ViTacGen: Robotic Pushing with Vision-to-Touch Generation

Zhiyuan Wu[1], Yijiong Lin[2], Yongqiang Zhao[1], Xuyang Zhang[1], Zhuo Chen[1], Nathan Lepora[2], Shan Luo[1]

***Abstract*— We propose ViTacGen, a novel robot manipulation framework designed for visual robotic pushing with vision-to-touch generation in reinforcement learning to eliminate the reliance on high-resolution real tactile sensors, enabling effective zero-shot deployment on visual-only robotic systems. Specifically, ViTacGen consists of an encoder-decoder vision-to-touch generation network that generates contact depth images, a standardized tactile representation, directly from visual image sequence, followed by a reinforcement learning policy that fuses visual-tactile data with contrastive learning based on visual and generated tactile observations. We validate the effectiveness of our approach in both simulation and real world experiments, demonstrating its superior performance and achieving a success rate of up to 86%.**

I. Introduction

Robotic pushing is a fundamental manipulation task that requires tactile feedback to capture subtle contact forces and dynamics between the end-effector and the object [1], [2]. However, real tactile sensors often face hardware limitations such as high costs and fragility [3], [4], and deployment challenges involving calibration and variations between different sensors [5], while vision-only policies struggle with satisfactory performance [6].

Humans, in contrast, have the remarkable ability to infer tactile states from visual information [7]. Inspired by this, we propose ViTacGen, a novel robot manipulation framework designed for visual robotic pushing with vision-to-touch generation in reinforcement learning to eliminate the reliance on high-resolution real tactile sensors. Specifically, our method introduces two key components: an encoder-decoder vision-to-touch generation network **VT-Gen** that synthesizes tactile contact depth images [5], a standardized tactile representation, directly from visual image sequence, and an RL network **VT-Con** that learns robust policies through feature fusion and contrastive learning based on visual and generated tactile observations. Our ViTacGen is trained in simulation by two steps: (1) training VT-Gen to generate tactile contact depth images on paired visual and tactile data collected from expert trajectories by a pre-trained RL network [8] with visual and tactile observations, and (2) incorporating the frozen VT-Gen to train VT-Con based on visual and generated tactile observations. ViTacGen can perform zero-shot deployment on visual-only robotic systems, which eliminates the reliance on high-resolution real tactile sensors but still captures subtle dynamics and interactions between the end-effector and the object. Additionally, we solve the problem of manufacturing variations across different tactile sensors using contact depth maps as a standardized tactile representation. We conduct extensive experiments in both simulation and real world settings to validate the effectiveness of the proposed framework. The results demonstrate that ViTacGen achieves superior performance, achieving a success rate of up to 86% and outperforming baseline methods.

[1]King's College London
[2]University of Bristol

II. Methodology

As illustrated in Fig. 1, our proposed ViTacGen framework comprises two key components: **VT-Gen** for encoder-decoder vision-to-touch generation and **VT-Con** to learn robust policies through feature fusion and contrastive learning based on visual and generated tactile observations.

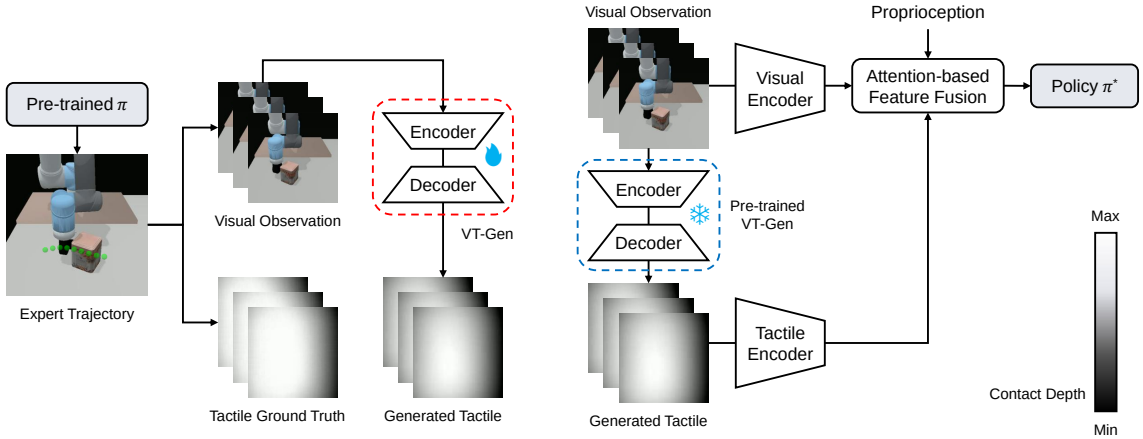
A. VTGen: Encoder-decoder Vision-to-touch Generation

Humans have the remarkable ability to infer tactile states from visual information [7]. Inspired by this, we propose a VT-Gen for encoder-decoder vision-to-touch generation to infer the tactile sensation of the contact region directly from a image sequence of visual input. As depicted in Fig. 1 (a), given a visual image sequence $\mathcal{V} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\}$, it is processed by VT-Gen to generate a current tactile contact depth image $\boldsymbol{c}^{gen}$, which is then repeated N times to form a generated tactile contact depth sequence $\mathcal{C}$ with N frames to align with the length of the visual sequence, serving as the tactile observation for VT-Con.

B. VT-Con: RL with Visual-Tactile Contrastive Learning

1) Feature Extraction: As shown in Fig. 1 (b), we consider the visual sequence $\mathcal{V} = \{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\}$ for visual observation, and the generated tactile sequence $\mathcal{C} = \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_N\}$ for tactile observation, where N denotes the number of frames in the visual and tactile sequences, respectively, empirically set to 3 for capturing the velocity and acceleration information. Each element $\boldsymbol{v}_i$ or $\boldsymbol{c}_i$ is represented as a tensor resized to the same dimension $\mathbb{R}^{H\times W\times C}$, where H, W, and C correspond to the height, width, and channel count of the frames. To extract features from the visual and tactile inputs, we first concatenate $\mathcal{V}$ and $\mathcal{C}$ respectively along their channel dimensions. Then, we employ two structurally identical Convolutional Neural Networks (CNNs) $\mathcal{E}^v$ and $\mathcal{E}^c$ as encoders for the visual and tactile modalities, resulting in a visual feature map $\boldsymbol{f}^v$ and a tactile feature map $\boldsymbol{f}^c$.

2) Contrastive Learning between Vision and Touch: In order to more efficiently fuse the encoded visual and tactile features $\boldsymbol{f}^v$ and $\boldsymbol{f}^c$, we optimize $\mathcal{E}^v$ and $\mathcal{E}^c$ with Momentum Contrast (MoCo) [9] for contrastive learning between vision and touch, inspired by [8].



(a) **VT-Gen**: Encoder-decoder Vision-to-Touch generation (b) **VT-Con**: Reinforcement Learning with Vision-to-Touch Generation

Fig. 1. The workflow of our proposed ViTacGen comprises two components: a VT-Gen for vision-to-touch generation, and a VT-Con for reinforcement learning on visual and generated tactile contact depth images with contrastive learning.

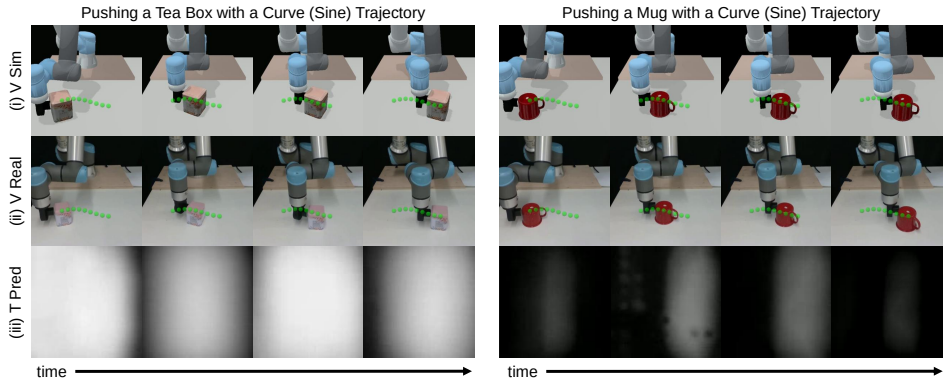


Fig. 2. Demonstration of ViTacGen in real world scenarios, where we present a tea box and a mug with a curve (sine) trajectories. The trajectories are visualized in simulation and projected onto both simulated and real world environments to demonstrate the system's performance.

3) Attention-based Visual-Tactile Feature Fusion: We introduce an attention-based visual-tactile feature fusion operation [10], to integrate $\boldsymbol{f}^v$ and $\boldsymbol{f}^c$, which can be formulated as follows:

$$\boldsymbol{f}^{fuse} = \boldsymbol{C}[\mathcal{A}_1^{cm}(\boldsymbol{f}^v, \boldsymbol{f}^c), \ldots, \mathcal{A}_h^{cm}(\boldsymbol{f}^v, \boldsymbol{f}^c)]\boldsymbol{w}_0, \quad (1)$$

where $\boldsymbol{f}^{fuse}$ refers to the fused feature map. Along with the fused feature map $\boldsymbol{f}^{fuse}$, the flattened and MLP-processed proprioception, *i.e.*, robot's TCP coordinates, are concatenated to form the complete observation vector for robot reinforcement learning and can be compatible with various RL algorithms, such as SAC and PPO.

III. Experiments

To validate the effectiveness of ViTacGen, we conducted comprehensive evaluations in both simulated and real world environments. Our simulation experiments were performed using Tactile Gym 2 [5]. For our VT-Gen, we collect visual and tactile data from 1,000 manipulation sequences and train our model for 200 epochs with a batch size of 64, splitting the data in a ratio of 7:2:1 for training, validation, and testing. For VT-Con, we employ the Soft Actor-Critic (SAC) [11] algorithm from Stable Baselines 3 [12] as our RL backbone, with a max episode length of 350. We train our model for 1,000,000 time-steps with a batch size of 64 and a buffer size of 20,000 for experience replay. The success termination criterion is defined as the distance between the object center and the goal center being less than 2.5 cm. For optimizer we employ Adam [13], with a learning rate of 1e-4, and an epsilon value of 1e-8. We provide qualitative results in Fig. 2. More results are available in our demo videos.

IV. Conclusion

This paper presents ViTacGen, a novel robot manipulation framework designed for visual robotic pushing with vision-to-touch generation in reinforcement learning to eliminate the reliance on high-resolution real tactile sensors, inspired by human's remarkable ability of predicting tactile states from vision to optimize manipulation. Specifically, we introduce an encoder-decoder vision-to-touch generation network that generates contact depth images, a standardized tactile representation, directly from visual image sequence, followed by a reinforcement learning policy that fuses visual-tactile data with contrastive learning based on visual and generated tactile observations. ViTacGen enables zero-shot deployment on visual-only robotic systems. We conduct extensive evaluation in both simulated and real world environments, demonstrating the effectiveness of our proposed methods.

REFERENCES

[1] B. Deng, Y. Lin, M. Yang, and N. F. Lepora, "Coarse-to-fine robotic pushing using touch, vision and proprioception," *IEEE Robotics and Automation Letters*, 2024.

[2] K.-T. Yu, M. Bauza, N. Fazeli, and A. Rodriguez, "More than a million ways to be pushed. a high-fidelity experimental dataset of planar pushing," in *2016 IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE, 2016, pp. 30–37.

[3] W. Yuan, S. Dong, and E. H. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[4] E. Donlon, S. Dong, M. Liu, J. Li, E. Adelson, and A. Rodriguez, "Gelslim: A high-resolution, compact, robust, and calibrated tactile-sensing finger," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1927–1934.

[5] Y. Lin, J. Lloyd, A. Church, and N. F. Lepora, "Tactile gym 2.0: Sim-to-real deep reinforcement learning for comparing low-cost high-resolution robot touch," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 10 754–10 761, 2022.

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

[7] F. N. Newell, A. T. Woods, M. Mernagh, and H. H. Bülthoff, "Visual, haptic and crossmodal recognition of scenes," *Experimental Brain Research*, vol. 161, no. 2, pp. 233–242, 2005.

[8] F. Lygerakis, V. Dave, and E. Rueckert, "M2curl: Sample-efficient multimodal reinforcement learning via self-supervised representation learning for robotic manipulation," *arXiv preprint arXiv:2401.17032*, 2024.

[9] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2020.

[11] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.

[12] A. Raffin, A. Hill, A. Gleave, A. Kanervisto, M. Ernestus, and N. Dormann, "Stable-baselines3: Reliable reinforcement learning implementations," *Journal of Machine Learning Research*, vol. 22, no. 268, pp. 1–8, 2021.

[13] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.