Unsupervised Adversarial Domain Adaptation for Sim-to-Real Transfer of Tactile Manipulation Skills

Xingshuo Jing[1], Yongqiang Zhao[1], Jiaqi Jiang[2], Boyi Duan[1], Kun Qian[1], Shan Luo[2]

***Abstract*— Transferring optical tactile skills learned from simulated environments to the real world benefits many robotic tactile applications, which can reduce the cost of data collection. However, the models purely trained on simulated data are often difficult to generalize well to the unseen real world due to the domain gap between the faultless training images and testing images with unpredictable manufacturing defects or natural wear. In this paper, we propose an Adaptively Correlation-attentive and Task-related Network (ACTNet) for tactile image transfer, a novel unsupervised adversarial domain adaptation method to narrow the domain gap for pixel-level tactile perception tasks. An adaptively correlative attention mechanism is introduced to improve the generator, which is capable of leveraging global information and focusing on salient regions. We also construct a task-related constraint loss based on the robotic insert-and-pullout tactile manipulation task to facilitate the zero-shot sim-to-real transfer of the reinforcement learning-based policy.**

I. INTRODUCTION

Tactile sensing is essential for tactile-based robotic manipulation tasks [1], [2]. Such tactile sensing is vital in situations of poor illumination or heavy occlusions, where the visual appearance of objects can be brittle. Since the soft elastomer in real-world tactile sensors is fragile when being heavily used in data collection, learning the pixel-level perception tasks in a simulator will be an alternative solution, which can avoid the high cost of collecting real-world data. However, the imperfections of reality such as the scratches and other object deformation, which are representative of characterizing objects via touch sensing, can hardly be precisely reflected in the simulation. Due to the significant reality gap, the generalization performance of the model trained in a simulator will be greatly reduced when it is deployed in a real-world environment. In our previous work [3], as shown in Fig. 1, we propose a novel tactile-motor policy learning method to generalize tubular object manipulation skills from simulation to reality. Although domain-randomization-based approaches [4] try to reduce the tactile domain gap by adding Gaussian noise or texture mapping, it is only able to cover a small fraction of the real-world distribution. The domain-adaptation-based approaches [2], [5], [6] with more flexibility learn the probability distribution between the simulated and real-world domains by the adversarial training of the generator and discriminator. However, the gentle press will produce less distinctive features, which poses a challenge to the generation network purely based on Convolutional Neural Network (CNN). Besides, the traditional image adaptation methods tend to treat a tactile image as a whole, without attention to the salient region which reflects the effective touching area. These issues will deteriorate the sim-to-real generalization performance of tactile-based policies.

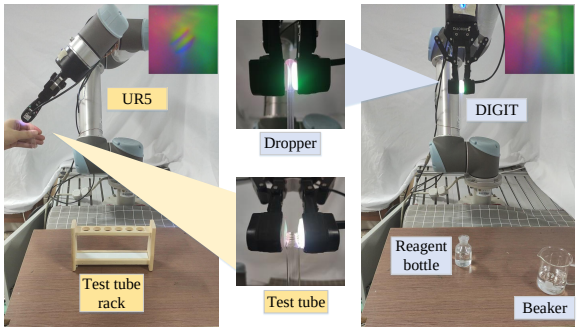

Fig. 1. Tactile-guided human-robot collaborative tube placing (left) and robotic pipetting (right).

We propose ACTNet to narrow the domain gap for pixel-level tactile tasks, which can generate high-quality tactile images. We optimize the design of network modules by introducing an Adaptively Correlative Attention (ACA) mechanism and construct a task-related loss by introducing an additional structural edge-consistency constraint. Compared with previous methods, this method can leverage global information and focus on salient regions, and enable the generated image to contain the necessary structural details from the real-world domain image. The proposed method is further applied to a sim-to-real robotic insert-and-pullout tactile manipulation task that is free from visual perception due to poor illumination.

II. METHODOLOGY

A. Transfer of tactile images

Network design: We introduce a novel generator network, which augments the decoder efficiency of U-Net with the ACA module built from multi-head transformers. The ACA block is mainly constructed as an adaptively correlative attention operation of the skip connection based on the attention given to a high-level feature map. The spatial attentive feature, the cross-correlative feature and the self-correlative feature are combined by three adaptive trainable weights to generate the final output of the ACA module. We consider this design mainly due to the reason that ACA possesses a greater ability to capture global context information and thus focuses on regions with salient features and adapts to the spatial variation of the touching region in tactile images.

[1]School of Automation, Southeast University and the Key Laboratory of Measurement and Control of CSE, Ministry of Education, No.2, Sipailou, Nanjing 210096, China. [2]Department of Engineering, King's College London, London, WC2R 2LS, United Kingdom. E-mail: kqian@seu.edu.cn, shan.luo@kcl.ac.uk. This work was supported by the Jiangsu Province Natural Science Foundation (No.BK20201264), Zhejiang Lab (No.2022NB0AB02), and the National Natural Science Foundation of China (No. 61573101).

It is mainly owing to the reason that these skip connections lack the semantic richness that can be found deeper in the network if they insure to keep high-resolution information. This design is to turn off uncorrelated or noisy areas from the skip connection features. In addition, we follow [7] as our discriminator. Edge features are also used to represent the structural information of tactile images. We train the BDCN [8] edge detection model to extract the edge feature for the task-related loss.

Objective function: The LSGAN [7] is applied as the texture-based adversarial loss $L_{texture}$. We follow [9] to compute the cycle-consistency loss L_{cycle} and the mapping-consistency loss L_{map}. Besides, a task-related constraint loss L_{task} is proposed to make the structural fineness in the sim-to-real image generation more aligned for a tactile task. For the background region, we ignore the structural details of generated images and only consider the constraint of color similarity. In particular, for the foreground region, attention is paid to the texture information directly reflected by brightness and contrast, and the edge structural information within the area, such as scratches, spikes, or lines. This is essential for the tactile manipulation task. The overall loss $\mathscr{L}_{overall}$ consists of four parts, i.e., $L_{texture}$, L_{cycle}, L_{map}, and L_{task}. Therefore, the optimal target models ($G^*_{S\to R}$ and $G^*_{R\to S}$) can be obtained by alternating iterative training.

B. Robotic insert-and-pullout tactile manipulation policy sim-to-real learning

We further apply our proposed ACTNet to the robotic insert-and-pullout tactile manipulation task for sim-to-real policy transfer. We use the pose of the end-effector, the 1D deflection angle θ obtained by tactile sensing, together with the wrench reading to model the state of the tasks. The wrench reading of the robot is used to determine whether there are problems such as failure to insert and jamming. The robotic action space is composed of 6D pose increment of the end-effector. Besides, the robot insert and pullout tasks all consist of a sequence of goal reach and pose control actions. The pose control is to ensure that there are no excess collisions or contact between the manipulated tubes and the hole during the insert and pullout process.

The pose θ describes the angle at which the manipulated tube is deflected from the tactile image's vertical main axis. We design an object in-hand pose estimation network, which takes ConvNeXt as the backbone and is added with a fully connected layer for the angle estimation. By converting synthetic tactile images via the sim-to-real ACTNet generator, we obtain the generated pseudo-reality tactile images, which are used to train the estimator. As a result, the superior performance of the ACTNet generator guarantees the RL policy trained in the simulator adapts well to the real-world robot, without any fine-tuning using real-world data.

III. Estimation and Conclusion

We qualitatively demonstrate the results of tactile image adaptation in the sim-to-real direction, as shown in Fig. 2. The first three columns are the transfer results of GelSight sensor, and the last three columns are the results of DIGIT

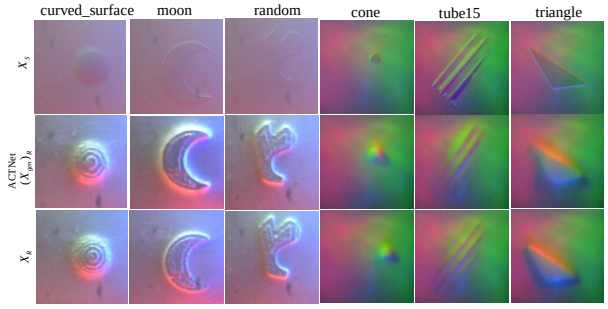

Fig. 2. Tactile image generation results in the sim-to-real direction.

sensor. The generated image by ACTNet generator is greatly approximate to the real-world image, which has a high quality and abundant structure information. Compared with the FID and PSNR metrics between raw real-world images and raw synthetic images, our method realizes an performance improvement of 133.76 and 8.58 respectively. Moreover, for the robotic insert-and-pullout tactile manipulation task, we construct two real-world scenarios, i.e., human-robot collaborative placing for tubes and robotic pipetting for droppers. The task success rate of seen, unseen and all test tubes in the first experiment reach 83.33%, 70.00%, and 78.00%, respectively. The task success rate of droppers in the second experiment reach up to 90.00%.

The estimation results demonstrate that our proposed method has a great capability of generating texture and edge structure of tactile images. Our method can effectively narrow the domain gap between optical tactile images in real and simulated environments, and further achieve the sim-to-real transfer of the robotic tubular object manipulation task. This contributes to training cross-domain transferable models for similar tasks in a simulated environment, independent of real-world image data and labels.

References

[1] S. Luo, J. Bimbo, R. Dahiya, and H. Liu, "Robotic tactile perception of object properties: A review," *Mechatronics*, vol. 48, pp. 54–67, 2017.

[2] T. Jianu, D. F. Gomes, and S. Luo, "Reducing tactile sim2real domain gaps via deep texture generation networks," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 8305–8311.

[3] Y. Zhao, X. Jing, K. Qian, D. F. Gomes, and S. Luo, "Skill generalization of tubular object manipulation with tactile sensing and sim2real learning," *Robotics and Autonomous Systems*, vol. 160, p. 104321, 2023.

[4] D. F. Gomes, P. Paoletti, and S. Luo, "Generation of gelsight tactile images for sim2real learning," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 4177–4184, 2021.

[5] X. Jing, K. Qian, X. Xu, J. Bai, and B. Zhou, "Domain adversarial transfer for cross-domain and task-constrained grasp pose detection," *Robotics and Autonomous Systems*, vol. 145, p. 103872, 2021.

[6] W. Chen, Y. Xu, Z. Chen, P. Zeng, R. Dang, R. Chen, and J. Xu, "Bidirectional sim-to-real transfer for gelsight tactile sensors with cyclegan," *IEEE Robotics and Automation Letters*, 2022.

[7] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2794–2802.

[8] J. He, S. Zhang, M. Yang, Y. Shan, and T. Huang, "Bi-directional cascade network for perceptual edge detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3828–3837.

[9] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.