Attention for Robot Touch: Tactile Saliency Prediction for Robust Sim-to-Real Tactile Control

Yijiong Lin, Mauro Comi, Alex Church, Dandan Zhang, Nathan F. Lepora
Department of Engineering Mathematics and Bristol Robotics Laboratory
University of Bristol, Bristol BS8 1UB, U.K.
{yijiong.lin, n.lepora}@bristol.ac.uk

Abstract

To improve the robustness of tactile robot control in unstructured environments, we propose and study a new concept: *tactile saliency* for robot touch, inspired by the human touch attention mechanism from neuroscience and the visual saliency prediction problem from computer vision. In analogy to visual saliency, this concept involves identifying key information in tactile images captured by a tactile sensor. While visual saliency datasets are commonly annotated by humans, manually labelling tactile images is challenging due to their counterintuitive patterns. To address this challenge, we propose a novel approach comprised of three interrelated networks: 1) a Contact Depth Network (ConDepNet), which generates a contact depth map to localize deformation in a real tactile image that contains target and noise features; 2) a Tactile Saliency Network (TacSalNet), which predicts a tactile saliency map to describe the target areas for an input contact depth map; 3) and a Tactile Noise Generator (TacNGen), which generates noise features to train the TacSalNet. Experimental results in contact pose estimation and edge-following in the presence of distractors showcase the accurate prediction of target features from real tactile images. Overall, our tactile saliency prediction approach gives robust sim-to-real tactile control in environments with unknown distractors. Videos for all the experiments are presented in: `https://sites.google.com/view/tactile-saliency/`.

1 Introduction

High-resolution tactile sensing is seeing greater use in robot manipulation as a complement to vision, due to its ability to reveal fine-grained details in local contact [1, 2]. Despite its potential, the research community has predominantly focused on tasks with idealized experimental conditions, disregarding the impact of noise and distractors [3, 4, 5]. This has resulted in a limited understanding of how to achieve robust tactile control in unstructured environments, where unexpected stimuli can impair the accuracy of controllers or policies relying on tactile sensing (as shown in Fig. 1b), making it difficult to achieve precise control in tactile-oriented tasks such as contour following and tactile exploration [6, 7, 8]. Therefore, it is crucial to develop a methodology that can effectively distinguish between target and noise features in tactile feedback, enabling robust tactile control in unstructured environments.

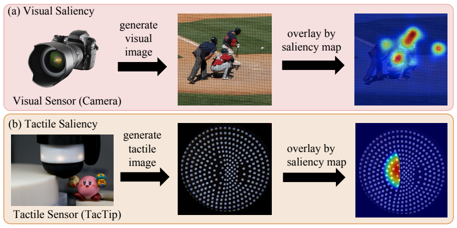

Figure 1: Visual saliency vs tactile saliency: (a) an example of visual saliency map (right) and its source visual image (mid) from SALICON dataset[10]; (b) an example of tactile saliency map (right) and its source real tactile image (mid) obtained by a TacTip (left) contacting a target edge (white cylinder) and a distractor (pink toy) in an edge-following task in a cluttered scene.

To better describe this problem, we propose a new concept for robot touch: *tactile saliency*. Analogous to *visual saliency* in computer vision, we define tactile saliency to describe the critical regions of

NeurIPS 2023 Workshop on Touch Processing: a new Sensing Modality for AI .

interest for a robot in a tactile image obtained by a tactile sensor. For example in Fig. 1b, a tactile saliency map can indicate a target feature of edge from a real tactile image captured by a tactile sensor during a contour-following task in an unstructured environment. However, collecting and labelling tactile saliency data presents unique challenges compared to visual saliency data. Unlike the latter, which is typically collected through human labelling using eye trackers or mouse clicks [9], it is challenging for humans to distinguish between target and noise features in a raw tactile image. Additionally, tactile sensors are often soft and delicate, making it impractical to collect a large amount of data through direct physical contact with various stimuli while avoiding damaging the sensors, particularly if the goal is to learn the joint distribution of target and noise features.

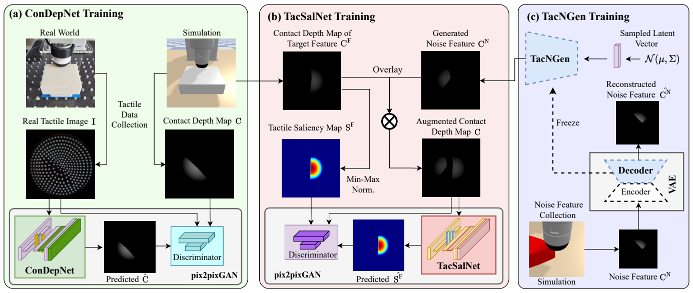

Figure 2: Overview of the 3-stage approach for tactile saliency prediction.

To address these challenges, here we propose a novel approach for tactile saliency prediction (Fig. 2) comprised of three interrelated networks: a) a Contact Depth Network (ConDepNet), which generates a contact depth map to localize deformation for an input real tactile image that contains target and noise features in a simplified format; b) a Tactile Saliency Network (TacSalNet), which predicts a tactile saliency map to describe the target areas for an input contact depth map; c) and a Tactile Noise Generator (TacNGen), which generates noise features to train the TacSalNet, making it more generalizable to unseen contact depth maps.

2 Methodology

2.1 Tactile Saliency Prediction

Given a real grey-scale tactile image $\mathrm{I} = \{\mathrm{I}_{ij} \in [0,1] \mid i \in \{1,...,w\}, j \in \{1,...,h\}\}$ of size $w \times h$ and a target feature type F, we define a *tactile saliency map* $\mathrm{S}^{\mathrm{F}} = \left\{\mathrm{s}^{\mathrm{F}}_{ij} \in [0,1] \mid i \in \{1,...,w\}, j \in \{1,...,h\}\right\}$ as a matrix of probabilities where $\mathrm{s}^{\mathrm{F}}_{ij}$ is the probability of pixel I_{ij} being part of the feature F (e.g. an edge) in the tactile image I. We define a mapping $\psi_{\mathrm{F}}(.) : \mathbb{R}^{w\times h} \to \mathbb{R}^{w\times h}$ as a function that maps a tactile image I to a tactile saliency map S^{F},

$$\psi_{\mathrm{F}}(\mathrm{I}) := \mathrm{S}^{\mathrm{F}}. \tag{1}$$

Our aim is to learn a tactile saliency prediction model to approximate ψ_{F}, which can be used to predict a tactile saliency map S^{F} representing the probabilities of target feature areas in image I. However, unlike visual saliency datasets, it is challenging for humans to construct a tactile saliency dataset by accurately labelling real tactile images with tactile saliency maps. This is due to the inherent nature of marker-based tactile images, which can present counterintuitive patterns that are difficult for humans to identify and separate the target features from the noise features. Thus, it is impractical to learn a tactile saliency model ψ_{F} for F that directly predicts S^{F} from I.

2.2 Contact Depth Prediction

While it is challenging to learn ψ_{F} to predict S^{F} directly from I, we can alternatively predict S^{F} from a simplified tactile image that only represents contact areas, giving a more straightforward and interpretable representation. Here, we define a simplified tactile image of I as a *contact depth map* $\mathrm{C} = \left\{\mathrm{c}_{ij} \in [0,1] \mid i \in \{1,\dots,w\}, j \in \{1,\dots,h\}\right\}$ where c_{ij} describes the contact depth level of the tactile skin in pixel I_{ij}. We define a mapping $\phi : \mathbb{R}^{w\times h} \to \mathbb{R}^{w\times h}$ as a function that maps a real tactile image I to a contact depth map C,

$$\phi(\mathrm{I}) := \mathrm{C}. \tag{2}$$

Our aim is to learn a contact depth prediction model G_{C} (referred to as *ConDepNet*) to approximate ϕ to predict a contact depth map C representing the contact areas in tactile image I. To collect a dataset $\mathcal{D}_{\Gamma,\Phi} = \left\{(\mathrm{I},\mathrm{C}) \mid (\mathrm{I} \in \Gamma, \mathrm{C} \in \Phi\right\}$ with auto-labelling, we leverage Tactile Gym [4], a simulator for tactile robotics based on rigid-body physics. We use this to generate simulated tactile images rendered as the contact depth map captured by a simulated tactile sensor when contacting stimuli. Following a general image-to-image translation approach in [12], we use pix2pix GAN to learn G_{C} with an objective in an adversarial training manner.

2.3 Predicting Tactile Saliency from Contact Depth

To achieve our primary aim of mapping saliency, we define a mapping $\delta_{\mathrm{F}} : \mathbb{R}^{w \times h} \to \mathbb{R}^{w \times h}$ as a function that maps a contact depth map C to a tactile saliency map S^{F} for a given target feature F:

$$\delta_{\mathrm{F}}(\mathrm{C}) := \mathrm{S}^{\mathrm{F}}. \tag{3}$$

In other words, we can solve Eq. 1 with a composite function of Eq. 2 and Eq. 3, because a contact depth map C preserves the contact information in a real tactile image I,

$$\delta_{\mathrm{F}}(\phi(\mathrm{I})) = \delta_{\mathrm{F}}(\mathrm{C}) = \mathrm{S}^{\mathrm{F}} = \psi_{\mathrm{F}}(\mathrm{I}). \tag{4}$$

Thus our aim reduces to learning a tactile saliency prediction model $G_{\mathrm{S}^{\mathrm{F}}}$ (referred to as *TacSalNet*) to approximate δ_{F} for a target feature F that predicts a saliency map S^{F} representing the target areas in C. Similar to ConDepNet, we also apply pix2pix GAN to learn $G_{\mathrm{S}^{\mathrm{F}}}$.

2.4 Tactile Noise Generator

We can use a simulated tactile sensor to interact with noise stimuli in simulation to collect a set of noise contact depth maps Φ^{N} for learning a TacSalNet. However, this is inefficient as various stimuli CAD models are required. Also, the noise patterns are unlike those we see in contact depth maps generated from a ConDepNet in real-world experiments. Hence, we propose a generative model τ to generate random noise, which we call Tactile Noise Generator (TacNGen), as shown in Fig. 2c.

3 Experiments and Results

3.1 TacSalNets with Different Tactile Noise Generation

First, we conduct an ablation study to compare the performance of our tactile saliency network (TacSalNet) when the noise is generated using our proposed noise generation model (TacNGen) versus a 2-dim multivariate Gaussian distribution. For conciseness, we will refer to the TaSalNet trained with noise generated from TacNGen as **TacSalNet-1**, and the one trained with Gaussian noise as **TacSalNet-2**. The evaluation process uses the same pose ranges as those used during data collection. We evaluate the performance of both TacSalNets with four commonly-used metrics in visual saliency research, and the testing results are reported in Table 2. The results demonstrate that the TacSalNet-1 outperforms the TacSalNet-2 in three metrics and one the same, indicating the effectiveness of our TacNGen in improving the performance of tactile saliency prediction compared to domain-agnostic stochastic noise generation. Additionally, we found that even though they are only trained with straight-edge features, the TacSalNet-1 can generalize well to unseen corner-edge features while the TacSalNet-2 tends to preserve the noise features. Thus, we only consider the TacSalNet-1 (TacNGen-based) for the evaluation of our whole framework in the contact pose prediction task and the edge-following task.

Table 1: MAEs of the trajectories from the ground truth for the edge-following task.

Objects	Pose-based PID		Image-based deep RL	
	w/o TSN-1	w/ TSN-1	w/o TSN-1	w/ TSN-1
Square	Fail	0.77mm	Fail	1.04mm
Foil	Fail	0.53mm	Fail	0.94mm
Flower	Fail	0.76mm	Fail	1.03mm
Volute	Fail	0.91mm	Fail	1.21mm

Table 2: Real-world validation with similarity metrics for TacSalNet-1 and TacSalNet-2 over 1k samples. Bold numbers denote the best results.

	AUC-J ↑	SIM ↑	CC ↑	NSS ↑
TacSalNet-1	**0.995**	**0.984**	**0.957**	**4.629**
TacSalNet-2	**0.995**	0.972	0.936	4.288

3.2 Tactile Saliency Prediction for Contact Pose Estimation

In our second experiment, we evaluate the performance of tactile saliency prediction in improving tactile pose prediction accuracy in the presence of distractors. To achieve this, we apply the TacSalNet-1 to a tactile PoseNet, which is a Convolutional Neural Network that predicts pose from tactile images [7], and investigate its impact on predicting the target edge 2D pose (position y and orientation R_z) from a tactile image when the TacTip statically contacts the target edge and distractors (see Fig. 3a). Here, we present three everyday objects considered as distractors varied in shapes, each distractor is fixed next to the target edge, with distances ranging between $[7, 14]$ mm.

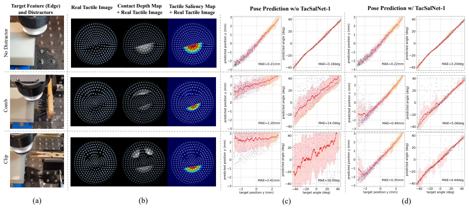

Figure 3: Real-world evaluation of tactile saliency prediction based on the tactile PoseNet prediction accuracy on the target feature (edge) distracted by various everyday objects.

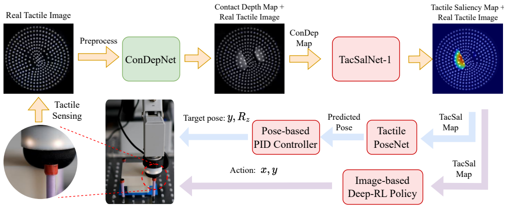

Figure 4: The diagram of two sim-to-real tactile control methods augmented with saliency prediction for the edge-following task with distractors. The networks in red are trained solely in simulation.

The experimental results show that the presence of the distractors significantly impairs the standard PoseNet performance, resulting in mean absolute errors (MAEs) of 1.84-2.41 mm and 14.0°-30.0° for predicting position y and orientation Rz respectively (Fig. 3c). However, when the TacSalNet-1 is applied, the PoseNet can maintain its accuracy with position y MAE of 0.26-0.40 mm and orientation R_z MAE of 4.44°-5.06° (Fig. 3d). To demonstrate the generalizability of our method, we randomly selected real tactile images with paired contact depth and saliency maps induced by static contact with different distractors (Fig. 3b). In the last pair of tactile images of Fig. 3b, we see the TacSalNet-1 accurately predicts the target edge shape even in the presence of unseen distracting contacts.

3.3 Tactile Saliency Prediction in the Edge Following Task

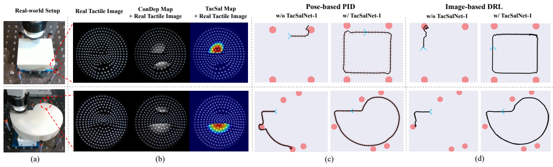

Figure 5: Evaluation of the proposed tactile-saliency-based control framework in the real-world 2D edge-following task.

In our third experiment, we investigate the performance of the tactile saliency prediction in improving the robustness of different tactile control methods during an edge-following task that involves distractors. Specifically, the robot is tasked to control its tactile sensor to follow along the edges of four target objects chosen to have distinct edge features. These target objects are surrounded by at least four distractors fixed next to their edges (Fig. 5a). In this task, we focus on the effect of the distractor as the edge changes in curvature, considering just one type (the bolt) given the effectiveness of the TacSalNet-1 for all distractors considered in Sec. 3.2.

We consider two distinct state-of-the-art tactile control methods: 1) tactile pose-based PID control [11], and 2) tactile image-based deep reinforcement learning [4]. Note that both the pose-based PID controller and the image-based deep-RL policy are learned solely in simulation with the contact depth map as input, and apply to the real world without fine-tuning. We augment them with TacSalNet-1 (pipeline in Fig. 4). The accuracies of the pose-based PID controller and the image-based deep-RL policy with tactile saliency prediction from 80 real-world tests (10 for each object with each method combination) are 0.5-0.9 mm and 0.94-1.21 mm respectively (Table 1), compared to an overall distance traveled of 300-520 mm. Overall, the saliency prediction gives a significant improvement in the robustness of sim-to-real tactile control.

4 Discussion and Future Work

In this work, we introduce the concept of *tactile saliency* inspired by visual saliency, to describe the target features from real tactile images captured by optical tactile sensors, for improving the accuracy and robustness of tactile control in the presence of distractors. To develop a generalizable tactile saliency prediction network, we propose a 3-stage approach. Notably, we only train the ConDepNet using real-world data, while the TacSalNet and TacNGen are trained solely in simulation.

Our proposed approach offers several key advantages. Firstly, it eliminates the need for labour-intensive human labelling, resulting in a more efficient and practical data collection process. Secondly, with the high fidelity of generated tactile noise, it enables accurate prediction of target features from contact depth maps in the presence of unseen noise, enhancing the accuracy of the contact pose prediction and the robustness of tactile control in unstructured environments. Thirdly, the TacSalNet can be seen as a simple plug-in module that can extend readily for various types of sim-to-real tactile control methods, such as the ones we considered here[4, 11].

References

[1] N. F. Lepora, “Soft biomimetic optical tactile sensing with the tactip: A review,” IEEE Sensors Journal, vol. 21, no. 19, pp. 2131–2143, 2021.

[2] A. C. Abad and A. Ranasinghe, “Visuotactile sensors with emphasis on gelsight sensor: A review,” IEEE Sensors Journal, vol. 20, no. 14, pp. 7628–7638, 2020.

[3] N. F. Lepora, A. Church, C. De Kerckhove, R. Hadsell, and J. Lloyd, “From pixels to percepts: Highly robust edge perception and contour following using deep learning and an optical tactile sensor,” IEEE Robotics and Automation Letters, vol. 4, no. 2, pp. 2101–2107, 2019.

[4] A. Church, J. Lloyd, N. F. Lepora et al., “Tactile sim-to-real policy transfer via real-to-sim image translation,” in 5th Annual Conference on Robot Learning, 2021.

[5] L. Pecyna, S. Dong, and S. Luo, “Visual-tactile multimodality for following deformable linear objects using reinforcement learning,” in 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2022, pp. 3987–3994.

[6] Z. Kappassov, J. A. C. Ramon, and V. Perdereau, “Tactile-based task definition through edge contact formation setpoints for object exploration and manipulation,” IEEE Robotics and Automation Letters, vol. 7, no. 2, pp. 5007–5014, 2022.

[7] N. F. Lepora and J. Lloyd, “Optimal deep learning for robot touch: Training accurate pose models of 3d surfaces and edges,” IEEE Robotics & Automation Magazine, vol. 27, no. 2, pp. 66–77, 2020.

[8] Y. Lin, J. Lloyd, A. Church, and N. Lepora, “Tactile gym 2.0: Sim-toreal deep reinforcement learning for comparing low-cost high-resolution robot touch,” vol. 7, no. 4. IEEE, August 2022, pp. 10 754–10 761.

[9] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, “Review of visual saliency detection with comprehensive information,” IEEE Transactions on circuits and Systems for Video Technology, vol. 29, no. 10, pp. 2941–2959, 2018.

[10] M. Jiang, S. Huang, J. Duan, and Q. Zhao, “Salicon: Saliency in context,” in IEEE conference on computer vision and pattern recognition, 2015, pp. 1072–1080.

[11] N. F. Lepora and J. Lloyd, “Pose-based tactile servoing: Controlled soft touch using deep learning,” IEEE Robotics & Automation Magazine, vol. 28, no. 4, pp. 43–55, 2021.

[12] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.