# Feature-level Sim2Real Regression of Tactile Images for Robot Manipulation

Boyi Duan[1], Kun Qian[1*], Yongqiang Zhao[1], Dongyuan Zhang[1], Shan Luo[2]

***Abstract*—When the robot tactile-motor policy trained in the simulator is transferred to the real world, model's performance has a great potential for degradation due to the domain gap between simulated and real tactile images. Currently pixel-level domain adaptation methods for tactile images, extensions of the GAN-based image transfer algorithms, provide relatively general solutions. However, these GAN-based methods are characterized by high network structure complexity, unstable and costly training process because of the adversarial training relationship between the generator and discriminator. On the contrary, feature-level domain adaptation designed based on specific regression tasks can obtain more stable results with less cost. In this paper, we propose a feature-level unsupervised sim2real framework for tactile images, called STR-Net, to narrow the domain gap for feature-level tactile perception tasks.**

## I. Introduction

TACTILE sensing is essential for tactile-based robotic manipulation tasks [1]. Such tactile sensing is vital in situations of poor illumination or operating small objects with heavy occlusions by gripper. For those learning-based tactile-motor manipulation skills, most of them are trained in simulator in order to avoid great damage for robots. However, due to the significant reality gap between simulated and real tactile images, the generalization performance of the model trained in a simulator will be greatly reduced when it is deployed in a real-world environment. Therefore, the core of sim2real for tactile-motor policy is decreasing the discrepancy between simulated and real tactile images.

Many recently proposed works have demonstrated the effectiveness of pixel-level domain adaptation, a mainstream approach, for tactile images with GAN-based methods [2], [3]. However, while these pixel-level transfer represent a versatile approach suitable for tactile sensing, it necessitates the design of additional task-specific networks for different tactile image tasks, like classification, regression, or feature extraction. Specifically, subsequent classification/regression networks must use the output of the pixel-level transfer network as input, which can lead to error accumulation and thus degrade classification or regression performance. Also, instead of only collecting RGB images, mask images are needed as well [2], [3]. Furthermore, Generative Adversarial Networks(GAN) inherently exhibit training instability, contributing to increased structural complexity and training costs.

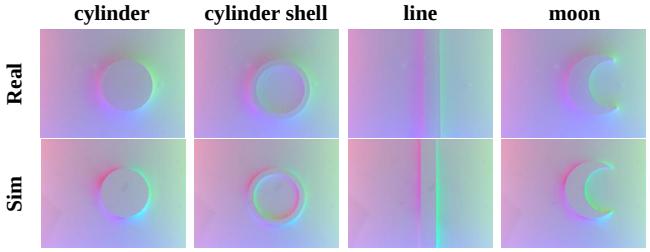


Fig. 1. Real Gelsight tactile image samples (top) and simulated tactile image samples by FOTS (bottom).

This study explores the possibility of achieving a faster and more robust tactile image sim2real transfer directly at the feature level for regression tasks. Our defined task is to estimate the pose of cylindrical objects in hand inspired by [1]. The proposed network, STR-Net(Siamese Tactile Regression Network), is trained using unpaired tactile images collected in simulator and reality. The simulated images are automated labeled by simulator while the real images are without any labels.

## II. FOTS simulator for tactile image simulation

In this paper, we use our proposed FOTS (Fast Optical Tactile Simulator) to generate simulated tactile images. Specifically, we present a more robust method that utilizes a multilayer perceptron to simulate the optical system of sensors by mapping contact gradients to illumination intensities. Corresponding planar shadows are then generated for each light source. Image samples are shown in Fig.1

With regard to lighting simulation, our method uses real-world data to simulate the intrinsic noise of the real sensors, without requiring any prior knowledge of the sensors' hardware layout. It also allows for greater generalization and robustness in the simulation, with an additional advantage of improved learning performance through the use of batch normalization. For shadow simulation, we use the planar hard shadow generation method [4] to simulate the shadow from each light source separately, which eliminates platform dependencies and can be easily implemented on other sensors.

Experimental results demonstrate that FOTS outperforms other methods in terms of image generation quality and rendering speed, achieving 28.6 fps for optical simulation on an Intel Core i7-9700k 8-Core Processor CPU.

*Corresponding author

[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. Email: kqian@seu.edu.cn. 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).

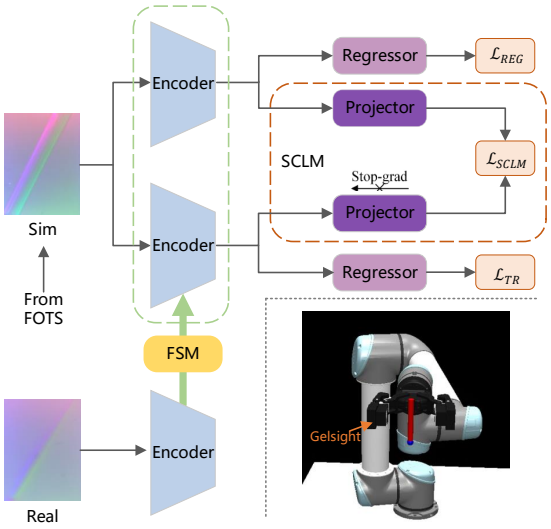


Fig. 2. An overview of STR-Net during training and the task scenario.

## III. Unsupervised Feature-Level Domain Adaptation Method

**Network architecture during training phase.** In the training phase, STR-Net consists of three main parts: a Siamese Encoder, a Feature Stylization Mechanism (FSM) and a Style Consistency Learning Module(SCLM). An overview of STR-Net is shown in Fig.2. The manipulation task scenario, pose regression of in-hand cylindrical objects, is also displayed in the lower right corner.

When training, the siamese network randomly takes one batch of simulated tactile images as input while the third encoder takes one batch of unpaired real tactile images as input. FSM is inspired by [5], which is embedded in the second encoder to realize feature-level domain adaptation. Finally, the siamese encoders will output the simulated features and real-stylized simulated features, respectively. Inspired by [6], we design a contrastive learning loss $\mathcal{L}_{SCLM}$ in image-level for visual representation learning without negative tactile sample pairs behind the encoder. In order to avoid the degenerated solution and yielding a collapsed representation, two approaches can be implemented. Firstly, an asymmetric architecture for the siamese encoder is established, where one representation needs to pass through an additional non-linear projector, resulting in a different feature embedding. Besides, similar to [6], we adopt the stop-gradient operation on them.

Feeding the finally obtained simulated features into regressor, we can get the corresponding predicted results. MSE loss $\mathcal{L}_{REG}$ is performed for supervised loss in simulation domain.

Meanwhile, although FSM helps alleviating the style differences between the simulated and real domain, it may cause a loss of semantic content. Hence, we embed task-relevant loss $\mathcal{L}_{TR}$ to capture pose features related to real domain. To be specific, obtained real-stylized simulated features are also sent into the regressor. Finally, the total loss function of STR-Net consists of three parts, $\mathcal{L}_{SCLM}$, $\mathcal{L}_{REG}$ and $\mathcal{L}_{TR}$.

**Network architecture during testing phase.** Once the network is trained, we apply the STR-Net to unseen real tactile images. During testing, only a single encoder is included. The siamese encoder becomes a single branch and all the FSM in CNN are deactivated, replaced by the original BN layer. Finally, the features from the encoder are fed into the regressor to get the predicted in-hand pose.

## IV. Results and Conclusion

Following previous work [7], Mean Absolute Error (MAE) and Mean Square Error (MSE), which are widely applied in regression tasks, are used as evaluation metrics to validate the proposed method. In this paper, their unit is radians.

RSD [7] and CycleGAN [8] are outstanding unsupervised domain adaptation methods in feature-level and pixel-level, respectively. We quantitatively analyzed and compared STR-Net with above algorithms in sim2real direction on our dataset. The results are as follows. RSD exhibits an MAE of 0.353 and an MSE of 0.413. For CycleGAN, employing a model trained on the simulated dataset to evaluate the transferred unseen real images results in an MAE of 0.353 and an MSE of 0.224. Compared with these domain adaptation methods, STR-Net presents a better performance, demonstrating higher accuracy and reliability with an MAE of 0.160 and an MSE of 0.088. Furthermore, while CycleGAN's training duration extends beyond 8 hours, STR-Net requires less than 3 hours, highlighting its efficiency. In addition, the images in simulator and reality are collected by different grasping forces, which means the size of the foreground area is not constant. Thus our method is applicable to grasping columnar object in different sizes with different forces.

The results demonstrate that our proposed method has effectiveness for cross-domain tactile images regression task. STR-Net brings a new idea for narrowing the domain gap between optical tactile images in real and simulated environments. Compared with pixel-level domain adaptation methods, our approach needs less training cost and has better task-specific performance. In the future, we will try to combine this method with robot skills based on reinforcement learning algorithms to achieve sim2real transfer of tactile-motor policies.

## References

[1] 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.

[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] X. Jing, K. Qian, T. Jianu, and S. Luo, "Unsupervised adversarial domain adaptation for sim-to-real transfer of tactile images," *IEEE Transactions on Instrumentation and Measurement*, 2023.

[4] A. Woo, P. Poulin, and A. Fournier, "A survey of shadow algorithms," *IEEE Computer Graphics and Applications*, vol. 10, no. 6, pp. 13–32, 1990.

[5] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 1501–1510.

[6] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 15 750–15 758.

[7] X. Chen, S. Wang, J. Wang, and M. Long, "Representation subspace distance for domain adaptation regression." in *ICML*, 2021, pp. 1749–1759.

[8] 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.