2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

# Efficient Tactile Sensing-based Learning from Limited Real-world Demonstrations for Dual-arm Fine Pinch-Grasp Skills

Xiaofeng Mao, Yucheng Xu, Ruoshi Wen, Mohammadreza Kasaei,
Wanming Yu, Efi Psomopoulou, Nathan F. Lepora, Zhibin Li

***Abstract*— Imitation learning for robot dexterous manipulation, especially with a real robot setup, typically requires a large number of demonstrations. In this paper, we present a data-efficient learning from demonstration framework which exploits the use of rich tactile sensing data and achieves fine bimanual pinch grasping. Specifically, we employ a convolutional autoencoder network that can effectively extract and encode high-dimensional tactile information. Further, we develop a framework that achieves efficient multi-sensor fusion for imitation learning, allowing the robot to learn contact-aware sensorimotor skills from demonstrations. The ablation studies on encoded tactile features highlighted the effectiveness of incorporating rich contact information, which enabled dexterous bimanual grasping with active contact searching. Extensive experiments demonstrated the robustness of the fine pinch grasp policy directly learned from few-shot demonstration, including grasping of the same object with different initial poses, generalizing to ten unseen new objects, robust and firm grasping against external pushes, as well as contact-aware and reactive re-grasping in case of dropping objects under very large perturbations. Furthermore, the saliency map analysis method is used to describe weight distribution across various modalities during pinch grasping, confirming the effectiveness of our framework at leveraging multimodal information. The video is available online at: https://youtu.be/BlzxGgiKfck.**

## I. INTRODUCTION

Dexterous robot manipulation has the capability to work across a range of tasks and environments. However, enabling dexterous manipulation in robots, particularly in a manner that is comparable to human capabilities, remains an unsolved challenge. Currently, numerous studies utilize visual feedback to enable robots to perform dexterous manipulation tasks such as box flipping [1], object rotating [2], reconfiguring and grasping objects from ungraspable poses [3], and door opening [4]. However, these visual-based methods have limitations, as the visual data could be influenced by occlusion and lighting variations. Consequently, it is very important to investigate how to incorporate tactile information for the enhancement of dexterous manipulation in robotic systems.

Tactile sensing plays a vital role in capturing detailed information about contact surfaces, including the distribution of contact forces and their variations during force-sensitive tasks – which is indispensable for achieving dexterous handling of lightweight objects with irregular surfaces, shapes, and deformable properties. Especially during close-range interaction between hands and objects, visual occlusion restricts the ability to perceive detailed information of the contact surfaces, during which tactile sensors become valuable for providing essential information of these unseeable surfaces. Integrating tactile sensing into motor learning of dexterous grasping can enhance the rich and precise sensing of surface contacts and interaction dynamics, provide irreplaceable and direct feedback when manipulating objects, and enable more robust and precise manipulation tasks [5], [6]. It is crucial to explore how robots can leverage this information to achieve dexterous manipulation abilities.

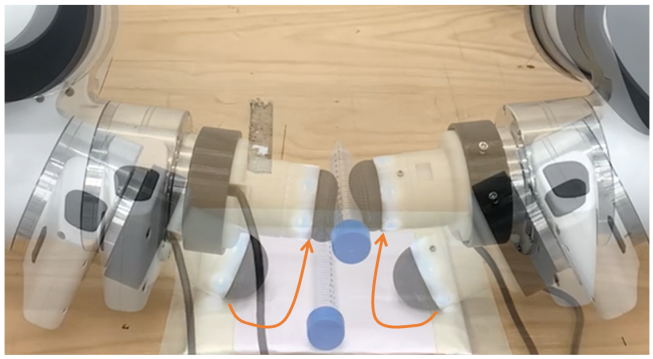


Fig. 1: Autonomous dexterous grasping with soft tactile sensors, including pre-grasp, press, roll-lift, and firm grasp.

The canonical hardware for robot manipulation incorporates Force/Torque sensors that can only measure the 6-degree-of-freedom (DoF) wrench at each end-effector. Soft optical-based tactile sensors can provide abundant and discriminative contact information by quantifying the deformation of the soft materials using a camera system [6]. Currently, several soft tactile sensors have been developed, including TacTip [7], DigiTac [8], Gelsight [9], and DIGIT [10]. However, how to use high-dimensional data from tactile sensors for robot contact-rich tasks remains open research.

The complex and non-trivial deformation of soft tactile sensors during dexterous grasping tasks presents a considerable challenge. Humans can deal with soft contacts, quickly adapt to new tasks, and produce skills of dual-arm coordination for manipulating objects. Learning from Demonstration (LfD) offers an intuitive, efficient method for acquiring human skills through synchronized tactile informa-

Xiaofeng Mao, Yucheng Xu, Mohammadreza Kasaei and Wanming Yu are with the School of Informatics, University of Edinburgh, UK. Wanming Yu is with Oxford Robotics Institute, University of Oxford, UK. Ruoshi Wen is with the Touchlab Limited, Edinburgh, U.K. Efi Psomopoulou and Nathan F. Lepora are with the Department of Engineering Mathematics, University of Bristol, UK. Zhibin Li is with the Department of Computer Science, University College London, UK. Corresponding author's email: xiaofeng.mao@ed.ac.uk

tion, encoding rich state-action mapping and enabling robots to learn human sensorimotor skills while responding to tactile and proprioceptive feedback. Additionally, the issue of errors accumulating in contact-rich tasks during Learning from Demonstration (LfD), due to the lack of direct feedback, can be addressed by incorporating rich tactile feedback in real-time. The challenge involves effectively fusing high-dimensional data with robot proprioceptive states for sample-efficient human dexterous manipulation behavior learning.

### A. Contribution

In this work, we present a framework to teach bimanual, dexterous sensorimotor skills. To handle the complex dynamics involved with deformable tactile sensors, we employ behavior cloning (BC) to learn from human teleoperated demonstrations. A convolutional autoencoder (CAE) is trained in a self-supervised manner to extract essential features from the rich tactile data, which are then integrated with the robot's proprioceptive state. This multimodal fusion enables the robot to efficiently acquire feedback-driven dexterous grasping skills through a few human demonstrations. The proposed framework is validated by pinch grasp tasks on a dual-arm setup equipped with TacTips sensors [7] and has achieved the successful retrieval of a small, cylindrical object on a table using few-shot demonstrations. Our experimental results show that the policy, learned from few-shot human demonstration data, can achieve stable grasping of unseen objects with different diameters, masses, and materials.

Furthermore, the robustness of the framework against external disturbances has been validated, with the learned policy demonstrating stable grasping under external disturbance, as well as the capacity to autonomously execute successful re-grasping in case of a large external force that pushes off the object. We applied saliency map analysis [11] and revealed how the learned policy uses different sensory modalities in a variable way throughout the dexterous pinch grasp process. This analysis demonstrates the capability and effectiveness of our proposed network to efficiently use high-dimensional data and autonomously segment the long-horizon data into several distinct fine-skills for execution according to different contact situations.

## II. Related Works

During robotic dexterous manipulation, tactile sensors can provide rich contact information which is not easily accessed via visual information, thereby playing a crucial role in enhancing the dexterous grasping capabilities [12]. Soft deformable tactile sensors can perform contact-rich interactions with the environment and manipulate delicate objects safely [13]. With optical-based tactile sensors, the orientation of the contact surface can be inferred from the tactile image, enabling stabilization of the pinch grasp by rolling the sensor on the contact surface and applying desired grasping forces [14]. The study in [15] explores utilizing tactile feedback to achieve bimanual tasks including bi-pushing, bi-reorienting, and bi-gathering, conducted within a simulated environment using reinforcement learning, and successfully achieves simulation-to-reality transfer. In contrast, our work focuses on using both deformable properties and rich surface contact information provided by the tactile sensor to achieve dexterous bimanual pinch grasping of small, fragile, and delicate objects.

One open question with high-dimensional tactile sensors is how to extract useful information from them. The works in [16] estimate 6D contact wrenches from tactile images and the estimated wrenches that can be used as feedback to the grasping controllers within the classical control theory. Deep neural networks can also be used to process tactile images. The works in [17] show that contact poses can also be detected from tactile images, which was then combined with goal-driven methods to achieve non-prehensile object stable pushing. The works in [18] introduce Autoencoder networks [19] to compress the high-dimensional tactile images into low-dimensional latent vectors which can be used for several down-stream tasks, such as object classification. In our work, we similarly employ an autoencoder to extract and encode tactile images. However, we extend its utility to achieve data-efficient learning from human demonstration by fusing the encoded tactile feature with robot proprioceptive information.

Moreover, although deformable tactile sensors facilitate area contact, potentially improving grasp stability and protecting delicate objects, the dynamics of the deformable sensor cannot be neglected. The work proposed in [13] combines 3D geometry of the tip of a deformable tactile sensor with robot proprioceptive action to learn the tactile sensor membrane dynamics and predict the deformation conditioned on robot action. Data-driven method can be used to learn the dynamics and combined with the Model Predictive Control (MPC) methods to achieve tactile servoing [20]. Insights from human intrinsic understanding may prove valuable in leveraging deformable sensors to achieve dexterous dual-arm manipulation tasks. LfD is an intuitive and effective way to learn human skills from collected demonstrations, which is very helpful for tasks requiring high-level skills, such as intricate coordination between two arms [21], [22]. By utilizing action chunking with transformer (ACT), a framework trained as a conditional Variational Autoencoder (CVAE) to learn a generative model over action sequences, the study in [23] enables the dual-arm robot to learn 6 complex tasks with only 10 minutes demonstration. Additionally, the work in [24] proposed to train a self-supervised tactile encoder and learn non-parametric policies from tactile and vision features, without integrating with the robot proprioceptive state. In contrast, our work incorporates the fusion of tactile features with robot pripriceptive state, facilitating the end-to-end learning of the manipulation tasks using BC method.

## III. Methods

### A. System Overview

Teleoperation through a physical robot is a viable approach for generating real demonstration data that can be executed on a physical system, and it was shown to be effective in performing fine dexterous grasping [25]. As shown in Fig. 2,

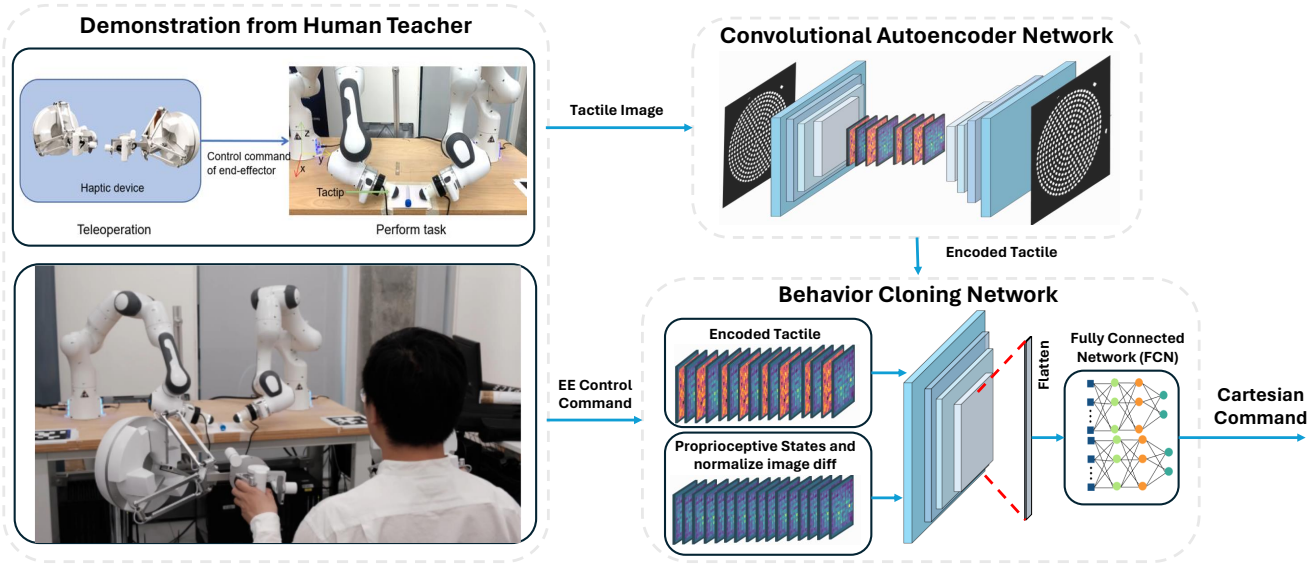



Fig. 2: Architecture detailing the teleoperation system for demonstrations and the LfD framework.

the overall architecture incorporates a teleoperation system for the collection of human demonstration data and a dual-arm setup for executing pinch grasp tasks. The teleoperation system consists of two haptic devices (Force Dimension Sigma 7) for human operators to control the dual-arm robot [26]. The dual-arm robot system includes two Franka Emika Panda arms each with a TacTip [7] installed on the end-effector of each arm. The Tactips capture contact information between end-effector and objects as 2D tactile images. Task-Space Sequential Equilibrium and Inverse Kinematics Optimization (SEIKO) runs in the backend to guarantee the physical constraints and safety of the dual-arm robot [27].

The Learning from Demonstration (LfD) framework (see Fig. 2) is composed of two distinct networks: 1) a Convolutional AutoEncoder (CAE) network to extract the latent features from tactile images; 2) a Behavior Cloning (BC) network to learn the policy of dexterous dual-arm grasping with tactile sensing from human demonstrations.

### B. Demonstration Dataset of Bimanual Manipulation

In our implementation, the haptic devices allow operators to adjust the 6D pose simultaneously, providing an intuitive way to demonstrate bimanual grasping skills on a dual-arm robot. During the demonstration, a human operator teleoperates the dual-arm robot to complete the grasping task by sending Cartesian commands to the two end-effectors via two haptic devices. The human demonstration data are recorded automatically during the entire grasping.

### C. Tactile Feature Extraction

The Tactip used in this work is an optical tactile sensor with a soft hemispherical tip, which was 3D-printed in one piece combining an elastic skin with 330 rigid white markers (pins) [7]. When the soft tip deforms during contact with objects, the white pins start to move away from their initial positions. The displacement of these pins reflects the complex deformation of the soft surface. An inner camera captures and projects the displacement to an array of white pins on a black background in the image plane. Raw tactile RGB images are firstly resized to 256×256 pixels using linear interpolation and converted to grayscale images, which are then cropped using a circle mask and converted to binary images by thresholding. A median filter is applied to denoise the binary images.

We propose to use a self-supervised learning method – convolutional autoencoder network to extract robust features that can represent the contact properties from the preprocessed tactile images. Eight convolutional layers are used in the CAE network to extract the spatial information represented by the displacement of the pins. The CAE network consists of an encoder and a decoder, formulated as follows:

$$\begin{aligned} g_\Theta(\cdot) &: \mathcal{X} \rightarrow \mathcal{H} \\ f_\Phi(\cdot) &: \mathcal{H} \rightarrow \hat{\mathcal{X}} \end{aligned} . \tag{1}$$

The encoder $g_\Theta(\cdot)$ projects each tactile image $\gamma_t$ in the high-dimensional input space $\mathcal{X}$ (256×256) to 16 feature maps $\gamma_l$ in the low-dimensional latent space $\mathcal{H}$ (16×16), then the decoder $f_\Phi(\cdot)$ reconstructs that image from the same feature maps to the output space $\hat{\mathcal{X}}$ (256×256). The binary cross-entropy loss function is used as the reconstruction loss between the input images $\mathcal{X}$ and the reconstructed images $\hat{\mathcal{X}}$ to update the network parameters via back-propagation:

$$\begin{aligned} L_{CAE}(\gamma_t, \gamma_p) = -(\gamma_t \log \gamma_p) + (1-\gamma_t)\log(1-\gamma_p) \\ \gamma_l = g_\Theta(\gamma_t), \gamma_p = f_\Phi(\gamma_l) \end{aligned} , \tag{2}$$

where $\gamma_p$ is the reconstructed image by the decoder network.

### D. Behavior Cloning Network

We propose and design a BC network to learn the behaviors of coordinated manipulation skills of bimanual

grasping from human demonstration data. Dexterous bimanual grasping skills can be considered into two categories: (1) adaptive interaction with objects, and (2) dual-arm motion coordination. To capture these skills, we have designed the input to our network to include encoded tactile feature maps, tactile image differences, and the robot's proprioceptive state. The encoded feature maps and tactile image differences capture the human-object interaction skills. The robot's proprioceptive state, on the other hand, offers insights into the coordination of movements between both arms. These inputs collectively serve to reflect the complexity and adaptability of dexterous grasping skills.

Following this idea, we use the encoded tactile feature maps $l_t$, the proprioceptive state $\phi_t$, and the tactile image difference $e_t$ as input to the BC network to represent and learn fine human skills. The discrete-time state-action pair set $G = \{(s_0, a_0), (s_1, a_1), ..., (s_t, a_t), ...\}$ is created to train the BC network, where $s_t = (l_t, \phi_t, e_t)$ denotes the robot state and $a_t$ denotes the Cartesian commands of the two arms at time $t$.

Using such data of multiple modalities as input to train a network requires a well-crafted embedding structure [28]. A common way of fusing a 2D feature map and a 1D feature vector is to flatten the 2D feature map into a 1D vector and concatenate the flattened vector and the 1D feature vector [29]. However, we found that the flattening projection results in the *loss of spatial correlation of tactile information*. For fine dexterous pinch grasping of small objects, the spatial information provided by tactile images is essential in understanding the contact situation and the object being contacted. To preserve the spatial information of the encoded tactile feature maps, we specifically tile the proprioceptive state of robots and the tactile image difference to *match* the dimension of the tactile feature maps, so as to keep the spatial information of the encoded tactile feature maps.

We then concatenate the tactile feature maps, the tiled proprioceptive state maps, and the tactile image difference on each feature channel. The convolutional layers in the BC network first filter the input feature maps (46×16×16) to a feature map (1×8×8), which is then flattened and fed into a fully connected network (FCN). The FCN network outputs a vector $\hat{\boldsymbol{a}} \in \mathbb{R}^{12}$ as the predicted Cartesian pose commands of the two arms, including 3D position and 3D orientation for each arm.

The loss function used to train the BC network consists of two parts, which are formulated as:

$$\begin{gathered} L_{BC}(\boldsymbol{a}, \hat{\boldsymbol{a}}) = \|\boldsymbol{a} - \hat{\boldsymbol{a}}\|^2 + \left\|\boldsymbol{d} - \hat{\boldsymbol{d}}\right\|^2 \\ \hat{\boldsymbol{a}} = \psi(l, \phi, e; \Phi_{conv}, \Phi_{fcn}) \end{gathered} \tag{3}$$

where $\boldsymbol{a} \in \mathbb{R}^{12}$ is the Cartesian pose commands of the two arms from the human demonstration dataset, and $\hat{\boldsymbol{a}} \in \mathbb{R}^{12}$ is the predicted Cartesian pose command by the BC network $\psi(\cdot; \Phi_{conv}, \Phi_{fcn})$, parameterized by $\Phi_{conv}$ and $\Phi_{fcn}$; $l$, $\phi$ and $e$ denote the tactile feature maps, the proprioceptive state maps and the tactile image difference, respectively. The second term $\left\|\boldsymbol{d} - \hat{\boldsymbol{d}}\right\|^2$ is added to learn the dual-arm coordination skills from human demonstrations, where $\boldsymbol{d} \in \mathbb{R}^3$ is the relative position between the two end-effectors, and $\hat{\boldsymbol{d}} \in \mathbb{R}^3$ is the predicted relative position between the two end-effectors by the BC network.

## IV. Experiments and Results

### A. Experimental Setup and Data Collection

We validate the performance of LfD with tactile sensing for robot dexterous manipulation on the challenging task: the retrieval of small and fragile objects from the desk using dual-arm pinch grasp and and ensuring a stable grasp throughout the process. During dexterous grasping, external vision can be easily occluded by the end-effector, potentially leading to inaccurate object estimation. Therefore, our experiments operate without using external visual sensors. By default, the starting position of the object lies between two robot hands, and the whole demonstration is operated in the task space.

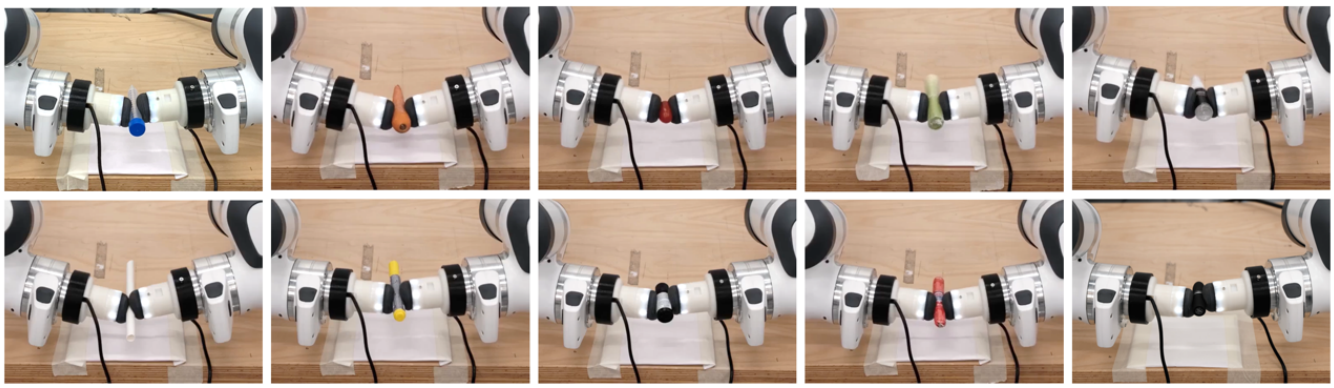
We collected 5 demonstrations for this task. The human demonstration dataset collected in the grasping task includes three main components: the Cartesian commands, the proprioceptive states, and the tactile feedback (i.e., tactile images provided by the TacTip sensors). The Cartesian commands and the proprioceptive states of the two arms are collected at a frequency of 1000 Hz. Two Tactips record the tactile image pairs at a frequency of 60 Hz. For each demonstration, about 1500 tactile images are recorded. Before using the collected dataset to train the networks, several pre-processing methods are used to process the raw data. The proprioceptive states of the two arms and the tactile images, collected at different sampling rates, are synchronized using a linear interpolation method to align their timestamps. A median filter is then applied to smooth the Cartesian commands $a_t$, i.e., the 6D poses of two end-effectors. For raw tactile images, the structural similarity index measure (SSIM) [30] is used to quantify the difference between the current frame and the original frame, serving as a preliminary metric for estimating contact forces.

### B. Implementation detail

Our proposed model is developed using PyTorch [31]. For the training of the CAE, we utilized a dataset that comprised 20 trials, including demonstrations of random behavior unrelated to our experiment. This dataset was collected using our robot setup, with each trial yielding approximately 1500 images captured during the demonstration phases. The trained CAE exhibits a satisfactory reconstruction quality, with a Mean Squared Error (MSE) loss of 0.015 and a Structural Similarity Index Measure (SSIM) of 0.934. The model training process for CAE, which involved 100 iterations, was completed in approximately two hours using an NVIDIA 1080 GPU. In the case of the BC network, the model was trained using data from 5 collected demonstrations, with the training process involving over 1,000 iterations and taking approximately 5 minutes to complete.

(a) Robustness of the learned control policy against external disturbances.

(b) Successful re-grasping using the learned policy.



(c) Successful grasping of unseen objects using the learned policy.

Fig. 3: Generalization of the learned policy and its robustness to external disturbances.

### C. Design of Validation Tasks

*1) Learning grasping vial:* The human demonstrator performs teleoperation of dual-arm robots to grasp a plastic vial (a test tube with $\Phi = 15.65$mm) that is horizontally placed on the table. A Behavior Cloning (BC) network is trained using the gathered demonstration data, and the trained policy is tested on dual-arm robots to validate its generalization on unseen initial poses. During the evaluative phase, we positioned the test tube between the end-effectors to evaluate the performance of the learned policy given variations in the starting position, specifically alterations of up to $\pm 20$ degrees and displacements of up to $\pm 2$ centimetres in the objects' locations.

*2) Generalization to unseen objects:* To evaluate the generalizability of the trained policy to unseen objects with a variation of radius, weight, or even materials (e.g., soft and fragile objects), a set of test experiments have been conducted using multiple objects of different radii ranging from 11.7mm to 28.6mm.

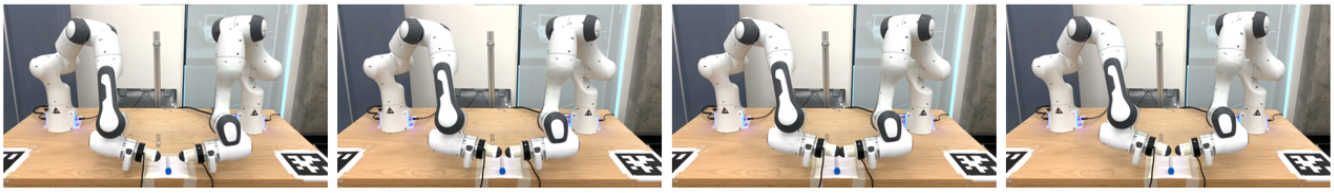
*3) Robustness against external disturbance:* We also validate the robustness of the trained policy against external disturbances. We applied random external pushes from the left, right, up, and down directions on the grasped object to test if the two arms can coordinate their end-effectors' poses to ensure the balance of the object.

*4) Re-grasping capability:* The re-grasping experiments are conducted to test if the trained controller is contact-awareness and can perceive the loss of contact with the object in order to make necessary adjustments according to the tactile feedback and react to grasping failures. After the successful normal grasping, we severely pushed the object away to break its static equilibrium, and the object dropped down between two end-effectors again.

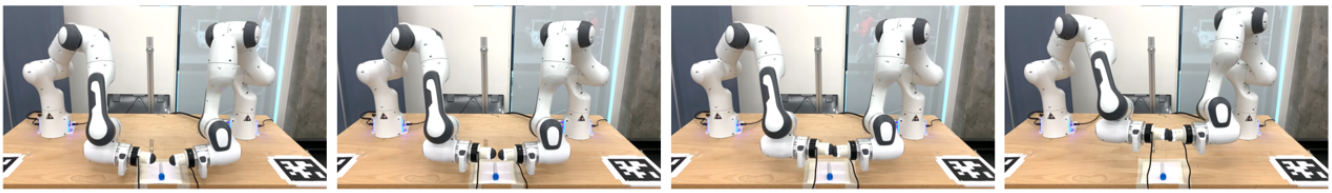
### D. Results of Grasping Tasks

The BC network trained on human demonstration data is deployed on a real dual-arm robot to verify its performance by the designed tasks. In both grasping tasks, the learned control policy achieved a $95\%$ success rate, even when the initial poses of the tube were different from their original pose in the demonstration. Occasionally, the policy might not succeed in grasping the object on the first attempt. However, it consistently adapts the behavior based on the contact situation with objects and will attempt to repeat the grasping action upon failure. The policy fails to grasp an object only when the object falls outside the workspace boundaries.

The dual-arm robot can make prompt adjustments and enable stable dexterous grasping by learning from only few demonstrations. In the process of lifting the object, the dual-arm robot achieves stable grasping by constantly twiddling the "fingertips" (tips of TacTip sensors) and adjusting the



(a) Unsuccessful grasping of the baseline policy using exactly the same BC network trained with unchanged tactile feedback.



(b) Unsuccessful grasping of the baseline FCN policy trained with only the proprioperceptive information of end-effectors' poses and positions.

Fig. 4: Results of the comparison study. The policy trained with both comparison frameworks bypass the object directly, manoeuvring the end-effectors directly to the desired end-poses without making any physical contact, grasping attempts, or interactions with the tube.

object to the central position. The process of retrieving an object from the table and adjusting its pose to maintain balance requires very fine movements and interactions supported by rich tactile information, where a 6-axis force/torque information is not sufficient to discern different contact situations in this scenario.

We evaluate the robustness of the learned policy against external disturbances. It can be seen from Fig. 3a that the dual-arm robot can make a proper adjustment to adapt to pushes. Although the pose of the two-arm robot in contact with the object was changed each time while being pushed, the dual-arm robot can always fine-adjust the object reactively to the center of the fingertips (Tactip sensors), roll and move the object to the desired position. Compared with the manually programmed behavior, this serves as a feedback policy that has been successfully acquired from human dexterity skills, which enables the dual-arm robot to autonomously adjust the posture and ensure a stable grasp quickly. It is noteworthy that such active rolling adjustment has not been specially demonstrated by any separate trials, but rather, this behavior was successfully captured by the rich tactile data during the demonstration of pick-lift grasping.

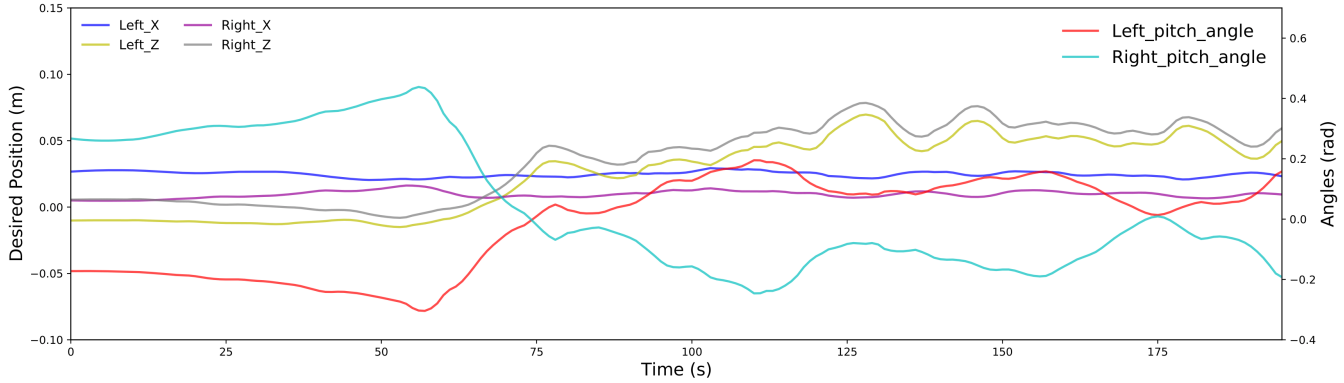
To examine the reaction in the presence of an unknown situation, i.e., grasping failures, the learned policy demonstrated contract-awareness of the falling object, i.e., loss of contact according to the tactile feedback, and thus controls the robot to restart the grasping process, which was not explicitly programmed or demonstrated by the prior LfD data. The result of the re-grasping experiments in Fig. 3b shows that the tactile-based control learned from human demonstrations is very effective in performing robotic dexterous bimanual manipulation tasks autonomously and quickly without the need for explicit manual programming or complex planning.

The policy also achieves successful grasping of previously unseen objects, as shown in Fig 3c. Although the test objects have a variety of sizes and weights compared with the object used in the demonstration, the policy can still perform stable grasping. The experiment results show that the trained policy can generalize to unseen objects with similar cylindrical shapes but with different sizes and weights.

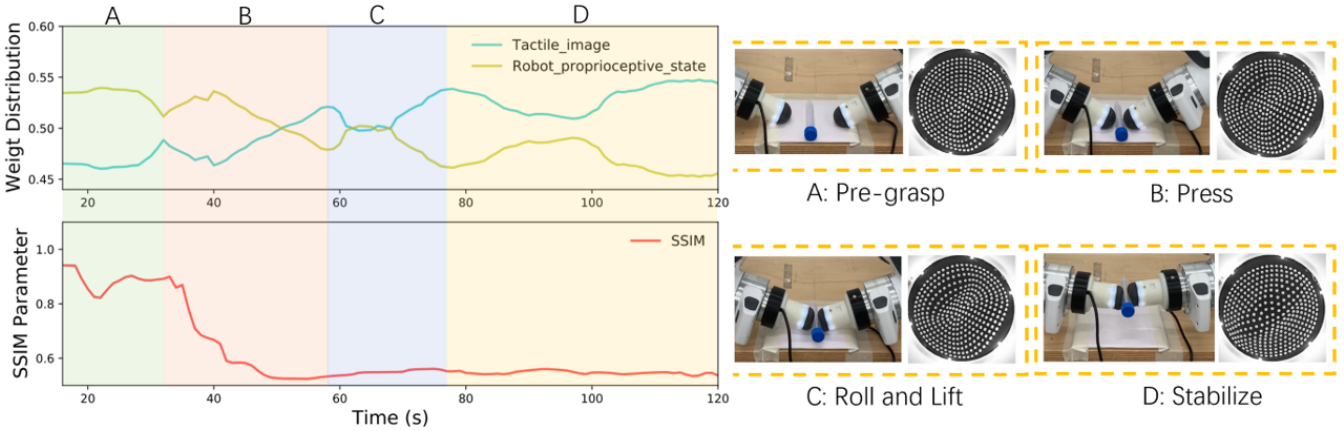
### *E. Comparison Study*

We conducted a comparison study to validate that successful grasping is achieved by the active use of tactile sensing. Besides training a BC network using the structure shown in Fig. 2, we also train two different BC networks for comparison. The first one has exactly the same BC network structure but with frozen tactile images as input to CAE, meaning that the encoded image feature input stays unchanged during both the training and testing. The second one has an FCN structure and uses the poses of two end-effectors (both positions and orientation) as the input to train the network. The proposed BC network demonstrates convergence to a loss of 0.04 on the testing set. In contrast, the network employing frozen tactile information achieves convergence with a loss of 0.5, while the FCN converges to a loss value of 1. These results prove that the effective integration of tactile information significantly enhances the convergence rate and leads to a reduced loss value in the final model.

We also compared all the grasping performances of the real dual-arm robot. As shown in Fig. 4, both BC network structures without using the tactile information failed in grasping the tube: robot arms failed to approach the object from its initial pose, and instead, they bypassed the object and moved towards the desired end-poses, showing no contact-awareness. The experimental results indicate that tactile feedback plays an essential role in providing contact information for initiating contacts, generating appropriate adjustments, lifting and retrieving to the desired target locations, enabling the dual-arm robot to perform very fine and dexterous contact-rich skills.




(a) The output of the desired positions and angles of the BC network for the dual-arm robot during dexterous grasping.




(b) Relative weight changes of each modality and its corresponding tactile information.

Fig. 5: The output of the learned policy and the weight changes during the grasping.

### F. Interpretability

To explicitly show how much different modalities influence the entire operation, we use the saliency map method for calculating the weight distribution. The procedure for calculating this distribution is formulated as follows [32]:

$$W_i = \frac{N(I)}{N(I)+N(J)}, W_j = \frac{N(J)}{N(I)+N(J)} \quad , \tag{4}$$

where $W_i$ and $W_j$ are the weight distributions of each modality. $N(\cdot)$ represents the normalization process. $I$ is the importance of the tactile information that is calculated by adding all the absolute values of weight that the learned policy distributed to tactile features. $J$ is calculated in the same way by adding all the absolute values of weight that are distributed to robot proprioceptive state features.

The comprehensive process of dexterous pinch grasping can be subdivided into four primary stages: pre-grasp, pressing, rolling and lifting, and stabilization. Each of these stages utilizes tactile feedback in a distinct manner. In Fig 5b, the weight changes during the complete dexterous pinch grasping process are depicted. Initially, as the end-effector moves toward the objects without any contact deformation on the tactile sensor, the weight of the robot's proprioceptive state exceeds that of the tactile information. When the tactile sensor comes into contact with the desk and is prepared for a pre-grasp pose, the weight of the tactile information increases (stage A). As the end-effector moves towards the object and initiates contact, the weight attributed to the tactile information increases, exceeding that of the proprioceptive state (stage B). During the roll and lift phase, the weight of the tactile information initially decreases, subsequently achieving equilibrium with the proprioceptive state (stage C). This indicates that during the lifting phase, the learned policy necessitates both tactile information for successful in-hand manipulation and proprioceptive information for effective dual-arm coordination. Finally, upon successfully lifting the tube, the weight reverts to the tactile information, facilitating the stabilization of the tube (stage D).

## V. Conclusion And Future Work

In this work, we introduce a tactile-driven LfD framework that demonstrates promising results in bimanual pinch grasping with a limited number of real robot demonstrations. Our exploration into leveraging the latest compliant tactile sensors has led to the development of the presented encoding methods that can effectively extract and capture high-dimensional contact sensing from soft tactile sensors, together with the fusion with proprioceptive feedback. The interesting outcome is to confirm the possibility of learning from real robot data directly, eliminating the necessity for large datasets and extensive training time, if the right data is effectively used.

Our comparison studies showed that without tactile sensing, dexterous motor skills cannot be learned by few-shot demonstrations with traditional robot sensing which is rather limited. Despite the deformable sensor offering more compliant behavior during grasping, the extracted feature from the tactile image is negligible. Our approach demonstrates remarkable robustness in the presence of external pushes and is able to re-grasp the object if it drops. This ability was not explicitly illustrated in the initial demonstrations, emerging instead as a natural consequence of contact-aware sensorimotor skills through state-action mapping.

Meanwhile, one apparent limitation is that the skill needs to be trained on a specific task, and it can be generalized and robust only around neighbourhood situations within a category of similar tasks: generalization applies to new/unseen objects that are similar to the demonstrated object of certain variations. Another limitation is that the robot's performance is based on blind grasping and re-grasping, and has not yet utilized external visual perception. In the future, integration of the current framework with stereo vision could extend the versatility and dexterity of object manipulation. Overall, our proposed LfD framework provides an attractive solution for learning from a few demonstrations with tactile sensing and supports broad real-world applications in contact-rich manipulation tasks.

## REFERENCES


[1] H. Zhu, A. Gupta, A. Rajeswaran, S. Levine, and V. Kumar, "Dexterous manipulation with deep reinforcement learning: Efficient, general, and low-cost," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3651–3657.

[2] S. P. Arunachalam, S. Silwal, B. Evans, and L. Pinto, "Dexterous imitation made easy: A learning-based framework for efficient dexterous manipulation," *arXiv preprint arXiv:2203.13251*, 2022.

[3] Z. Sun, K. Yuan, W. Hu, C. Yang, and Z. Li, "Learning pregrasp manipulation of objects from ungraspable poses," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9917–9923.

[4] Y. Qin, B. Huang, Z.-H. Yin, H. Su, and X. Wang, "Dexpoint: Generalizable point cloud reinforcement learning for sim-to-real dexterous manipulation," in *Conference on Robot Learning*. PMLR, 2023, pp. 594–605.

[5] Q. Li, O. Kroemer, Z. Su, F. F. Veiga, M. Kaboli, and H. J. Ritter, "A review of tactile information: Perception and action through touch," *IEEE Transactions on Robotics*, vol. 36, no. 6, pp. 1619–1634, 2020.

[6] N. F. Lepora, "Soft biomimetic optical tactile sensing with the tactip: A review," *IEEE Sensors Journal*, vol. 21, no. 19, pp. 21 131–21 143, 2021.

[7] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, "The TacTip Family: Soft Optical Tactile Sensors with 3D-Printed Biomimetic Morphologies," *Soft Robotics*, vol. 5, no. 2, pp. 216–227, 4 2018.

[8] N. F. Lepora, Y. Lin, B. Money-Coomes, and J. Lloyd, "Digitac: A digit-tactip hybrid tactile sensor for comparing low-cost high-resolution robot touch," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9382–9388, 2022.

[9] W. Yuan, S. Dong, and E. Adelson, "GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force," *Sensors*, vol. 17, no. 12, p. 2762, 11 2017.

[10] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer *et al.*, "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[11] K. Simonyan, A. Vedaldi, and A. Zisserman, "Deep inside convolutional networks: Visualising image classification models and saliency maps," *arXiv preprint arXiv:1312.6034*, 2013.

[12] J. Jiang and S. Luo, "Robotic perception of object properties using tactile sensing," in *Tactile Sensing, Skill Learning, and Robotic Dexterous Manipulation*. Elsevier, 2022, pp. 23–44.

[13] M. Oller, M. P. i Lisbona, D. Berenson, and N. Fazeli, "Manipulation via membranes: High-resolution and highly deformable tactile sensing and control," in *Conference on Robot Learning*. PMLR, 2023, pp. 1850–1859.

[14] E. Psomopoulou, N. Pestell, F. Papadopoulos, J. Lloyd, Z. Doulgeri, and N. F. Lepora, "A robust controller for stable 3d pinching using tactile sensing," *IEEE Robotics and Automation Letters*, vol. 6, no. 4, pp. 8150–8157, 2021.

[15] Y. Lin, A. Church, M. Yang, H. Li, J. Lloyd, D. Zhang, and N. F. Lepora, "Bi-touch: Bimanual tactile manipulation with sim-to-real deep reinforcement learning," *IEEE Robotics and Automation Letters*, 2023.

[16] N. F. Lepora, A. Church, C. De Kerckhove, R. Hadsell, and J. Lloyd, "From pixels to percepts: Highly robust edge perception and contour following using deep learning and an optical biomimetic tactile sensor," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 2101–2107, 2019.

[17] J. Lloyd and N. F. Lepora, "Goal-driven robotic pushing using tactile and proprioceptive feedback," *IEEE Transactions on Robotics*, vol. 38, no. 2, pp. 1201–1212, 2021.

[18] M. Polic, I. Krajacic, N. Lepora, and M. Orsag, "Convolutional Autoencoder for Feature Extraction in Tactile Sensing," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 3671–3678, 10 2019.

[19] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," California Univ San Diego La Jolla Inst for Cognitive Science, Tech. Rep., 1985.

[20] S. Tian, F. Ebert, D. Jayaraman, M. Mudigonda, C. Finn, R. Calandra, and S. Levine, "Manipulation by feel: Touch-based control with deep predictive models," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 818–824.

[21] Z. Fu, T. Z. Zhao, and C. Finn, "Mobile aloha: Learning bimanual mobile manipulation with low-cost whole-body teleoperation," *arXiv preprint arXiv:2401.02117*, 2024.

[22] H. Kim, Y. Ohmura, and Y. Kuniyoshi, "Goal-conditioned dual-action imitation learning for dexterous dual-arm robot manipulation," *IEEE Transactions on Robotics*, 2024.

[23] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[24] I. Guzey, B. Evans, S. Chintala, and L. Pinto, "Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play," *arXiv preprint arXiv:2303.12076*, 2023.

[25] R. Wen, K. Yuan, Q. Wang, S. Heng, and Z. Li, "Force-guided high-precision grasping control of fragile and deformable objects using sEMG-based force prediction," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2762–2769, 2020.

[26] R. Wen, Q. Rouxel, M. Mistry, Z. Li, and C. Tiseo, "Collaborative bimanual manipulation using optimal motion adaptation and interaction control: Retargeting human commands to feasible robot control references," *IEEE Robotics & Automation Magazine*, 2023.

[27] Q. Rouxel, K. Yuan, R. Wen, and Z. Li, "Multicontact motion retargeting using whole-body optimization of full kinematics and sequential force equilibrium," *IEEE/ASME Transactions on Mechatronics*, vol. 27, no. 5, pp. 4188–4198, 2022.

[28] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[29] B. Akbulut, T. Girgin, A. Mehrabi, M. Asada, E. Ugur, and E. Oztop, "Bimanual rope manipulation skill synthesis through context dependent correction policy learning from human demonstration," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3904–3910.

[30] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[31] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," 2017.

[32] W. Yu, C. Yang, C. McGreavy, E. Triantafyllidis, G. Bellegarda, M. Shafiee, A. J. Ijspeert, and Z. Li, "Identifying important sensory feedback for learning locomotion skills," *Nature Machine Intelligence*, vol. 5, no. 8, pp. 919–932, 2023.