ViTac-Tracing: Visual-Tactile Imitation Learning of Deformable Object Tracing

Yongqiang Zhao[1], Shan Luo[1]

***Abstract*— Deformable objects often appear in unstructured configurations. Tracing deformable objects helps bringing them into extended states and facilitating the downstream manipulation tasks. Due to the requirements for object-specific modeling or sim-to-real transfer, existing tracing methods either lack generalizability across different categories of deformable objects or struggle to complete tasks reliably in the real world. To address this, we propose a novel visual-tactile imitation learning method to achieve one-dimensional (1D) and two-dimensional (2D) deformable object tracing with a unified model. Our method is designed from both local and global perspectives based on visual and tactile sensing. Locally, we introduce a weighted loss that emphasizes actions maintaining contact near the center of the tactile image, improving fine-grained adjustment. Globally, we propose a tracing task loss that helps the policy to regulate task progression. On the hardware side, to compensate for the limited features extracted from visual information, we integrate tactile sensing into a low-cost teleoperation system considering both the teleoperator and the robot. Extensive ablation and comparative experiments on diverse 1D and 2D deformable objects demonstrate the effectiveness of our approach, achieving an average success rate of 80% on seen objects and 65% on unseen objects. Demos, code and datasets are available at `https://sites.google.com/view/vitac-tracing`.**

I. INTRODUCTION

Deformable object manipulation has gained increasing attention in recent years due to its numerous real-world applications, such as cable management, cloth handling, and assistive dressing [1]. A common characteristic of deformable objects is that they commonly appear in unstructured configurations, e.g. a folded-over shoelace or a crumpled towel, where their geometry and task-relevant features are difficult to observe directly (Fig. 1). Tracing transforms deformable objects into an extended configuration by following their edge with fingers from one end to the other [2], [3]. Existing methods typically rely on object-specific models or reinforcement learning in simulation with accurate deformable simulators and carefully shaped rewards, limiting their robustness and generality on real-world objects.

With the visual and tactile information, we propose an imitation learning method for deformable object tracing. This work marks the first step towards a unified model for tracing both 1D and 2D deformable objects. Our method is designed from both the local and global perspectives of the task. To improve fine-grained action adjustment and reduce the risk of dropping the object, we propose a local center loss that prioritizes actions that center the object in the tactile image. Since various deformable manipulation tasks require tracing stopping at vision-detected accurate locations, e.g., stopping when inserting a cable into a clip during cable routing [4], we further design a global task loss to regulate the task progression. Experiments validate the effectiveness of individual components and demonstrate the generalizability of our method, achieving an average success rate of 80% on seen objects and 65% on unseen objects.

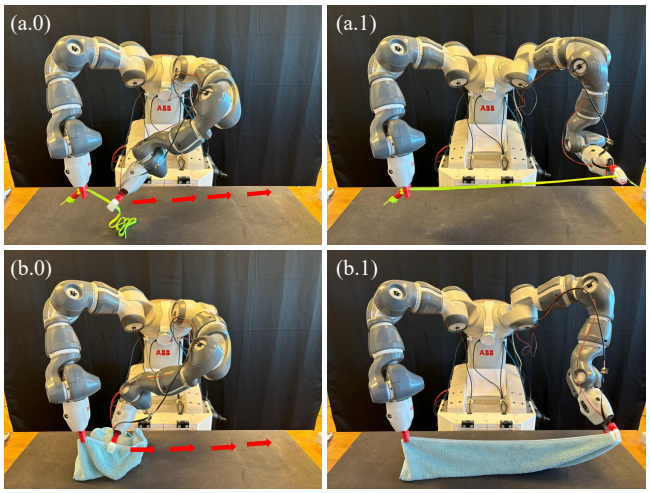

Fig. 1. (a) 1D deformable object tracing. (b) 2D deformable object tracing. Using our proposed method, an ABB YuMi traces the objects by sliding a gripper along the 1D deformable object or an edge of the 2D deformable object and transforming them from the unstructured configurations on the left to the extended states on the right.

In summary, the contributions of this work are as follows:

- We propose a novel visual-tactile imitation learning framework for deformable object tracing, enabling real robots to trace on diverse 1D and 2D deformable objects through a unified policy;
- We introduce a budget-efficient visual–tactile teleoperation system with multi-modal feedback to enrich perception of both the robot and the teleoperator;
- Extensive experiments validate the effectiveness of the individual components of our method and demonstrate the performance on seen objects as well as the generalizability to unseen objects.

II. METHODOLOGY

The proposed visual-tactile imitation learning framework is illustrated in Fig. 2.

To obtain expert demonstrations, we develop a visual–tactile teleoperation system on a dual-arm ABB YuMi. as shown in Fig. 3. One arm holds the deformable object, and the other performs tracing. A top-down RGB stereo camera

[1]Robot Perception Laboratory, Department of Engineering, King's College London, WC2R 2LS, United Kingdom.

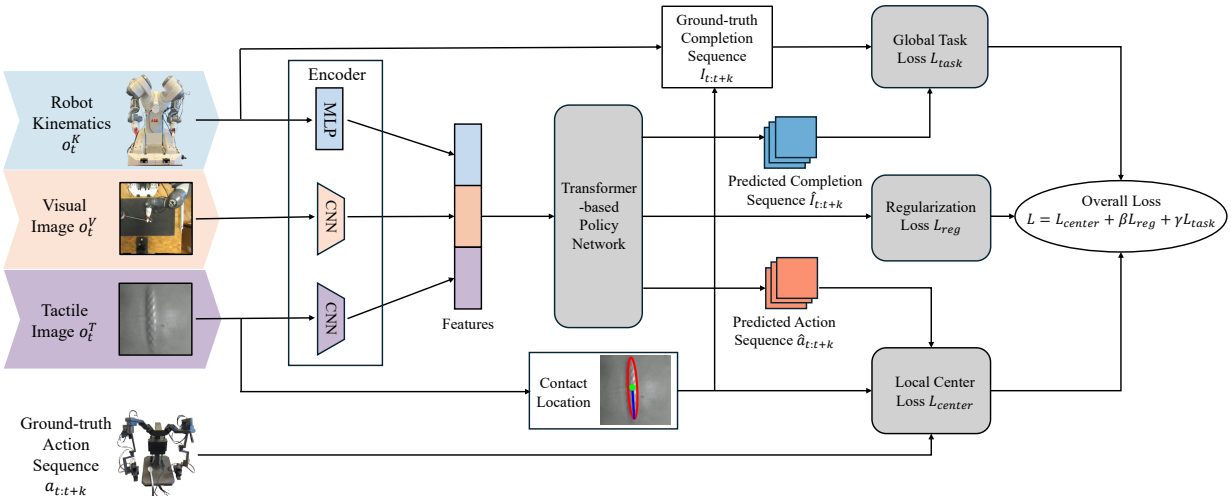

Fig. 2. Overview of the proposed tracing policy learning framework. The inputs include robot kinematics o_t^K, visual image o_t^V, and tactile image o_t^T collected from the follower robot, while the ground truth consists of action sequence $a_{t:t+k}$ recorded from the leader robot. Input features are first extracted using a Multilayer Perceptron (MLP) and Convolutional Neural Networks (CNNs). These features are then concatenated and fed into a Transformer-based policy network, which is trained using a combination of three loss functions: local center loss, global task loss, and regularization loss.

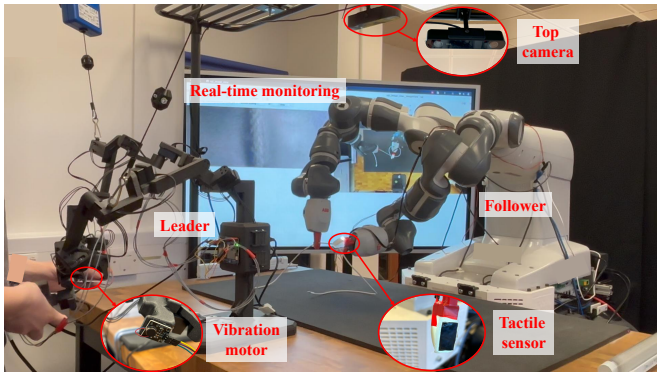

Fig. 3. Visual-tactile teleoperation system for collecting demonstrations. On the robot side, a top-view camera is installed, and a tactile sensor is mounted on the follower robot's gripper. On the teleoperator side, visual and tactile images are monitored in real time, and vibration motors are mounted on the leader robot's gripper.

provides a global view of the scene, while a compact vision-based tactile sensor integrated into the gripper provides high-resolution local contact feedback. Streaming both visual and tactile images to the human operator enables reliable, repeatable demonstrations across diverse deformable objects.

For policy learning, we adopt the Action Chunking Transformer (ACT) to predict short Cartesian action sequences from fused proprioceptive, visual, and tactile inputs. To better exploit tactile information, we introduce a local center loss that reweights imitation loss based on the distance between the contact patch and the tactile image center, biasing the policy toward actions that keep contact within the high-fidelity region. To regulate global task progress, we define a completion index from the tactile contact location relative to the fixed grasp point and train a second head with a global task loss to predict this index, helping the policy decide when to stop tracing.

TABLE I

EXPERIMENTAL RESULTS TO VALIDATE THE EFFECTIVENESS OF INDIVIDUAL COMPONENTS.

Methods	Success rate	Robot collision	Early stopping	Over-tracing	Object dropping
Joint Space	70.00%	2.50%	10.00%	5.00%	12.50%
w/o Vision	65.00%	10.0%	5.00%	20.0%	0.00%
w/o Tactile	60.00%	5.00%	12.50%	2.50%	20.00%
w/o Center Loss	65.00%	10.00%	2.5%	0.00%	22.50%
w/o Task Loss	67.50%	7.50%	7.50%	17.50%	0.00%
Ours	**80.00%**	5.00%	5.00%	7.50%	2.50%

III. EXPERIMENTAL RESULTS

We evaluate ViTac-Tracing on multiple 1D and 2D objects and on unseen test objects, as Tab. I shows. The unified policy achieves high success rates on seen objects and maintains robust performance on unseen geometries, with most failures due to termination rather than contact loss. These results show that combining visual–tactile imitation learning with tailored local and global objectives enables a single policy to perform contact-rich deformable object tracing, moving toward general visuomotor control for deformable manipulation.

IV. CONCLUSIONS

In this work, we present an Imitation Learning (IL)-based approach for deformable object tracing using a single unified policy with both visual and tactile sensing, implemented in a teleoperation system with multi-modal feedback. To capture both local corrections and global task progression, we introduce a center loss and a task loss. Trained on demonstrations from four 1D and 2D deformable objects, the model achieves an overall success rate of 80%, and ablation studies validate the contribution of each component. Tests on two unseen objects further demonstrate generalizability, achieving a 65% success rate.

REFERENCES

[1] J. Zhu, A. Cherubini, C. Dune, D. Navarro-Alarcon, F. Alambeigi, D. Berenson, F. Ficuciello, K. Harada, J. Kober, X. Li *et al.*, "Challenges and outlook in robotic manipulation of deformable objects," *IEEE Robotics & Automation Magazine*, vol. 29, no. 3, pp. 67–77, 2022.

[2] Y. She, S. Wang, S. Dong, N. Sunil, A. Rodriguez, and E. Adelson, "Cable manipulation with a tactile-reactive gripper," *The International Journal of Robotics Research*, vol. 40, no. 12-14, pp. 1385–1401, 2021.

[3] N. Sunil, S. Wang, Y. She, E. Adelson, and A. R. Garcia, "Visuotactile affordances for cloth manipulation with local control," in *Conference on Robot Learning*. PMLR, 2023, pp. 1596–1606.

[4] J. Luo, C. Xu, X. Geng, G. Feng, K. Fang, L. Tan, S. Schaal, and S. Levine, "Multistage cable routing through hierarchical imitation learning," *IEEE Transactions on Robotics*, vol. 40, pp. 1476–1491, 2024.