Canonical Representation and Force-Based Pretraining of 3D Tactile for Dexterous Visuo-Tactile Policy Learning

Tianhao Wu, Jinzhou Li*, Jiyao Zhang*, Mingdong Wu, Hao Dong

Abstract— **Tactile sensing plays a vital role in enabling robots to perform fine-grained, contact-rich tasks. However, the high dimensionality of tactile data, due to the large coverage on dexterous hands, poses significant challenges for effective tactile feature learning, especially for 3D tactile data, as there are no large standardized datasets and no strong pretrained backbones. To address these challenges, we propose a novel canonical representation that reduces the difficulty of 3D tactile feature learning and further introduces a force-based self-supervised pretraining task to capture both local and net force features, which are crucial for dexterous manipulation. Our method achieves an average success rate of 80% across three fine-grained, contact-rich dexterous manipulation tasks in real-world experiments, demonstrating effectiveness and robustness compared to other methods. Further analysis shows that our method fully utilizes both spatial and force information from 3D tactile data to accomplish the tasks. The videos can be viewed at `https://3dtacdex.github.io`.**

I. INTRODUCTION

Human hands play a crucial role in daily life [1], enabling a wide range of tasks such as opening boxes and flipping objects. This level of dexterity is essential for integrating robots into everyday human activities. Visual base imitation learning has shown great potential in teaching dexterous hands to perform various tasks [2], [3], [4]. While simpler tasks like pick-and-place operations can achieve high success rates, more fine-grained and contact-rich tasks—such as in-hand manipulation, remain significantly challenging. These tasks involve precise control of force, nuanced coordination of different fingers, and continuous feedback during manipulation. A key factor in successfully executing such tasks is tactile sensing [5].

To enable dexterous hands to perceive contact, current approaches typically equip them with tactile sensors. These sensors can be mainly categorized into vision-based tactile sensors, such as GelSight [6], DIGIT [7], and distributed tactile sensors, like uSkin [8]. Distributed tactile sensors are particularly well-suited for various robotic structures due to their small size, which allows for easy integration. Their robustness also makes them reliable in diverse environments, leading to widespread use in many systems [9], [10], [11].

However, distributed tactile sensors typically cover a large area of the dexterous hand [12], resulting in high-dimensional input. Additionally, different dexterous hands often use different types of distributed tactile sensors with varying sensor distributions, leading to a lack of large-scale standardized datasets. This poses unique challenges in learning tactile features effectively to improve dexterous manipulation.

Authors are with the Center on Frontiers of Computing Studies, School of Computer Science, Peking University, Beijing 100871, China, also with PKU-Agibot Lab, School of Computer Science, Peking University, Beijing 100871, China, and also with National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University, Beijing 100871, China.

* indicates equal contribution.

Corresponding to hao.dong@pku.edu.cn.

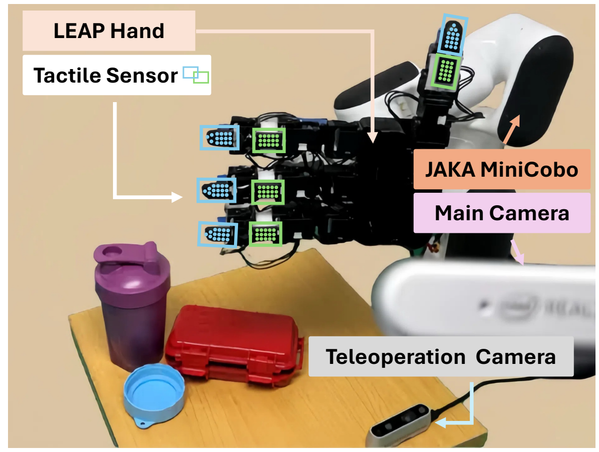

Fig. 1. **Real Robot System.** Our system uses a single camera and distributed tactile sensors to achieve fine-grained, contact-rich tasks. Note the teleoperation camera is only used for data collection, without policy learning.

Considering the power of visual backbones, many works [13], [4] convert tactile data into 2D images to reduce the complexity of learning useful tactile features. While this transformation aids feature learning, it leads to the loss of spatial information between different taxels. To preserve these spatial relationships, most approaches represent 3D tactile data as a graph and use graph neural networks (GNNs)[14] to encode tactile signals[12], [15]. However, these methods primarily focus on specific tasks and require a large amount of data to learn effective features. Inspired by the success of pretraining strategy in vision-based learning, T-DEX [4] collects tactile play data through self-supervised interaction with various objects. This pretraining improves the efficiency of feature learning and enhances diverse downstream robotic manipulation tasks. However, it still relies on 2D images as the tactile representation. As a result, efficiently learning 3D tactile data features for dexterous manipulation remains a challenge.

To address the difficulty of 3D tactile feature learning, we first propose a novel canonical representation of 3D tactile data, which canonicalizes the coordinates of taxels in each sensor into a unified coordinate system. This canonicalization aligns the features of different sensors and reduces the feature space. Additionally, it amplifies the distances between taxels within the same sensor, facilitating the capture of more localized features. We further propose a force-based, self-supervised prediction task for pretraining 3D tactile data, given the importance of force usage in object manipulation. The pretraining tasks include both masked local force prediction and net force prediction, encouraging the encoder to learn features related to both local and net force relationships.

To demonstrate the effectiveness of our method, we incorporate the pretrained tactile encoder into an imitation learning framework and evaluate its manipulation performance in the real world on three fine-grained, contact-rich tasks: open box, reorientation, and flip. Comparative results demonstrate the effectiveness of our method compared to other baselines. Ablation studies confirm the importance of our proposed canonical representation and force-based pretraining. Additionally, our analysis shows that the policy effectively utilizes both the spatial and force information from the 3D tactile data.

In summary, our contributions are as follows: (i) We propose a novel canonical representation for 3D tactile data that effectively improves 3D tactile data feature learning. (ii) We propose a novel force-based self-supervised pertaining task on tactile play data, including local force and net force prediction, enhancing downstream dexterous manipulation policy learning. (iii) We demonstrate the effectiveness and robustness of our method through a range of real-world experiments using a dexterous hand.

II. RELATED WORK

A. Tactile for Dexterous Manipulation

Tactile sensing has been widely used to enhance dexterous manipulation. Current approaches primarily utilize either vision-based tactile sensors [6] or distributed tactile sensors [8]. Due to their larger size, vision-based tactile sensors are typically only mounted on the fingertips of dexterous hands [16]. In contrast, distributed tactile sensors can cover a larger area [10]. Thus, we choose to use distributed tactile sensors in this work.

There are two main approaches to learning manipulation policies with distributed tactile sensors. One is simulation-to-reality, where simulation can generate large amounts of tactile data, making the learning process more efficient. However, there is a significant gap between tactile data in simulation and the real world. To close the sim-to-real gap, most works use discrete tactile signals [17], [18], [19] or only activated tactile positions [12] as input, limiting the full potential of tactile sensors. The other approach is learning directly in the real world. Due to the large coverage area of distributed sensors, tactile input can be high-dimensional, especially for dexterous hands, posing challenges for efficient learning. Pretraining with play data has been proposed to improve efficiency [4], [9], but these works rely only on 2D tactile images. In contrast, we propose a pretraining strategy and representation specifically for 3D tactile data.

B. Tactile Representaiton

Different tactile representations convey various types of information and can be encoded using different methods. For low-dimensional tactile data, directly applying MLP to flattened tactile readings [20] or converted 3D vectors [21] can still capture useful tactile features. However, as the tactile sensor coverage increases, the dimensionality of the data grows significantly, making direct encoding of tactile readings inefficient. To leverage powerful visual backbones, some methods convert raw tactile readings into RGB images by mapping the tri-axis forces to three channels [22], [13], [23], using visual backbones for encoding. However, such 2D information changes the inherent spatial relationship of taxels in the same sensor and does not contain the spatial relationship of taxels in different sensors that are distributed on the different parts of the robot. To preserve spatial relationships, graph-based methods have been applied to tactile data by treating each taxel as a node, connecting them with either predefined [15] or dynamically changing graph [12]. Nevertheless, the representation in these works only use a subset of the available tactile information in 3D space, such as the 3D position [12] or 3D force of the taxels [15]. Our work fully leverages both the 6D pose and 3D force of each taxel, and we propose a novel canonical representation to more effectively learn features from such complex tactile data.

C. Tactile Pretraining

Due to the high-dimensional of tactile data, pretraining is an effective strategy for improving the efficiency of downstream task learning. Different pretraining strategies encourage the encoder to learn distinct features. Aligning vision and tactile data has been widely studied for pretraining to understand relationships between different data modalities [24], [21], [25], [26]. However, these approaches primarily focus on inter-modal pretraining for multi-modal learning. Our work focuses on intra-modal pretraining.

A common approach for intra-modal pretraining typically involves augmenting the data and encouraging the encoder to match the augmented data with the original [27], enhancing the encoder's ability to discriminate between different data patterns. However, most powerful intra-modal pretraining methods are designed for 2D images, which require representing tactile data as 2D images [4], [28], failing to fully leverage the spatial information in 3D tactile data. In contrast, we focus on pretraining for 3D tactile data, and instead of enhancing the encoder's discriminative ability, we encourage it to learn features related to force.

III. ROBOT SYSTEM SETUP

As shown in Fig. 1, our system consists of a 6-Dof JAKA MiniCobo robot arm and a 16-Dof Leap Hand [29] dexterous hand with four fingers. The Leap Hand is equipped with PaXini tactile sensors, Each finger has two type of sensors: one for the fingertip and another for the fingerpad. Both types of sensors have a 3x5 array of taxels, but taxel distribution is slightly different. Each taxel measuring tri-axial forces

$\mathbf{F} \in \mathbb{R}^3$. A single Intel RealSense D415 camera is mounted diagonally on the robot to capture visual information.

For expert demonstration collection, we use an additional Intel RealSense D415 camera with HaMeR [30] to track human hand pose, use Dexpilot [31] to retarget and teleoperate the robot. The robot arm is controlled with a target end-effector pose consisting of 3-Dof translation and 4-Dof quaternion, while the robot hand is controlled with 16-Dof target hand joint positions. Both demonstration collection and inference are performed at a frequency of 5 Hz.

IV. METHOD

We focus on the problem of leveraging 3D tactile data from distributed tactile sensors for learning visuo-tactile dexterous manipulation policies. To reduce the difficulty of learning features from complex 3D tactile data, we canonicalize the data into a unit coordinate system, as described in IV-A. We then pretrain the tactile encoder using self-supervised force-based prediction tasks to enhance local and net force feature learning, detailed in IV-B. This pretrained tactile encoder is subsequently used for visuo-tactile policy learning, as outlined in IV-C.

A. Canonical Tactile Representation

To preserve the spatial relationships of each taxel, we aim to use 3D tactile data instead of converting it into a 2D image. For each taxel of the distributed tactile sensor, in addition to the 3D force $\mathbf{F}$, we can also obtain the 6D pose $\mathbf{P} \in \mathbb{R}^6$ by computing forward kinematics. This information shows how the force is applied at every frame. However, since a large number of taxels are distributed across different parts of the fingers, using this 9D tactile representation results in a vast feature space, making it difficult to learn meaningful tactile features. Additionally, though the taxels within a sensor are distributed sparsely, the distances between them are very small (e.g., less than 2 millimeters), making it challenging to capture local features within the same sensor.

To address the challenges, we propose to canonicalize the 9D tactile representation. Specifically, we normalize each taxel's coordinate within the same sensor into a unit coordinate system (ranging from -1 to 1 for each axis) by computing the diagonal length of the original coordinates within the sensor's coordinate system. As shown in Fig. 3, the 3D position of each taxel in the unit coordinate system is denoted as $\mathbf{T} \in \mathbb{R}^3$. However, this representation only captures the spatial relationships between taxels within the same sensor, without accounting for the spatial relationships between different sensors. Therefore, we also include the 6D pose of each sensor's origin with respect to the hand's base, denoted as $\mathbf{P}^s$, into the representation. As a result, the representation for each taxel is represented as $\mathbf{R} = [\mathbf{P}^s, \mathbf{T}, \mathbf{F}]$. Although this representation has a higher dimension, it effectively reduces the feature space because the features of different sensors become more aligned due to the canonicalized coordinates. Additionally, this canonicalization amplifies the relative distance between taxels within the same sensor, making their features more distinguishable for the neural network. This facilitates the capture of more localized features for each taxel.

However, this representation still suffers from the inherent sparsity of the distributed tactile sensor. To address this, we utilize a graph neural network [32] to encode our proposed representation. We define the tactile information as a set of our proposed 12D representations, e.g., $\mathbf{S} = \{\mathbf{R}_1, \mathbf{R}_2, ..., \mathbf{R}_n\}$. Based on $\mathbf{S}$, we construct the graph $\mathbf{G} = (\mathbf{S}, \mathbf{E})$, where $\mathbf{E}$ represents the edges defined by the 4-neighbourhood of each tactile node.

B. Force-based Pretraining

While the canonical tactile representation can ease the difficulty of tactile feature learning, it does not ensure that the neural network will learn the features essential for manipulation and can be low-efficiency if trained on specific tasks only. Pretraining, however, can encourage the encoder to learn the inherent structures of the data [28] and improve the efficiency of learning for downstream tasks [4].

What kind of pretraining can we use for 3D tactile data to improve dexterous manipulation policies? When humans manipulate objects, we carefully apply force to achieve the desired object pose. Inspired by this, we propose pretraining the 3D tactile data based on force. When applying force to an object, it is essential to consider how each part of the finger applies local force so that the net force moves the object as intended. Consequently, we designed two force-based self-supervised pretraining tasks: the first predicts the local force, and the second predicts the net force. Since our robot system differs from T-DEX [4], we follow their method to collect our own play data for pretraining. We use GraphMAE [32] as the backbone for pretraining.

Local Force Prediction: To help the encoder learn the local features of each taxel, we design a masked force prediction task. Since the force applied to each taxel can propagate to its neighboring taxels, we randomly mask part of the tactile force in the play data and use the masked tactile data as input for the encoder. The encoder first encodes the tactile data, then decodes the latent representation to reconstruct the original tactile data, as shown in Fig. 2. We use the MSE loss to compute the loss between the reconstructed and original force values, only for the masked forces. This pretraining approach helps the encoder learn the relationships between local forces.

Net Force Prediction: To help the encoder understand the relationship between local and net forces, we design a self-supervised task for net force prediction. Given the original 3D tactile data, we compute the net force $\mathbf{F}_G^n$ based on each taxel's pose and force. This $\mathbf{F}_G^n$ serves as the target for prediction. As shown in Fig. 2, after predicting the local force, we substitute the predicted force values into the original tactile data and use the same encoder to encode this modified data into a latent representation. We then use an MLP to predict the net force $\mathbf{F}_P^n$. The MSE loss is calculated between $\mathbf{F}_G^n$ and $\mathbf{F}_P^n$. By further predicting the net force, the encoder learns to capture the feature of both local and net force, which benefits downstream tasks.

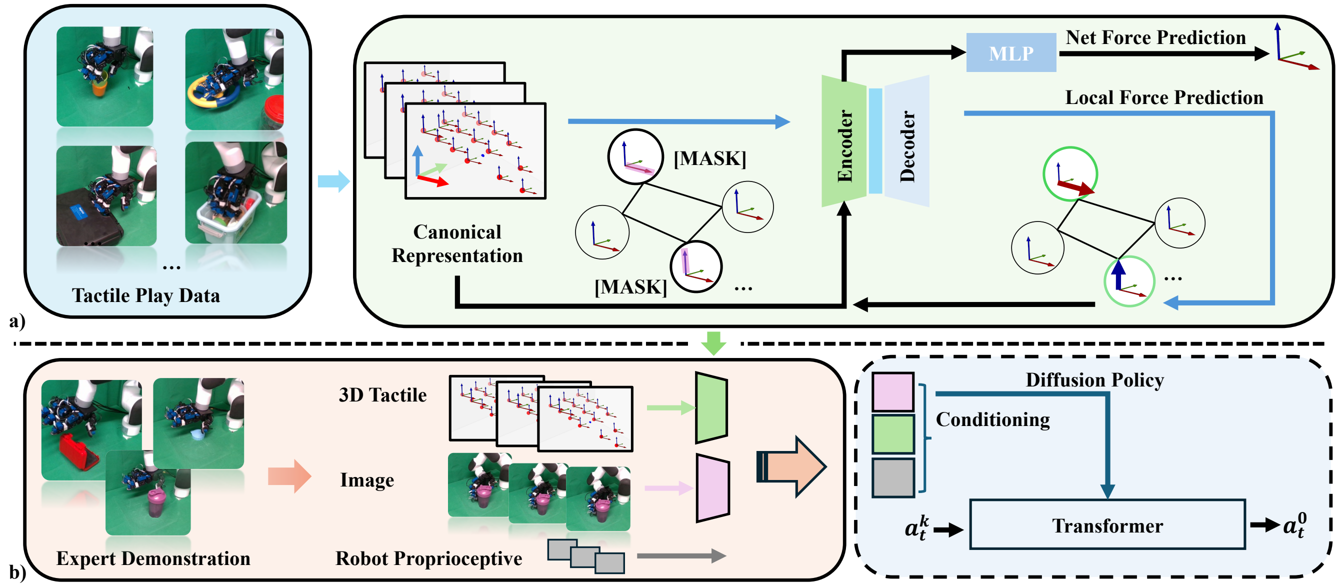

Fig. 2. **Pipeline.** a) Pretraining on play data with our canonical representation and force-based task. Local force prediction: a portion of the tactile force is randomly masked, encoded into a latent representation, and then decoded to predict the masked forces. Net force prediction: the predicted masked forces are substituted back into the original data and encoded again to predict the net force. The local force prediction and net force prediction share the same encoder. b) Incorporating the pretrained encoder within the imitation learning framework for downstream dexterous manipulation.

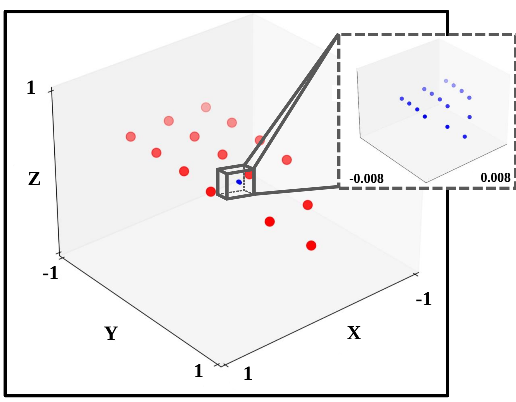

Fig. 3. **Comparison of Canonical Representation and Original Representation.** We visualize the coordinates of each taxel in the fingertip sensor before and after canonicalization. The blue dots represent the original taxel coordinates, which are difficult to distinguish when input to the encoder. In contrast, the red dots represent the taxel coordinates after canonicalization. With canonicalization, the coordinates of each taxel become more discriminative, and sensors of the same type have a consistent representation, reducing the feature space. Canonicalization with diagonal length maintains the inherent spatial relationships between taxels on each sensor pad, the same as the original representation.

C. *Visuo-Tactile Policy Learning*

After obtaining the pretrained tactile encoder, we utilize an imitation learning framework to learn visuo-tactile policy for dexterous manipulation tasks. Specifically, we adopt the diffusion policy [33] as our backbone. However, we replace the vision backbone from ResNet [34] to DinoV2 [35] for improved visual feature extraction. Unlike the original diffusion policy, we incorporate tactile data as an additional input, using the pretrained tactile encoder to encode this data. The output features are then concatenated with visual features as a condition for policy learning. We fine-tune the encoder during downstream tasks training, following diffusion policy [33]. In addition to visual and tactile inputs, we include the robot's proprioceptive state, including 3D position, 4D quaternion of the robot arm, and 16D joint position of the dexterous hand. Notably, our action space consists of target actions that humans intended to achieve during teleoperation, rather than states, as these actions implicitly capture the use of force, which is crucial for accomplishing a variety of tasks [36], [37].

V. EXPERIMENTS

We conduct comprehensive real-world experiments to validate the following questions:

- Does our canonical tactile representation can help to learn features of complex 3D tactile data?
- Does our force-based pertaining can improve visuo-tactile policy?
- What kind of role does the spatial information and force information of tactile data play during dexterous manipulation?

A. *Dexteous Manipulation Tasks*

We conduct experiments on three dexterous fine-grained, contact-rich manipulation tasks, as shown in Fig. 4. Each experiment run will be limited to a maximum of 500 steps. Each method will be evaluated on each task of 10 experiment runs.

1) Open Box: This task requires the robot to open a box using the thumb and index finger. The robot needs to first reach the box, then grasp the upper part with the thumb and index finger. After grasping, the robot needs to hold the upper part firmly and carefully adjust its fingers and wrist to gradually open the box without pushing it. The challenge

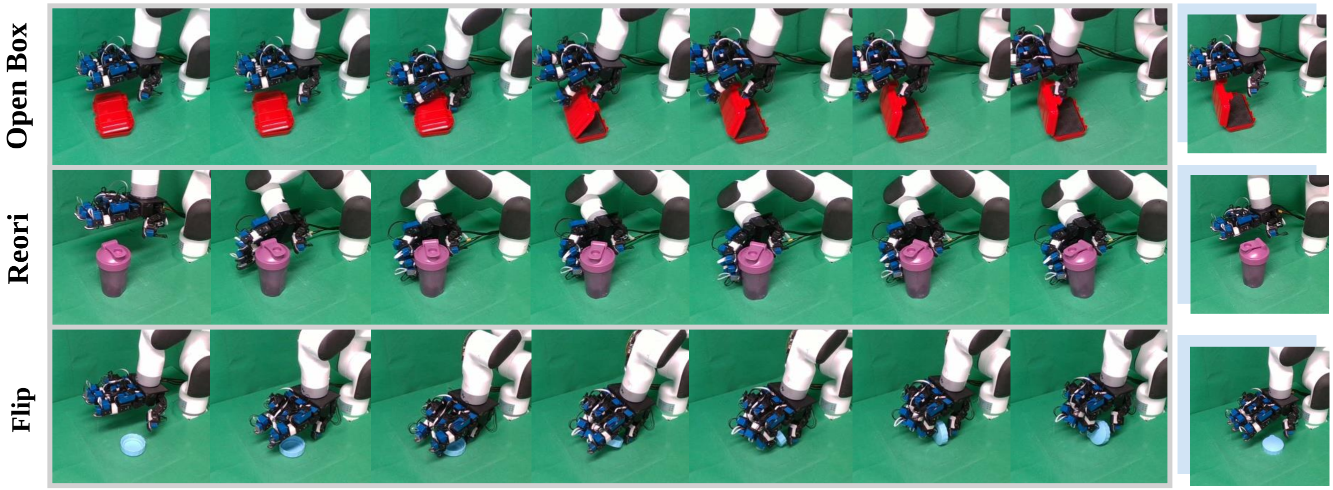

Fig. 4. **Visualization of Our Policy's Rollout on Three Fine-Grained, Contact-Rich Tasks.** Note this view is the view of robot's observation.

lies in maintaining a firm hold on the upper part of the box; otherwise, it may become loose and fall. The box will be placed in a random position and orientation within an 18x12 cm area for each experimental run. The task is considered successful if the upper part of the box does not fall after opening.

2) Reorientation: This task requires the robot to continuously reorient a bottle until it points in a specific direction. The robot needs to first reach to the bottle, and then coordinate the four fingers to reorient the bottle without pushing it down. The challenge lies in precise coordination among the fingers and the task is long-horizon. The bottle will be placed in an 18x12 cm area with a random position and orientation for each experimental run. The task is considered successful if the bottle is adjusted to point within 10 degrees of the goal direction.

3) Flip: This task requires the robot to flip a bottle cap. The robot needs to first reach the bottle cap, and then use the thumb and middle finger to grasp the bottle cap. After grasping, the robot needs to lift the middle finger and use the index finger to reorient the cap and may need further push it down to complete the task. The challenge lies in the precise coordination of the fingers and the correct application of force to manipulate the object. Additionally, there are severe occlusions during the manipulation process. The bottle cap will be placed in the same position but with random orientations, as a protruding part of the cap influences the manipulation. The task is considered successful if the bottle cap is flipped by 180 degrees.

B. Baselines

We compare our method with the following baselines, which all use the same visual backbone, diffusion policy backbone, visual observation, robot proprioceptive state, and action space as ours, but with different types of tactile representation and pertaining. 1) DP: We implement the diffusion policy without using tactile data for this baseline. 2) HATO: HATO [20] uses MLP to directly encode the tactile readings. We flatten 360-dimensional force values and use MLP to encode tactile data for this baseline. 3) T-DEX: T-DEX [4] convert the raw tactile readings into 2D image, and pretrain with BYOL [27]. For this baseline, we follow their procedure, first converting raw, tactile readings into 2D images, then pretrain on our own collected dataset and we use the pretrained encoder for diffusion policy learning. 4) GNN: We use the 9D tactile representation (e.g., the 6D pose of each taxel with respect to the hand's base with 3D tactile force values) as input for this baseline, use graph attention networks [38] to encode tactile, which is the same GNN backbone as ours. We do not pretrain for this baseline.

C. Manipulation Policy Comparsion

TABLE I
SUCCESS RATE OF DIFFERENT MANIPULATION POLICIES.

Method	Open Box	Reorientation	Flip	Avg
DP	90%	60%	20%	57%
HATO	70%	60%	10%	47%
T-DEX	80%	70%	40%	63%
GNN	0%	0%	0%	0%
Ours	**90%**	**70%**	**80%**	**80%**

As shown in Tab. I, our approach achieves the highest success rate across all tasks. While other baselines generally perform well on the open box and reorientation tasks, they struggle with the flip task. Interestingly, we found that even without tactile feedback, *DP* still achieves a high success rate on open box and reorientation tasks. This is mainly because these tasks do not involve significant occlusion or ambiguity during manipulation, allowing *DP* to successfully find and manipulate objects using visual input and robot state alone. In contrast, even with rich tactile information, *GNN* consistently fails across all tasks; once it reaches the object and attempts manipulation, the thumb either swings back and forth or the hand shakes, preventing further manipulation. Compared to *GNN*, *HATO*, which only uses tactile force values, is able to accomplish some tasks, demonstrating the difficulty of learning spatial and force information from 3D tactile data simultaneously. *T-DEX* performs better than the other baselines, showing that even 2D tactile data with pretraining can lead to high success rates, although it also struggles with the flip task.

The flip task requires extremely precise coordination between the fingers and relies on tactile feedback to ensure a firm grasp and accurate force application. For this task, we observed that *DP* hesitates to grasp the bottle cap and often reaches the maximum number of steps without succeeding, mainly due to the lack of tactile feedback. While *HATO* can reach the object accurately, it usually does not perform grasp or lift. *T-DEX* fails primarily due to an unstable initial grasp, which leads to difficulties during middle-finger lifting and index-finger reorientation. This underscores the importance of the spatial information provided by 3D tactile data.

D. Importance of Representation and Pretraining

TABLE II
SUCCESS RATE OF ABLATION. CR: OUR PROPOSED CANONICAL REPRESENTATION. PRE: OUR PROPOSED FORCE-BASED PERTAINING.

Method	Open Box	Reorientation	Flip	Avg
Ours w/o CR & PRE	0%	0%	0%	0%
Ours w/o PRE	60%	60%	50%	57%
Ours	**90%**	**70%**	**80%**	**80%**

To validate the effectiveness of our proposed canonical representation and force-based pretraining, we conduct ablation studies across all tasks. As shown in Tab. II, using the canonical representation of tactile data, even without pretraining, the trained policy achieves a 50% success rate. This indicates that the representation significantly reduces the complexity of learning from 3D tactile data. With force-based pretraining, the success rate increases to 80%, demonstrating that pretraining indeed enhances the performance of the visuo-tactile policy. By combining both canonical representation and pretraining, *Ours* can achieve robust success rate across both easy and difficult tasks.

E. Effect of Force-based Pretraining Tasks

TABLE III
SUCCESS RATE OF PERTAINING TASK. NF: NET FORCE PREDICTION. LF: MASKED LOCAL FORCE PREDICT.

Method	Open Box	Reorientation	Flip	Avg
Ours w/o NF	30%	30%	40%	33%
Ours w/o LF	70%	50%	30%	50%
Ours	**90%**	**70%**	**80%**	**80%**

To validate the effectiveness of our designed pretraining task, we conduct experiments using only one of the tasks as pretraining. As shown in Tab. III, omitting either task leads to a performance drop across all tasks, with net force proving to be more critical for achieving the tasks.

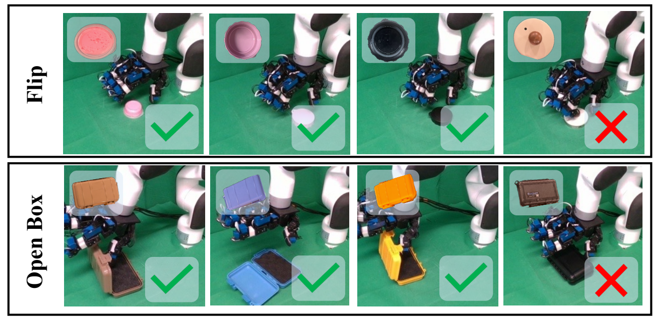

Fig. 5. **Visualization of Our Policy on Unseen Objects.**

F. Role of Spatial Information and Force Information

We conduct the ablation to validate the usage of spatial and force information of our policy. For the spatial information ablation, we kept the 6D pose information fixed at the initial state during manipulation, while for the force information ablation, we set the tactile force to zero throughout the manipulation. For both ablations, the robot failed to flip the bottle cap at all. In the spatial information ablation, we observed that once the robot reached the object and attempted to grasp it, the thumb began oscillating randomly, preventing the manipulation from continuing. In the force information ablation, although the robot reached the object and attempted to grasp it, it consistently failed due to an unstable grasp or continuously adjusting the grasp. This highlights that spatial information alone is insufficient; the policy still relies on force information to apply the correct force or perform subsequent manipulation.

G. Generlization

To validate the generalization of our method, we tested the policy with four unseen objects varying in color, geometry, and dynamics, with each object being tested twice for the open box and flip tasks. As shown in Fig. 5, our policy successfully opens the box 5 times out of 8 tries. For one failure of the open box task, although the hand opened the box to a certain degree that generally won't fall down, it fell down due to completely different friction properties of the box. For the flip task, the policy succeeded 6 times out of 8 tries, demonstrating the generalization ability of our method.

VI. CONCLUSIONS

In this work, we enhance 3D tactile feature learning by proposing a novel canonical representation that aligns differently distributed tactile sensor readings, reduces the feature space, and increases the discriminability of each taxel within the same sensor. We also introduce a force-based self-supervised pretraining task to encourage the use of both spatial and force information. Real-world experiments using the pretrained encoder for downstream dexterous, fine-grained, contact-rich tasks demonstrate the effectiveness and robustness of our methods.

Limitations and Future Work. Our policy shows limited generalization when encountering objects with significantly different shapes and dynamics. Quick adaptation using tactile could be a direction for future work.

REFERENCES

[1] F. M. Li, M. X. Liu, Y. Zhang, and P. Carrington, “Freedom to choose: Understanding input modality preferences of people with upper-body motor impairments for activities of daily living,” in *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*, 2022, pp. 1–16.

[2] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, “3d diffusion policy,” *arXiv preprint arXiv:2403.03954*, 2024.

[3] C. Wang, H. Shi, W. Wang, R. Zhang, L. Fei-Fei, and C. K. Liu, “Dexcap: Scalable and portable mocap data collection system for dexterous manipulation,” *arXiv preprint arXiv:2403.07788*, 2024.

[4] I. Guzey, B. Evans, S. Chintala, and L. Pinto, “Dexterity from touch: Self-supervised pre-training of tactile representations with robotic play,” in *Conference on Robot Learning*. PMLR, 2023, pp. 3142–3166.

[5] R. S. Johansson, C. Häger, and L. Bäckström, “Somatosensory control of precision grip during unpredictable pulling loads: Iii. impairments during digital anesthesia,” *Experimental brain research*, vol. 89, pp. 204–213, 1992.

[6] R. Patel, R. Ouyang, B. Romero, and E. Adelson, “Digger finger: Gel-sight tactile sensor for object identification inside granular media,” in *Experimental Robotics: The 17th International Symposium*. Springer, 2021, pp. 105–115.

[7] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, and R. Calandra, “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.

[8] T. P. Tomo, A. Schmitz, W. K. Wong, H. Kristanto, S. Somlor, J. Hwang, L. Jamone, and S. Sugano, “Covering a robot fingertip with uskin: A soft electronic skin with distributed 3-axis force sensitive elements for robot hands,” *IEEE Robotics and Automation Letters*, vol. 3, no. 1, pp. 124–131, 2017.

[9] I. Guzey, Y. Dai, B. Evans, S. Chintala, and L. Pinto, “See to touch: Learning tactile dexterity through visual incentives,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 13 825–13 832.

[10] R. Bhirangi, A. DeFranco, J. Adkins, C. Majidi, A. Gupta, T. Hellebrekers, and V. Kumar, “All the feels: A dexterous hand with large-area tactile sensing,” *IEEE Robotics and Automation Letters*, 2023.

[11] W. Hu, B. Huang, W. W. Lee, S. Yang, Y. Zheng, and Z. Li, “Dexterous in-hand manipulation of slender cylindrical objects through deep reinforcement learning with tactile sensing,” *arXiv preprint arXiv:2304.05141*, 2023.

[12] L. Yang, B. Huang, Q. Li, Y.-Y. Tsai, W. W. Lee, C. Song, and J. Pan, “Tacgnn: Learning tactile-based in-hand manipulation with a blind robot using hierarchical graph neural network,” *IEEE Robotics and Automation Letters*, vol. 8, no. 6, pp. 3605–3612, 2023.

[13] S. Funabashi, T. Isobe, S. Ogasa, T. Ogata, A. Schmitz, T. P. Tomo, and S. Sugano, “Stable in-grasp manipulation with a low-cost robot hand by using 3-axis tactile sensors with a cnn,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 9166–9173.

[14] F. Scarselli, M. Gori, A. C. Tsoi, M. Hagenbuchner, and G. Monfardini, “The graph neural network model,” *IEEE transactions on neural networks*, vol. 20, no. 1, pp. 61–80, 2008.

[15] S. Funabashi, T. Isobe, F. Hongyi, A. Hiramoto, A. Schmitz, S. Sugano, and T. Ogata, “Multi-fingered in-hand manipulation with various object properties using graph convolutional networks and distributed tactile sensors,” *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 2102–2109, 2022.

[16] H. Qi, B. Yi, S. Suresh, M. Lambeta, Y. Ma, R. Calandra, and J. Malik, “General in-hand object rotation with vision and touch,” in *Conference on Robot Learning*. PMLR, 2023, pp. 2549–2564.

[17] Y. Yuan, H. Che, Y. Qin, B. Huang, Z.-H. Yin, K.-W. Lee, Y. Wu, S.-C. Lim, and X. Wang, “Robot synesthesia: In-hand manipulation with visuotactile sensing,” in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6558–6565.

[18] Z.-H. Yin, B. Huang, Y. Qin, Q. Chen, and X. Wang, “Rotating without seeing: Towards in-hand dexterity through touch,” *arXiv preprint arXiv:2303.10880*, 2023.

[19] J. Yin, H. Qi, J. Malik, J. Pikul, M. Yim, and T. Hellebrekers, “Learning in-hand translation using tactile skin with shear and normal force sensing,” *arXiv preprint arXiv:2407.07885*, 2024.

[20] T. Lin, Y. Zhang, Q. Li, H. Qi, B. Yi, S. Levine, and J. Malik, “Learning visuotactile skills with two multifingered hands,” *arXiv preprint arXiv:2404.16823*, 2024.

[21] M. Zambelli, Y. Aytar, F. Visin, Y. Zhou, and R. Hadsell, “Learning rich touch representations through cross-modal self-supervision,” in *Conference on Robot Learning*. PMLR, 2021, pp. 1415–1425.

[22] S. Sundaram, P. Kellnhofer, Y. Li, J.-Y. Zhu, A. Torralba, and W. Matusik, “Learning the signatures of the human grasp using a scalable tactile glove,” *Nature*, vol. 569, no. 7758, pp. 698–702, 2019.

[23] C. Sferrazza, Y. Seo, H. Liu, Y. Lee, and P. Abbeel, “The power of the senses: Generalizable manipulation from vision and touch through masked multimodal learning,” *arXiv preprint arXiv:2311.00924*, 2023.

[24] Y. Li, J.-Y. Zhu, R. Tedrake, and A. Torralba, “Connecting touch and vision via cross-modal prediction,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 609–10 618.

[25] W. Yuan, S. Wang, S. Dong, and E. Adelson, “Connecting look and feel: Associating the visual and tactile properties of physical materials,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 5580–5588.

[26] S. Luo, W. Yuan, E. Adelson, A. G. Cohn, and R. Fuentes, “Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition,” in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 2722–2727.

[27] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar *et al.*, “Bootstrap your own latent a new approach to self-supervised learning,” in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 2020, pp. 21 271–21 284.

[28] V. Dave, F. Lygerakis, and E. Rückert, “Multimodal visual-tactile representation learning through self-supervised contrastive pre-training,” in *Proceedings/IEEE International Conference on Robotics and Automation*. Institute of Electrical and Electronics Engineers, 2024.

[29] K. Shaw, A. Agarwal, and D. Pathak, “Leap hand: Low-cost, efficient, and anthropomorphic hand for robot learning,” in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[30] G. Pavlakos, D. Shan, I. Radosavovic, A. Kanazawa, D. Fouhey, and J. Malik, “Reconstructing hands in 3d with transformers,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 9826–9836.

[31] A. Handa, K. Van Wyk, W. Yang, J. Liang, Y.-W. Chao, Q. Wan, S. Birchfield, N. Ratliff, and D. Fox, “Dexpilot: Vision-based teleoperation of dexterous robotic hand-arm system,” in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 9164–9170.

[32] Z. Hou, X. Liu, Y. Cen, Y. Dong, H. Yang, C. Wang, and J. Tang, “Graphmae: Self-supervised masked graph autoencoders,” in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2022, pp. 594–604.

[33] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, “Diffusion policy: Visuomotor policy learning via action diffusion,” in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[34] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[35] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby *et al.*, “Dinov2: Learning robust visual features without supervision,” *arXiv preprint arXiv:2304.07193*, 2023.

[36] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, “Learning fine-grained bimanual manipulation with low-cost hardware,” in *Proceedings of Robotics: Science and Systems (RSS)*, 2023.

[37] S. Yang, Y. Qin, R. Ding, X. Cheng, M. Liu, R. Yang, J. Li, S. Yi, and X. Wang, “Ace: A cross-platform and visual-exoskeletons system for low-cost dexterous teleoperation,” in *8th Annual Conference on Robot Learning*.

[38] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, “Graph attention networks,” in *International Conference on Learning Representations*, 2018.