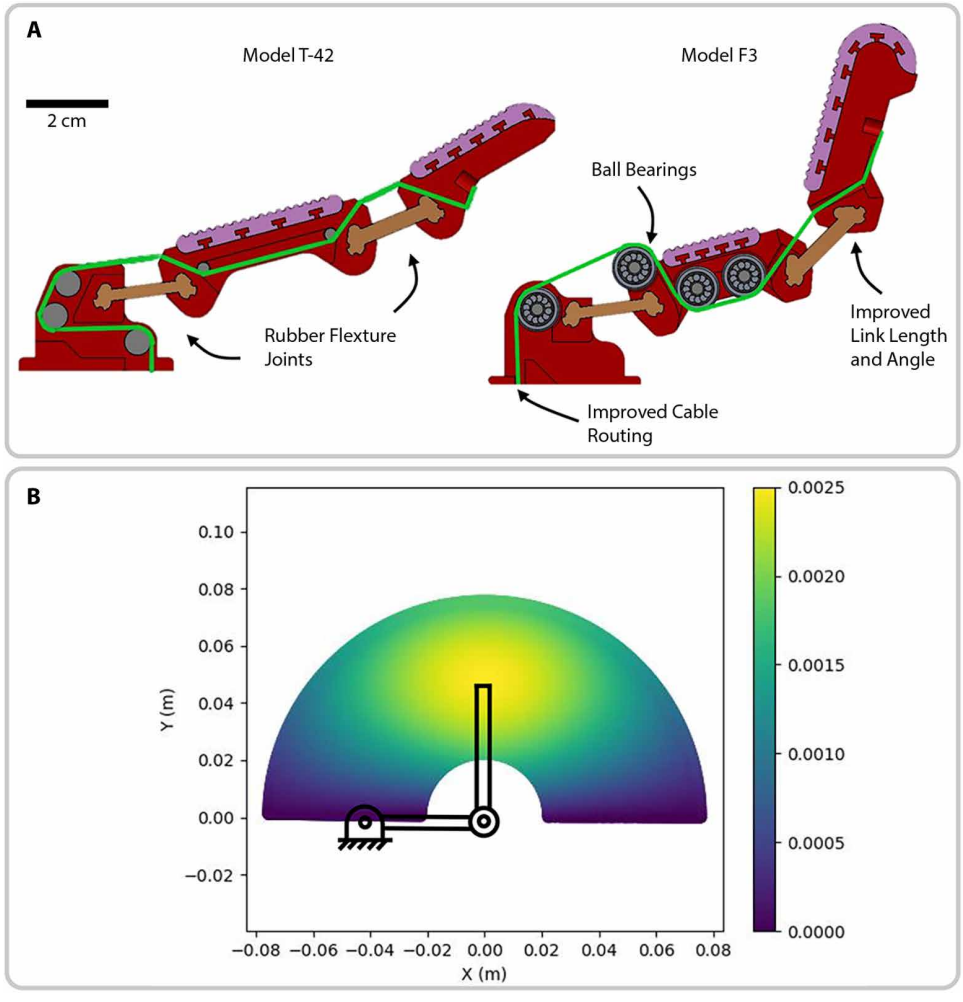

Fig. 6. F3 gripper finger design and manipulability optimization. (**A**) Our modified F3 finger compared with the original model T42 (*39*) finger in the rest configuration. Notice that when the finger is closed upon grasping objects, the distal joint angle will decrease to about 90°. A 2-cm scale bar is included to indicate size. (**B**) Visualization of the manipulability of the closed finger. We fix the total length of the arm to be 0.1 m and the proximal joint angle to be 0°. We iterate through a grid of proximal/distal link ratios and distal joint values, and the manipulability at each corresponding end-effector location is plotted. The best manipulability is achieved when the links are equal in length and the distal link angle is 90° (the black linkage plotted).

$$\mathbf{J_\theta}^{\mathrm{T}}\mathbf{F} = \mathbf{K\theta}$$

where $\mathbf{K} \in \mathbb{R}^{2x2}$ is the diagonal torsional spring constant matrix. Therefore, the relationship between fingertip force and the finger joint positions becomes

$$\mathbf{F} = \left(\mathbf{J_\theta}^{\mathrm{T}}\right)^{-1}\mathbf{K\theta}$$

Near a singular configuration (for example, when the finger is fully extended), a large change in external force **F** in the direction of poor manipulability results in essentially zero change in joint positions, which makes the identification of the external force from finger positions difficult. Therefore, we optimized the finger link length and angle by maximizing a manipulability measurement on the basis of the kinematic Jacobian (*33*)

$$\omega = \sqrt{\det\left(\mathbf{J_\theta}^{\mathrm{T}}\mathbf{J_\theta}\right)}$$

We fixed the total length of distal and proximal links of the finger at 0.1 m to maintain the overall form factor of the T42 gripper. In addition, given that the joint position of the proximal link does not affect the manipulability measure ω, we ran a grid search over possible link lengths and distal link joint angle. Shown in Fig. 6B, the best manipulability was achieved at equal lengths of the two links and 90° for the distal link joint angle. We therefore modified the fingers of the T42 accordingly to have distal and proximal links equal in length, shown in Fig. 6A. In addition, the resting angle of the proximal link angle was designed such that upon pinch-grasping objects, the proximal joint angle was about 90° (see Fig. 1 and Fig. 3 for reference). In contrast, the fingers in the original T42 design are near a singularity and therefore less compliant to contact forces at the fingertips in certain directions.

Friction on the tendon would keep the fingers in the same position even when experiencing different forces and was a key contributor to the uncertainty in force predictions. Therefore, another key design goal was for the fingers to have as little tendon friction as possible. The reduction in friction was achieved by first changing all metal pins for cable routing into low-friction ball bearings. Then, the routing of the tendon was optimized such that the tendon angle change when in contact with the bearings was minimized. This reduced the load on the ball bearings given the same tendon tension, which in turn lowered the friction in the bearings. As shown in Fig. 6A, four small ball bearings mounted on metal pins supported all major contacts between the tendon and the finger body. The routing of the tendon was also improved at the base to reduce sharp turns. The resulting design showed a major reduction in tendon friction compared with the original T42 model (now about 0.6 N versus 4.0 N).

Force estimator

Given a sequence of h RGB images $\mathrm{I}_{t,h} = \{\mathrm{I}_{t-h+1}, \cdots, \mathrm{I}_t\}$ taken by a wrist-mounted camera of an underactuated hand at time t, the goal was to predict as close as possible to the ground truth total 2D force $\mathbf{F} = \left[F_x, F_z\right]^{\mathrm{T}} \in \mathbb{R}^2$ acting on the gripper in the gripper frame at t. Shown in Fig. 7, our neural network architecture used a shared ResNet (*34*) to encode each RGB image of size 160 pixels by 90 pixels in the memory into a latent vector of 256 dimensions, after which the sequence of image feature vectors was fed into a transformer (*35*) to predict a sequence of forces, where all but the last was discarded. The neural network then predicted the force at time t as $\widehat{\mathbf{F}}_t = g_w\left(\mathrm{I}_{t,h}\right)$

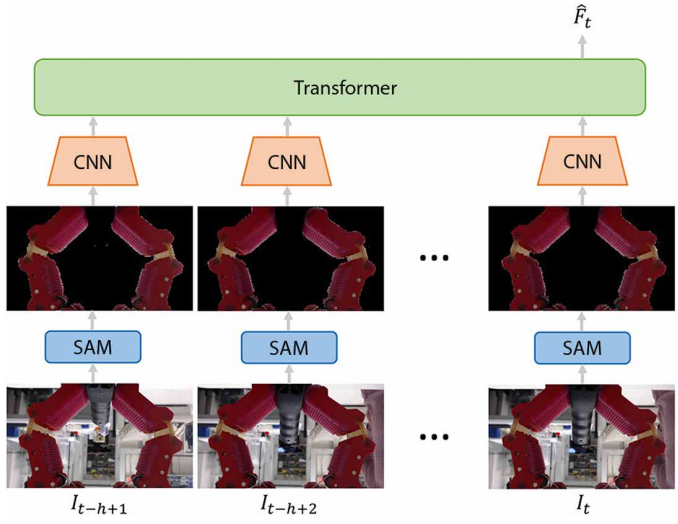

Fig. 7. Overview of the estimator neural network architecture. A sequence of images is first segmented by SAM to remove the visual distractors, after which the features are extracted by a shared CNN feature extractor. The sequence of feature vectors is then fed through a transformer, and the last output is kept as the force prediction.

where w was the neural network weights. We then used the Euclidean norm of the force error as the loss function to optimize the neural network

$$\mathcal{L} = \frac{1}{n}\sum_{i=1}^{n} \left\| g_w\left(\mathrm{I}_{i,h}\right) - \mathbf{F}_i \right\|^2$$

Training was then performed with the AdamW (*36*) optimizer using minibatches. Gradient clipping with a norm of one was applied.

One notable challenge of applying the force estimator in the real world is ensuring robustness to visual distractors from the objects being manipulated and the image background. We leveraged vision foundation models trained on web-scale image data to segment out the fingers, essentially removing all visual distractors. Specifically, we used the SAM with the ViT-B backbone and a fixed bounding-box prompt. Directly applying the model in a zero-shot fashion led to poor results. Therefore, we fine-tuned only the model decoder with a small custom dataset with labeled segmentation masks of the gripper fingers and a bounding-box prompt. We labeled a total of 15 images of the finger in various positions, augmented the background with random background images from the Massachusetts Institute of Technology indoor dataset (*37*), and added random object images from the Yale-Carnegie Mellon University–Berkeley dataset (*38*) in the foreground to get a total of 915 images. The fine-tuned SAM could robustly segment out the gripper, even for objects and backgrounds that had colors similar to that of the gripper.

To make the estimator robust to occlusion, during training, we optionally further occluded the segmented images with random convex polygons with four or five vertices, 10 to 80 pixels in height, and 80 to 160 pixels in width with a chance of 50%. The same polygon was applied to all images in the input sequence. As discussed in the Results, this augmentation led to a trade-off between the robustness and accuracy of the estimator in the ideal occlusion-free case, and we treat it as optional.

Implementation

Although the camera ran at 10 Hz throughout the paper, inputting the full sequence of recent images into the estimator was unnecessary. As a result, our estimator kept a buffer of all of the images in the past 20 s, uniformly downsampled it (1 Hz), and fed this sequence of 20 images to g_w. This downsampling frequency was determined by a grid search on a small dataset of a pilot study that minimized computation without hurting performance; the effect of the history length was discussed in the ablation studies from the Results.

During online use, for efficient implementation, instead of keeping the raw images in the memory, each camera image was processed by SAM and the ResNet feature extractor, after which the image feature vector was added to a queue of 200 image feature vectors from the last 20 s. Then, the image features were downsampled and fed to the transformer for inference at each time step during control. On a standard PC with an Intel i9-13900KF CPU, 64 GB of RAM, and a GeForce RTX 4090 GPU, the estimator took 0.068 s to extract the feature vector from the most recent image, add the feature vector to the queue, and predict the force.

Statistical analysis

The plot in Fig. 5 illustrates the mean and SD of force prediction errors for two different neural network architectures with different memory lengths (sample size $N = 3$ for data points). The mean is shown with solid circles, and the shaded regions represent $\pm$ SD.

Supplementary Materials

The PDF file includes:

Legend for movie S1

Other Supplementary Material for this manuscript includes the following:

Movie S1

REFERENCES AND NOTES

1. J. Bimbo, A. S. Morgan, A. M. Dollar, Force-based simultaneous mapping and object reconstruction for robotic manipulation. *IEEE Robot. Autom. Lett.* **7**, 4749–4756 (2022).
2. N. Doshi, O. Taylor, A. Rodriguez, "Manipulation of unknown objects via contact configuration regulation" in *Proceedings of IEEE International Conference on Robotics and Automation* (IEEE, 2022), pp. 2693–2699.
3. M. Toussaint, J.-S. Ha, D. Driess, Describing physics for physical reasoning: Force-based sequential manipulation planning. *IEEE Robot. Autom. Lett.* **5**, 6209–6216 (2020).
4. N. Chavan-Dafle, R. Holladay, A. Rodriguez, Planar in-hand manipulation via motion cones. *Int. J. Robot. Res.* **39**, 163–182 (2019).
5. S. Kim, D. K. Jha, D. Romeres, P. Patre, A. Rodriguez, "Simultaneous tactile estimation and control of extrinsic contact" in *Proceedings of IEEE International Conference on Robotics and Automation* (IEEE, 2023), pp. 12563–12569.
6. R. Ouyang, R. Howe, "Low-cost fiducial-based 6-axis force-torque sensor" in *IEEE International Conference on Robotics and Automation* (IEEE, 2020), pp. 1653–1659.
7. M. Rakotondrabe, I. A. Ivan, S. Khadraoui, P. Lutz, N. Chaillet, Simultaneous displacement/force self-sensing in piezoelectric actuators and applications to robust control. *IEEE/ASME Trans. Mechatron.* **20**, 519–531 (2014).
8. F. Feng, W. Hong, L. Xie, A learning-based tip contact force estimation method for tendon-driven continuum manipulator. *Sci. Rep.* **11**, 17482 (2021).
9. D. De Barrie, M. Pandya, H. Pandya, M. Hanheide, K. Elgeneidy, A deep learning method for vision based force prediction of a soft fin ray gripper using simulation data. *Front. Robot. AI* **8**, 631371 (2021).
10. S. Hirose, Y. Umetani, The development of soft gripper for the versatile robot hand. *Mech. Mach. Theory* **13**, 351–359 (1978).
11. J. Shintake, V. Cacucciolo, D. Floreano, H. Shea, Soft robotic grippers. *Adv. Mater.* **30**, 1707035 (2018).
12. A. M. Dollar, R. D. Howe, A robust compliant grasper via shape deposition manufacturing. *IEEE/ASME Trans. Mechatron.* **11**, 154–161 (2006).
13. W. Crooks, G. Vukasin, M. O'Sullivan, W. Messner, C. Rogers, Fin Ray® effect inspired soft robotic gripper: From the RoboSoft grand challenge toward optimization. *Front. Robot. AI* **3**, 00070 (2016).
14. F. Ilievski, A. D. Mazzeo, R. F. Shepherd, X. Chen, G. M. Whitesides, Soft robotics for chemists. *Angew. Chem. Int. Ed. Engl.* **50**, 1890–1895 (2011).
15. S. Puhlmann, J. Harris, O. Brock, RBO hand 3: A platform for soft dexterous manipulation. *IEEE Trans. Robot.* **38**, 3434–3449 (2022).
16. A. M. Dollar, R. D. Howe, "The SDM hand: A highly adaptive compliant grasper for unstructured environments" in *Experimental Robotics. Springer Tracts in Advanced Robotics*, vol. 54, O. Khatib, V. Kumar, G. J. Pappas, Eds. (Springer, 2009), pp. 3–11.
17. L. U. Odhner, L. P. Jentoft, M. R. Claffee, N. Corson, Y. Tenzer, R. R. Ma, M. Buehler, R. Kohout, R. D. Howe, A. M. Dollar, A compliant, underactuated hand for robust manipulation. *Int. J. Robot. Res.* **33**, 736–752 (2014).
18. L. U. Odhner, R. R. Ma, A. M. Dollar, Open-loop precision grasping with underactuated hands inspired by a human manipulation strategy. *IEEE Trans Autom Sci Eng* **10**, 625–633 (2013).
19. T. Z. Zhao, V. Kumar, S. Levine, C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware" in *Proceedings of Robotics: Science and Systems* (RSS Foundation, 2023).
20. C. Chi, Z. Xu, C. Pan, E. Cousineau, B. Burchfiel, S. Feng, R. Tedrake, S. Song, "Universal manipulation interface: In-the-wild robot teaching without in-the-wild robots" in *Proceedings of Robotics: Science and Systems* (RSS Foundation, 2024).
21. S. Cremer, L. Mastromoro, D. O. Popa, "On the performance of the Baxter research robot" in *Proceedings of IEEE International Symposium on Assembly and Manufacturing (ISAM)* (IEEE, 2016), pp. 106–111.
22. N. Elangovan, A. Dwivedi, L. Gerez, C.-M. Chang, M. Liarokapis, "Employing IMU and ArUco marker based tracking to decode the contact forces exerted by adaptive hands" in *Proceedings of 2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)* (IEEE, 2019), pp. 525–530.
23. J. A. Collins, P. Grady, C. C. Kemp, "Force/torque sensing for soft grippers using an external camera" in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2023), pp. 2620–2626.
24. J. A. Collins, C. Houff, P. Grady, C. C. Kemp, "Visual contact pressure estimation for grippers in the wild" in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (IEEE, 2023), pp. 10947–10954.
25. H. H. Nguyen, A. Baisero, D. Klee, D. Wang, R. Platt, C. Amato, "Equivariant reinforcement learning under partial observability" in *Proceedings of the 7th Conference on Robot Learning* (PMLR, 2023), pp. 3309–3320.
26. A. Baisero, C. Amato, Unbiased asymmetric reinforcement learning under partial observability. arXiv: 2105.11674 (2021).
27. Y. Xiao, S. Katt, A. T. Pas, S. Chen, C. Amato, "Online planning for target object search in clutter under partial observability" in *Proceedings of International Conference on Robotics and Automation (ICRA)* (IEEE, 2019), pp. 8241–8247.
28. A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo, P. Dollár, R. Girshick, Segment anything. arXiv: 2304.02643 [cs.CV] (2023).
29. A. S. Morgan, Q. Bateux, M. Hao, A. M. Dollar, "Towards generalized robot assembly through compliance-enabled contact formations" in *Proceedings of IEEE International Conference on Robotics and Automation*, (IEEE, 2023), pp. 8010–8016.
30. M. Y. Cao, S. Laws, F. R. Baena, Six-axis force/torque sensors for robotics applications: A review. *IEEE Sens. J.* **21**, 27238–27251 (2021).
31. L. Kim, Y. Li, M. Posa, D. Jayaraman, "Im2Contact: Vision-based contact localization without touch or force sensing" in *Proceedings of the 7th Conference on Robot Learning* (PMLR, 2023).
32. Q. Leboutet, J. Roux, A. Janot, J. R. Guadarrama-Olveram, G. Cheng, Inertial parameter identification in robotics: A survey. *Appl. Sci.* **11**, 4303 (2021).
33. K. M. Lynch, F. C. Park, *Modern Robotics* (Cambridge Univ. Press. 2017).
34. K. He, X. Zhang, S. Ren, J. Sun, "Deep residual learning for image recognition" in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2016), pp. 770–778.
35. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, "Attention is all you need" in *Proceedings of Advances in Neural Information Processing Systems* (ACM, 2017), pp. 6000–6010.
36. I. Loshchilov, F. Hutter, "Decoupled weight decay regularization" in *Proceedings of the International Conference on Learning Representations* (ICLR, 2019), pp. 1–8.
37. A. Quattoni, A. Torralba, "Recognizing indoor scenes" in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2009), pp. 413–420.
38. B. Calli, A. Walsman, A. Singh, S. Srinivasa, P. Abbeel, A. M. Dollar, Benchmarking in manipulation research: Using the Yale-CMU-Berkeley object and model set. *IEEE Robot Autom. Mag.* **22**, 36–52 (2015).
39. R. Ma, A. Dollar, Yale openhand project: Optimizing open-source hand designs for ease of fabrication and adoption. *IEEE Robot. Autom. Mag.* **24**, 32–40 (2017).

Acknowledgments: We thank I. Abraham for lending experiment equipment. We thank V. Patel, J. Grace, and S. Lai for suggestions on writing and figures. **Funding:** This work was supported by the Boston Dynamics AI Institute and US National Science Foundation FRR-2132823 (A.M.D.). **Author contributions:** Y.Z. created the method, developed software, created data collection scheme and collected data, designed and performed experiments, analyzed data, and wrote the paper; M.H. created the method, designed and made the hardware, developed software, created data collection scheme and collected data, designed and performed experiments, analyzed data, and wrote the paper; X.Z. created the method, developed software, designed the hardware, and wrote the paper; Q.B. developed software for the experiments and performed experiments; A.W. reviewed and developed software for the model and supervised the project; A.M.D. conceptualized the method, supervised the project, and obtained funding. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data needed to evaluate the conclusions in this study have been deposited in the database doi.org/10.5281/zenodo.15453922. The accompanying source code can be found at doi.org/10.5281/zenodo.15468193. The design files and assembly instructions for the hardware in this study are available at doi.org/10.5281/zenodo.15468204.

Submitted 17 May 2024
Accepted 28 May 2025
Published 25 June 2025
10.1126/scirobotics.adq5046

Science Robotics

Forces for free: Vision-based contact force estimation with a compliant hand

Yifan Zhu, Mei Hao, Xupeng Zhu, Quentin Bateux, Alex Wong, and Aaron M. Dollar

Sci. Robot. **10** (103), eadq5046. DOI: 10.1126/scirobotics.adq5046

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.