Optical Tactile Sim-to-Real Policy Transfer via Real-to-Sim Tactile Image Translation

Anonymous Author(s)
Affiliation
Address
email

Abstract: Simulation has recently become key for deep reinforcement learning to safely and efficiently acquire general and complex control policies from visual and proprioceptive inputs. Tactile information is not usually considered despite its direct relation to environment interaction. In this work, we present a suite of simulated environments tailored towards tactile robotics and reinforcement learning. A simple and fast method of simulating optical tactile sensors is provided, where high-resolution contact geometry is represented as depth images. Proximal Policy Optimisation (PPO) is used to learn successful policies across all considered tasks. A data-driven approach enables translation of the current state of a real tactile sensor to corresponding simulated depth images. This policy is implemented within a real-time control loop on a physical robot to demonstrate zero-shot sim-to-real policy transfer on several physically-interactive tasks requiring a sense of touch. Video results are available at *https://sites.google.com/view/tactile-sim-to-real*.

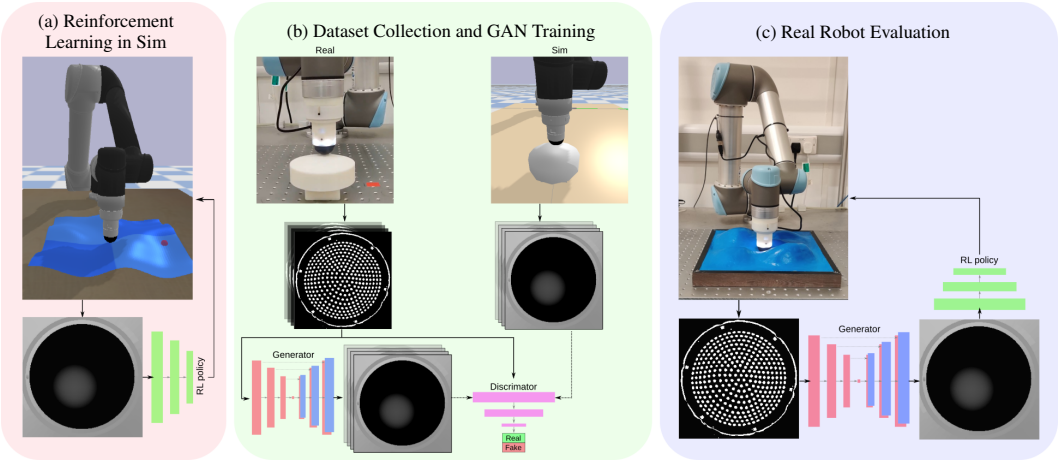

Figure 1: Overview of the proposed approach for sim-to-real transfer of learned tactile policies. a) Learn a policy in simulation directly from simulated tactile images. b) Train a GAN for translating between real and sim images using a dataset of image pairs collected by sampling a static environment with features similar to the environment used for reinforcement learning. c) Evaluate on the real robot by passing real images through the generator and then generated images through the RL policy.

1 Introduction

Learning algorithms have an innate appeal for robotics to enable general, complex behaviours that would be difficult to achieve with classic control methods. Large-scale data is often required for robot learning, and so physics-based simulation has a vital role to play for data collection. A common approach in learning-based robotics is to simulate data that would be impractical to collect in reality, learn control policies from this data, and then transfer learned skills to the physical system. Simulation also offers advantages such as avoiding damage during exploratory training, exploiting privileged information from the simulator, and use of open-sourcing for dissemination of research findings. However, physics engines necessarily approximate the real world to reduce computational costs, giving a 'sim-to-real gap' that impairs the performance of these policies applied to reality.

Submitted to the 5th Conference on Robot Learning (CoRL 2021). Do not distribute.

Reinforcement Learning (RL) robotics research is dominated by the use of proprioception and vision as the sensory inputs. Whilst this can lead to complex behaviours [1, 2, 3], other important sources of information have been underutilised. Specifically, humans use the sense of touch to accomplish complex manual tasks, utilising a granularity of detail unavailable to our other senses. Tactile data has several advantages as the main source of information for learning: it does not suffer from occlusion, particularly for fine manipulation; contact information can more detailed than visual images of an entire scene; and the observation space is constrained (e.g. to tactile images of markers), simplifying the translation of real to simulated data and, likewise, from simulated to real policies.

Here we make the following contributions aimed at bringing together reinforcement learning and tactile robotics: **1)** We provide an open-source suite of RL environments tailored to tactile robotics, utilising a simple and fast method of simulating a tactile sensor. **2)** We show that an RL method, Proximal Policy Optimisation (PPO) [4], acquires successful policies for all of these tasks. **3)** We demonstrate and validate a data-driven approach for zero-shot sim-to-real policy transfer via image translation between real and simulated tactile images on these tasks.

2 Related Work

Sim-to-Real Transfer: For computational efficiency, physics engines necessarily approximate the real-world dynamics, leading to a sim-to-real gap where dynamics and visuals differ between simulation and reality. This gap can make it difficult to transfer skills learned in simulation to the real world. Several methods have been proposed to mitigate this issue, namely *Domain Randomisation*, *Network Distillation* and *Domain Adaption* [3].

In this work we mainly use *Domain Adaption*, following James et al. [5] to train an image-conditioned generator network that translates between real and simulated images. When using vision, *Domain Randomisation* was necessary on simulated images to train a generator robust enough to generalise to real visual images. Here we exploit the more constrained image space of optical tactile sensors to train generator networks using a controlled tactile dataset. We perform some randomisations to make simulated tasks more difficult and RL policies more robust, aiming to keep within realistic conditions. Although privileged information from the simulator is used to construct a reward for improved learning, this information is not part of the observation in any task. As we perform zero-shot policy transfer from sim-to-real, *Network Distillation* is not required.

Tactile Sim-to-Real: Recently, several works have aimed to bridge between tactile simulation and reality. These split into two categories: simulating the physics of sensor deformation via Finite Element (FE) methods or replicating the captured sensor information via image rendering techniques. A concise overview of the sim-to-real approach applied to non-vision based sensors is given in [6].

FE methods have seen recent advancements in computational efficiency and accuracy. Narang et al. [7] simulated deformations of the BioTac sensor that, in combination with data-driven learning, can accurately regress the sensor output. Later, a 75-fold improvement in simulation speed was achieved with GPU-acceleration [6], with each simulated sensor output taking up to 5 seconds which is still impractical for sample-inefficient methods such as DRL. A similar approach was applied to a marker-based optical tactile sensor for regressing contact force fields [8, 9]. Later work [10] achieved zero-shot sim-to-real transfer of a complex policy learned in simulation via RL; however, task-specific approximation was needed for computational efficiency, e.g. assuming planar motion.

Most comparable to our work, Ding et al. [11] use an elastic deformation approach on the same optical tactile sensor (the TacTip) as used here. They focussed on supervised learning from simulated pin positions rather than tactile images, and with domain randomisation could accurately predict edge position and orientation for use in contour following. The authors noted that the simulation speed is restricted by Unity's collision detection module.

A computationally efficient approach is to render images from specific optical tactile sensors with rigid-body physics approximating the contact dynamics. Depth images from a GelSight tactile sensor were rendered using the Gazebo physics engine with additional Gaussian smoothing, light rendering and image calibration to help match real tactile images [12]. On a supervised classification task, a drop of 39% accuracy was found for direct sim-to-real transfer, reduced to 6.5% with texture augmentation as a form of domain randomisation. A similar approach is proposed in [13] with extended image-rendering techniques applicable to a range of GelSight-type sensors on more complex physically-interactive tasks; however, so far no sim-to-real quantitative results have been reported.

Our work here focuses on a high-resolution tactile simulation environment appropriate for RL of complex control policies. By relying on rigid-body approximations to the contact dynamics, our physics simulation is significantly faster than FE methods without task-specific approximations. Our simulated TacTip images use depth image-rendering techniques like those used for the GelSight [12, 13]. However, a key novelty is that our simulated depth images do not closely resemble the real TacTip images, because depth is non-trivially related to marker shear for the TacTip (and closely related to image shading for the GelSight). Instead, we close the sim-to-real gap by relying on a later domain adaption phase. Hence, we keep the simulation agnostic to the sensor transduction, which should facilitate future application of these methods to other types of tactile sensor.

3 Tactile Simulation

In this work, we utilise PyBullet's synthetic camera rendering to capture depth images within a virtual optical tactile sensor, based on the CAD files used to 3D print a real TacTip (see e.g. [14, Fig. 4]). When gathering tactile images, we take the difference between the current depth image and a reference depth image taken from when the sensor is not in contact. This difference produces a penetration depth map that generalises to arbitrary sensor shapes. Noise is removed from the image by zeroing values below a set tolerance of 10^{-4}, followed by a re-scaling from $[0, \mathrm{max_penetration}]$ to $[0, 255]$. An artificial border is also overlaid onto the tactile image to bring it closer to real tactile images and to provide a reference point that transforms with augmentations.

In order to achieve the computational efficiency necessary to generate data at large scales, we approximate the soft tip of the tactile sensor with rigid body physics simulation. We limit the contact stiffness and damping used during collision detection as this allows penetration of objects into the simulated sensor tip in an approximate manner to the deformation of the real tip.

We describe our simulation method in relation to the list of desiderata proposed by Wang et al. [13].

High Throughput: We use PyBullet's GPU rendering functionality to offer fast simulation. On a PC with an Nvidia 2080Ti, we can achieve up to 1000 fps when rendering single tactile images at 128×128-resolution. Multiple PyBullet physics engines can also run in parallel, so during training we used 10 vectorised environments to increase throughput.

Flexible: Whilst this study is predominantly based on the TacTip sensor, we do not attempt to simulate images accurate to any specific tactile sensor. Instead, we simulate only useful tactile features, relying on a later image-translation stage to map from real to simulated images. We expect similar results are possible with a broad range of other high-resolution optical tactile sensors, including sensors of the Gelsight type [15, 16, 17], the Soft-Bubble [18] and optical shear-based sensors [19].

Realistic: Instead of aiming to realistically simulate the physical properties of any specific sensor, which is both difficult and computationally expensive, we simulate only the desired properties of an idealised tactile sensor, currently focussing only on contact geometry. Our data-driven approach helps bridge the sim-to-real gap, so there is no need to generate synthetic images that match images from a real sensor to high precision.

Ease of use: The simulation suite and the sim-to-real translation tools will be open-sourced. The simulation only requires the commonly-used PyBullet, so our approach should have a low barrier of entry. Our approach should also extend readily to different tasks and other tactile sensors.

4 Reinforcement Learning Environments

Each task is provided with a set of observation spaces to allow for verification of the environments, comparisons between tasks, and an examination of multi-modal visuotactile control. Four standard types of observation space are considered:

Env State: Comprises state information from the simulator, which is difficult information to collect in the real world. We use this to give baseline performance for a task that is expected to act as an upper limit. The information in this state varies between environments but commonly includes tool center point (TCP) pose, TCP velocity, goal locations and the current state of the environment.

Tactile: Comprises images (128×128) retrieved from the simulated optical tactile sensor attached to the end effector of the robot arm (Figure 2 right). Where tactile information alone is not sufficient to solve a task, this observation can be extended with state information retrieved from the simulator.

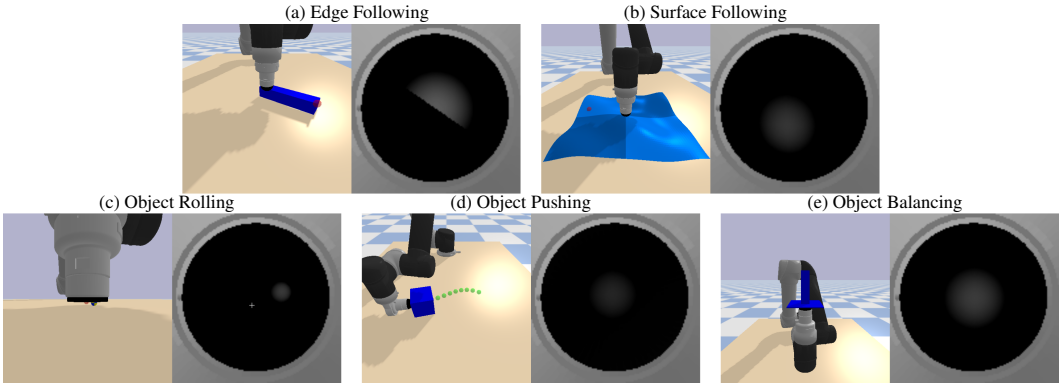

Figure 2: Tactile RL environments. (a) Traversing a randomly oriented edge while maintaining a set pose. (b) Traversing a randomly-generated 3D surface while maintaining a set penetration and normal orientation. (c) Manipulating a ball from a random initial position to a goal location. (d) Manipulating a cube along a randomly-generated trajectory. (e) Stabilising an object on the sensor tip.

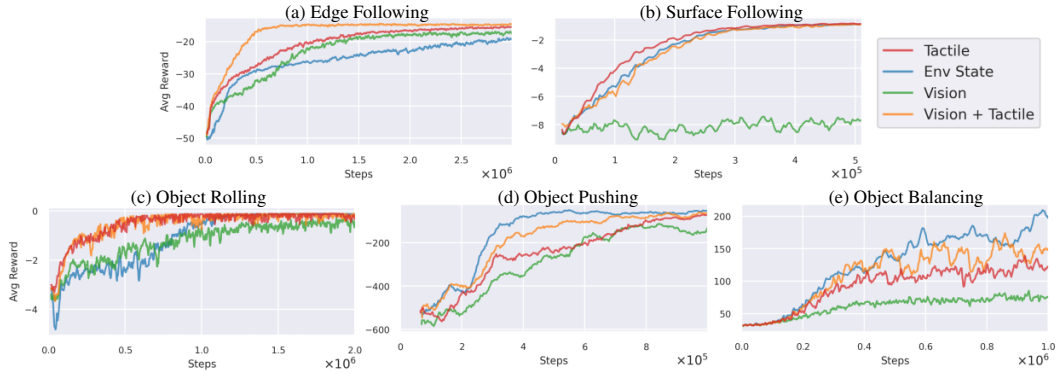

Figure 3: The DRL method PPO learns successful policies for the five considered RL environments. Panels show mean reward during training, smoothed with a window size of 50 and averaged over 3 seeds. Task reward commonly consists of negative distances between target and current poses, further details in Appendix C.

This can only include information that could be easily and accurately captured in the real world, such as the TCP pose that is available on industrial robotic arms and the goal pose.

Visual: Comprises RGB images (128×128) retrieved from a static, simulated camera viewing the environment (Figure 2 left). Only a single camera was used, although this could be extended to multiple cameras. As the simulated environment differs visually from the real-world environment, sim-to-real using RGB observations is challenging, requiring an approach like that of [5, 20].

Visual + Tactile: Combines the RGB visual and tactile image observations to into a 4-channel RGBT image. This case demonstrates a simple method of multi-modal sensing.

4.1 Tactile Exploration Environments

In this set of tasks, the robot interacts with a physical environment that stays static, with the objective to learn policies to safely explore physical areas. These policies can be used to report information about novel objects, simplify human-control methods and improve operational safety of the robots. Two exploration environments have been created, for edge and surface following. More complete details on these environments are given in Tables 5 and 6 of Appendix C.

Edge Following: The edge-following environment is used to train an agent that maintains its pose relative to an edge whilst traversing the edge towards a goal location. Previous work has demonstrated that robust 2D contour following can be achieved with supervised learning techniques [21]. Here we demonstrate this task can also be completed via sim-to-real reinforcement learning.

Surface Following: The surface-following environment is used to train an agent that maintains a set contact penetration of the sensor whilst orientating the TCP normal to an undulating 3D surface generated using OpenSimplex noise [22]. For ease of use, we automatically direct the sensor in the direction of the goal, analogous to the pose-based tactile servo control methods introduced in [23].

4.2 Non-prehensile Manipulation Environments

In this set of tasks, the objective is to learn policies that manipulate external objects in a desired manner. As our focus is on single sensors in this work, we consider non-prehensile manipulation. However, the simulation does support rendering of multiple tactile images, which could in principle be used simulate tactile grippers, manipulators or multiple robot arms in future work. Three manipulation tasks are proposed: object rolling, object pushing and object balancing. More details on the non-prehensile manipulation environments are given in Tables 7, 8 and 9 of Appendix C.

Object Rolling: The object-rolling environment requires the manipulation of small spherical objects into a goal position within the TCP coordinate frame. The agent must learn a policy to roll the object from a random starting position to a random goal position. In this environment, a flat tactile sensor tip is used to simplify the motion needed to maintain contact with a spherical object.

Object Pushing: In this task, the objective is to push an object to a goal location using the tactile image, as considered in previous work using supervised learning techniques [24]. Here we consider a cube pushed along a randomly-generated trajectory. The initial pose of the cube is randomised within limits and the trajectory generated using OpenSimplex noise.

Object Balancing: This task is analogous to the well-known 2D inverted pendulum problem [25], where an unstable pole with flat base is balanced on the tip of a sensor that points upwards. A random force perturbation is then applied to the object to cause instability. The objective is to learn a policy that applies planar actions to counteract the rotation of the balanced pole.

4.3 Reinforcement Learning Results

This work uses the Stable-Baselines-3 [26] implementation of PPO to train all tasks (further details in Appendix B). For an initial comparison, we train all tasks using all available observation spaces, averaged over 3 seeds. Training results are given in Figure 3, deterministic evaluation results and more detailed plots are given in Appendix D.

Environment state observations are used to verify that the environment leads to desired policies. As no image rendering is performed, this runs faster than tactile or visual observations; hence, it was used to find stable hyperparameters. Although comprised of ideal state information, we find that tactile image observations can lead to more sample-efficient and stable training, which is most clearly visible in the Edge-Following and Object-Rolling environments (Figs 3a, 3c).

Policies trained exclusively using visual observations tended to perform worst, which we attribute to most tasks being targeted towards tactile data; for example, details of the contact can be obscured using a single external camera. The best visual agent performance was for the object pushing task, which is the coarsest manipulation challenge. Whilst the object rolling task appears to perform well with visual observations, there is a notable performance gap where visual polices do not accurately learn the desired behaviour. When combining visual and tactile observations, successful learning can take place. Despite the additional complexity in the observations, agents can learn well with notable improvements in sample efficiency for the Edge-Follow and Object-Push tasks (Figs 3a, 3d).

5 Real-to-Sim Image Translation

An advantage of using tactile images as the main form of observation is that the image space is more simplistic (e.g. markers on a uniform background) compared with the variety of visual images from a scene. Vision-based images are affected by features such as changing lighting conditions, shadows and texture that usually constitute a superfluous level of detail when solving a task. Simulating these visual features can make the learning more difficult because of the increased complexity of the observations and the added computation to render these images; conversely, ignoring these visual features would make image translation more difficult because of the greater simulation and reality gap, as they are still prevalent in reality, necessitating techniques such as image randomisation or RL task-aware training [5, 20]. These complexities are not present for the real and simulated tactile images considered here because the camera is confined within the enclosure of the tactile sensor.

Conversely, a disadvantage of using tactile images as the main observations is that the tasks require the robot to be in close proximity, or touching, the environment. Whilst this contact may be necessary for some tasks, such as manipulation, it does make exploration in reality a challenge because of the self-inflicted damage that can occur. The approach proposed here is to use a separate data-

collection stage, where the tactile sensor explores a series of configurations in a safe and controlled manner. As the tactile image space is relatively confined, an efficient exploration of a representative sample of configurations for a specific RL task becomes possible. Whilst some sensor configurations are not possible to sample in reality, such as large penetrations of the sensor that would cause damage, we aim to train a model that generalises to those unreachable configurations.

Specifically, in this work we treat the sim-to-real problem as a supervised image-to-image translation problem. The same data collection procedure is performed in both simulation and reality, producing a data set of real and simulated image pairs. Here we choose to transfer from real-to-sim images, because the real images are richer in information with details such as shear forces that we choose not to model in simulation. Generative Adversarial Networks (GANs) are the state of the art for realistic image generation. Here we use the pix2pix [27] architecture for image-to-image translation, which uses the U-net [28] architecture for the image conditioned generator and a standard convolutional network for the discriminator (Appendix E shows this architecture applied to tactile images).

5.1 Data Collection

Three data sets corresponding to distinct environments are collected for sim-to-real transfer, each with 5000 image pairs for training and 2000 pairs for validation. We approximate the task environment with a simplified static environment consisting of 3d-printed components and no specialist hardware. Per task, these datasets take $\sim$5 hours to collect on physical hardware and $\sim$100 seconds in simulation. Following a previous investigation of contact-induced shear with the same optical tactile sensor [29], we deliberately induce shear perturbations by randomly sliding the sensor during data collection (details given in Appendix F). Past work has shown this step is key to ensuring that the trained neural network outputs are insensitive to unavoidable motion-dependent shear during task performance (for more details, see [23]). We do not model any sensor shear in simulation to ensure that the real-to-sim image generation is also insensitive to shear on the real sensor.

5.2 Pix2Pix Training

The only change to the pix2pix architecture was to replace instance normalisation with spectral normalisation, with other default parameters sufficient for training on tactile images (full details in Appendix E). This change reduced droplet artefacts that were otherwise prevalent in our generated images. In addition, excluding the border from the simulated tactile images improved training, as otherwise the GAN focused on generating a realistic border instead of the tactile imprint. This is important because the imprint is the useful component when learning control policies. Instead, the border is re-added from a saved reference image after generation occurs.

Accurate tactile image generation is achieved with minimal difference between generated and target images (example images for each data set shown in Appendix F, Figure 11). For the edge, surface and probe datasets, mean SSIM scores across the full validation set are 0.9955, 0.9924 and 0.9942 respectively. The source of the small errors are highlighted in the shown SSIM image differences (Figure 11, right). The imprint borders in the generated images lack some sharp details generated in simulation, which will likely be due to a slight elastic stretching of the skin of the real tip that is not modelled in simulation. An approach such as Gaussian smoothing of the simulated images, as in [12], could reduce this effect, although we did not find it necessary to explore in the present work.

Crucially, the generator can interpret and generate image features useful to training RL policies. For example, edge orientation and position are accurately captured in the GAN. Also, the generator can generalise to images unseen during training. For example, all training data only had imprints from one source in each training image; however, during inference, multiple sources can be applied and the generator still produces realistic outputs (video demonstrations available *here*). Similarly, the generator appears to generalise outside the training data for penetration depth, which is important for helping to ensure that the sensor can avoid damage by not pressing too far into a surface.

5.3 Supervised Learning Comparison

For comparison to the only existing sim-to-real approach for the TacTip [11], we implement a supervised learning task to predict the radial displacement and polar angle of an edge pressed into the sensor (results in Appendix G). Domain adaptation with our proposed method of tactile simulation out-performs the previous approach using elastic deformation simulation and domain randomisation, with approximately a 3-fold improvement in Mean Absolute Error (MAE) of the predictions.

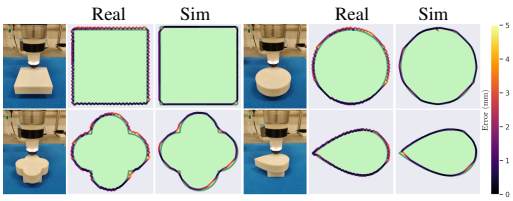

Figure 4: Comparison between sim and real performance for the Edge-Following RL environment. Using several flat shapes (circle, square, clover, foil) to evaluate the policy generalisation performance.

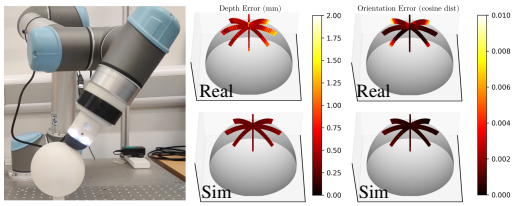

Figure 5: Comparison between sim and real performance for the Surface-Following RL environment. Positional error (left), Orientation error (right), Real data (top), simulated data (bottom).

6 Results: Sim-to-Real Policy Transfer

Successful sim-to-real policy transfer was achieved on 4 of the 5 proposed tasks. Qualitative demonstration videos are available at *https://sites.google.com/view/tactile-sim-to-real*

Edge Following: The learned edge-following behaviour is evaluated over several novel flat shapes in the environment (Figure 4). The learned policy successfully traverses all objects, generalising to novel features such as curved edges and sharp internal or external corners not present during training. As these shapes are 3D printed, we have access to the ground truth in the CAD files (accurate to the precision of the 3D printer). We compare the trajectories taken in simulation to those in reality, reporting the distance from each point of the trajectory to the nearest ground truth node (Table 1). The results show that the sensor maintains close proximity to the target edge.

Table 1: Quantitative results for edge following task. Mean distance from ground truth shape taken over 1000 evaluation steps.

	Real	**Sim**
Square	1.47mm	0.63mm
Circle	1.50mm	0.80mm
Clover	1.58mm	1.38mm
Foil	1.09mm	0.47mm

The drop in performance on the physical robot seems related to an oscillating behaviour also present in simulation. The movements are exaggerated in reality due to increased latency between capturing an observation and predicting an action. A likely cause is the action-range clipping used in the PPO algorithm, which results in the action extremes being more prevalent. Approaches such as squashing functions [30], beta distributions [31] or constraint-based RL [32] could mitigate this artefact.

Surface Following: To evaluate the surface-following behaviour, a spherical object is traversed from its centre outwards in a set direction ranging over 360° in 45° intervals (Figure 5). As with the edge following task, we use a 3D-printed shape whose ground truth is known from the CAD model. The sensor successfully and accurately traverses the object on the physical robot despite texture and frictional forces not modelled in simulation. Like the edge following task, the accuracy drops between sim and real, although the evaluated policies still exhibit the desired behaviour. A more extensive qualitative test is given in the supplementary video for the undulating surface shown in Fig.1.

Table 2: Quantitative results for surface following task. Mean distance from ground truth evaluated over 1000 steps.

	Real	**Sim**
Depth Error	0.57mm	0.30mm
Cosine Error	0.00118	0.00054

Object Rolling: The object-rolling behaviour is tested by manipulating ball bearings (on an extended dimension range of 2-8mm diameter) from random initial positions to random goal locations. During the evaluation, the position of the ball bearing relative to the sensor is obscured, so the imprint of the object is tracked on the tactile image using basic computer vision (blob detection). The target is reached when the pixel distance is less than a 5-pixel threshold (∼2mm). On the physical robot, this is performed for 25 trajectories with each bearing, resulting in 100 consecutive successful trials (subset shown in Figure 6). To compare with simulated results, we apply the same initial and target positions then perform the same pixel tracking. The sim and real results are similar, albeit with greater noise in the real trajectories.

Object Pushing: To evaluate this task we tracked the objects following the method given in [24]. Importantly Aruco markers were only used for obtaining quantitative results and not as part of the observation. This task was the most challenging to achieve accurate perfor-

Table 3: Quantitative results for object pushing task. Mean Euclidean distance from trajectory.

	Straight	**Curve**	**Sine**
Real Cube	11.5mm	11.7mm	13.6mm
Sim Cube	11.1mm	10.1mm	12.7mm
Real Tri	11.0mm	14.0mm	12.6mm
Sim Tri	21.8mm	10.2mm	11.0mm
Real Cyl	13.3mm	16.7mm	9.9mm
Sim Cyl	24.1mm	13.9mm	12.7mm

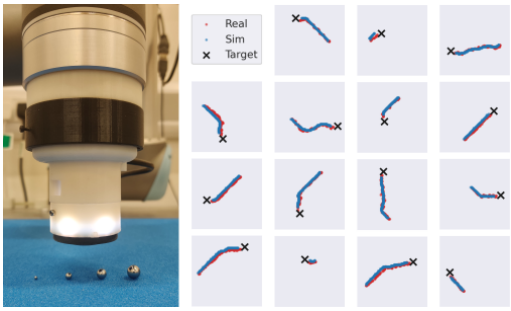

Figure 6: Real and simulated trajectories for policy evaluation of the object rolling task. A ball bearing is rolled an initial position to a goal location. Sizes range from 2-8mm diameter (columns: left to right).

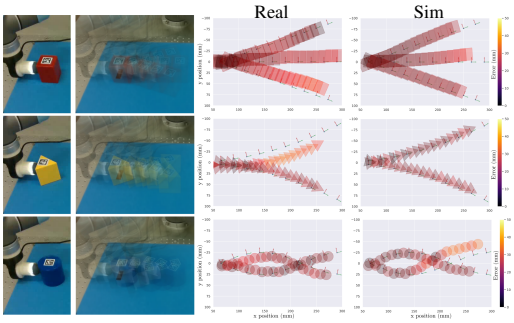

Figure 7: Evaluation of the pushing task. Trajectories show a cube (top), triangular prism (middle) and cylinder (bottom) pushed along straight, curved and sinusoidal trajectories respectively.

mance, likely due to the approximations of the simulated sensor tip, object being manipulated, and frictional interactions. Despite this, when evaluated in reality, task success was achieved across several trajectories with objects of different shape to that seen during training (examples in Figure 7). Performance both in simulation and reality was sensitive to initial conditions and perturbations, with compounding errors causing the object to veer off the trajectory. Future work could improve robustness by learning under additional dynamics randomisation, introduction of random force perturbations or by better matching the physical properties of the real objects.

Object Balancing: The physical task remains an outstanding challenge, because the successful training in simulation required physics parameters outside the capabilities of our UR5 robot.

7 Discussion and Future Work

In this work, we demonstrated that zero-shot sim-to-real policy transfer is a viable approach for tactile-based RL agents. To learn the policy, we created a fast method for simulating tactile images based on contact geometry. From these images, distinct policies were learned for several physically-interactive tasks requiring a sense of touch. We demonstrated that a data-driven model for real-to-sim image translation can be embedded into the control loop for successful sim-to-real policy transfer. There are several future directions for extending and improving this approach.

The tactile simulation of both the contact dynamics and captured information in the model could be improved. Our contact dynamics model used used rigid bodies with soft contacts as an approximation to the deformation of real tactile sensors. Extending this work with soft-body deformation should enable a more realistic simulation, which was not pursued in this initial study because of increased computational costs and instability issues with the simulators. The captured information in the present study focused exclusively on contact geometry, which in principle could be extended to include global shear information through contact data available in most physics simulators. Texture could be also included using rendering techniques common in photo-realistic image generation. Local shear of the sensor would be challenging to simulate, for example during incipient slip, due to the contact reduction commonly used in physics engines for computational efficiency; however, this capability may be possible in soft-body simulations where these local forces are required.

For translating real to simulated images, improvement could be made to both the GAN and training data. We used a conventional pix2pix [27] approach, which can be improved with extended methods [33]. In addition, a distinct data set was collected for training the GAN in each learning environment. Because of the constrained nature of the tactile image space, a general tactile data set could potentially be used train a single image-translation model for policy transfer over multiple tasks.

The environments considered here focused on skills achievable with only a single sensor and where the optimal behaviour is obvious for a human to interpret, as needed to verify that RL with a sim-to-real approach is a viable method. Future work could focus on tasks where ideal control policies are more complicated or unknown, so the RL framework can be fully exploited. An interesting topic would be to extend these methods to prehensile tasks such as grasping and dexterous manipulation, where tactile information will be valuable for learning desirable and robust policies. Although only one sensor was used in this current work, our tactile simulation does support multiple sensors, offering the opportunity to extend to more complex tasks involving hands with multiple tactile fingertips.

References

[1] I. Akkaya, M. Andrychowicz, M. Chociej, M. Litwin, B. McGrew, A. Petron, A. Paino, M. Plappert, G. Powell, R. Ribas, et al. Solving rubik's cube with a robot hand. *arXiv:1910.07113*, 2019.

[2] J. Lee, J. Hwangbo, L. Wellhausen, V. Koltun, and M. Hutter. Learning quadrupedal locomotion over challenging terrain. *Science Robotics*, 5(47), 2020.

[3] J. Ibarz, J. Tan, C. Finn, M. Kalakrishnan, P. Pastor, and S. Levine. How to train your robot with deep reinforcement learning: lessons we have learned. *IJRR*, 2021.

[4] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

[5] S. James, P. Wohlhart, M. Kalakrishnan, D. Kalashnikov, A. Irpan, J. Ibarz, S. Levine, R. Hadsell, and K. Bousmalis. Sim-to-real via sim-to-sim: Data-efficient robotic grasping via randomized-to-canonical adaptation networks. In *CVPR*, pages 12627–12637, 2019.

[6] Y. Narang, B. Sundaralingam, M. Macklin, A. Mousavian, and D. Fox. Sim-to-real for robotic tactile sensing via physics-based simulation and learned latent projections. *arXiv preprint arXiv:2103.16747*, 2021.

[7] Y. S. Narang, B. Sundaralingam, K. Van Wyk, A. Mousavian, and D. Fox. Interpreting and predicting tactile signals for the syntouch biotac. *arXiv preprint arXiv:2101.05452*, 2021.

[8] C. Sferrazza, T. Bi, and R. D'Andrea. Learning the sense of touch in simulation: a sim-to-real strategy for vision-based tactile sensing. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 4389–4396. IEEE, 2020.

[9] C. Sferrazza and R. D'Andrea. Sim-to-real for high-resolution optical tactile sensing: From images to 3d contact force distributions. *arXiv preprint arXiv:2012.11295*, 2020.

[10] T. Bi, C. Sferrazza, and R. D'Andrea. Zero-shot sim-to-real transfer of tactile control policies for aggressive swing-up manipulation. *IEEE Robotics and Automation Letters*, 2021.

[11] Z. Ding, N. F. Lepora, and E. Johns. Sim-to-Real Transfer for Optical Tactile Sensing. *ICRA*, pages 1639–1645, 2020.

[12] D. F. Gomes, P. Paoletti, and S. Luo. Generation of gelsight tactile images for sim2real learning. *IEEE Robotics and Automation Letters*, 6(2):4177–4184, 2021.

[13] S. Wang, M. Lambeta, L. Chou, and R. Calandra. Tacto: A fast, flexible and open-source simulator for high-resolution vision-based tactile sensors. *arxiv:2012.08456*, 2020.

[14] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora. The TacTip Family: Soft Optical Tactile Sensors with 3D-Printed Biomimetic Morphologies. *Soft Robotics*, 5(2):216–227, 2018. ISSN 2169-5172.

[15] M. K. Johnson and E. H. Adelson. Retrographic sensing for the measurement of surface texture and shape. In *CVPR*, pages 1070–1077. IEEE, 2009.

[16] A. Padmanabha, F. Ebert, S. Tian, R. Calandra, C. Finn, and S. Levine. Omnitact: A multi-directional high-resolution touch sensor. In *ICRA*, pages 618–624. IEEE, 2020.

[17] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, et al. Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation. *RAL*, 5(3):3838–3845, 2020.

[18] N. Kuppuswamy, A. Alspach, A. Uttamchandani, S. Creasey, T. Ikeda, and R. Tedrake. Soft-bubble grippers for robust and perceptive manipulation. *arXiv:2004.03691*, 2020.

[19] C. Sferrazza and R. D'Andrea. Design, Motivation and Evaluation of a Full-Resolution Optical Tactile Sensor. *Sensors*, 19(4):928, 2019. ISSN 1424-8220.

[20] K. Rao, C. Harris, A. Irpan, S. Levine, J. Ibarz, and M. Khansari. Rl-cyclegan: Reinforcement learning aware simulation-to-real. In *CVPR*, pages 11154–11163. IEEE, 2020.

[21] 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 biomimetic tactile sensor. *RAL*, 4(2):2101–2107, 2019.

[22] S. Kurt and S. A. Opensimplex noise. https://github.com/lmas/opensimplex, 2021.

[23] N. F. Lepora and J. Lloyd. Pose-based servo control with soft tactile sensing. *arXiv:2012.02504*, 2020.

[24] J. Lloyd and N. F. Lepora. Goal-driven robotic pushing using tactile and proprioceptive feedback. *arXiv:2012.01859*, 2020.

[25] A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE SMC*, 13(5):834–846, 1983.

[26] A. Raffin, A. Hill, M. Ernestus, A. Gleave, A. Kanervisto, and N. Dormann. Stable baselines3. Available: https://github.com/DLR-RM/stable-baselines3, 2019.

[27] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[28] O. Ronneberger, P. Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241. Springer, 2015.

[29] N. F. Lepora and J. Lloyd. Optimal deep learning for robot touch: Training accurate pose models of 3d surfaces and edges. *RAM*, 27(2):66–77, 2020.

[30] T. Haarnoja, A. Zhou, K. Hartikainen, G. Tucker, S. Ha, J. Tan, V. Kumar, H. Zhu, A. Gupta, P. Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv:1812.05905*, 2018.

[31] P.-W. Chou. The beta policy for continuous control reinforcement learning. Master's thesis, Pittsburgh: Carnegie Mellon University, 2017.

[32] S. Bohez, A. Abdolmaleki, M. Neunert, J. Buchli, N. Heess, and R. Hadsell. Value constrained model-free continuous control. *arXiv:1902.04623*, 2019.

[33] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *CVPR*, pages 8798–8807. IEEE, 2018.

[34] C. Chen, U. Jain, C. Schissler, S. V. A. Gari, Z. Al-Halah, V. K. Ithapu, P. Robinson, and K. Grauman. SoundSpaces: Audio-Visual Navigation in 3D Environments. In *ECCV*, volume 12351 LNCS, pages 17–36. Springer, 2019.

[35] S. James, Z. Ma, D. R. Arrojo, and A. J. Davison. RLBench: The Robot Learning Benchmark & Learning Environment. *RAL*, 5(2):3019–3026, 2019.

[36] Y. Zhu, J. Wong, A. Mandlekar, and R. Martín-Martín. robosuite: A Modular Simulation Framework and Benchmark for Robot Learning. *arXiv:2009.12293*, 2020.

[37] Y. Hu, L. Anderson, T. Li, Q. Sun, N. Carr, J. Ragan-Kelley, and F. Durand. Difftaichi: Differentiable programming for physical simulation. In *Proc. of ICLR*, 2020.

[38] E. Heiden, D. Millard, E. Coumans, Y. Sheng, and G. S. Sukhatme. NeuralSim: Augmenting Differentiable Simulators with Neural Networks. *arXiv:2011.04217*, 2020.

[39] E. Coumans and Y. Bai. Pybullet, a python module for physics simulation for games, robotics and machine learning. Available: http://pybullet.org, 2016–2019.

[40] J. Matas, S. James, and A. J. Davison. Sim-to-real reinforcement learning for deformable object manipulation. In *CoRL*, pages 734–743. PMLR, 2018.

[41] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[42] I. Kostrikov, D. Yarats, and R. Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv:2004.13649*, 2020.

[43] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas. Reinforcement learning with augmented data. In *Advances in Neural Information Processing Systems*, volume 33, pages 19884–19895. Curran Associates, Inc., 2020.

[44] D.-A. Clevert, T. Unterthiner, and S. Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

A Robotics Simulation

A.1 Physics Engine

Physics-based simulation has played a vital role in the field of robotics, enabling rapid prototyping, testing and experimentation. The accuracy and speed of simulation has scaled with available computation. Modern methods such as RL have leveraged this to simulate very large data sets that are otherwise impractical to collect from the real world, hence far more complex and general behaviours can be learnt. The majority of simulated robotics has focused on rigid-body dynamics and vision sensors. More recently, a range of specific environment suites have been introduced to bring simulation closer to reality or to facilitate research in an under-represented direction [34, 35, 36, 37, 38]. Collectively, these works highlight the importance of increasing the breadth and capabilities of simulation software available to researchers.

There are many choices for physics engines when developing a new suite. Here we choose the PyBullet [39] for its fast GPU rendering, support for deformable objects [40], fast and reliable kinematics and dynamics solvers, and its demonstrated sim-to-real success in robotics [5]. Importantly, PyBullet is open source and non-commercialised software which helps to improve accessibility and lower the barrier of entry for RL research. That said, the tools used here to build our tactile suite are available in other physics engines such as Mujoco, Gazebo, Nvidia-isaacgym and Unity ML agents.

A.2 Control

Throughout this work we use Cartesian velocity control where the control input is a desired velocity (twist) specified by a 6 DoF action constrained to the allowed modes of control of the task. A control rate specifies the frequency that new actions can be sent to the robot, with maximum velocity limits also imposed. For most of our experiments, we set a velocity control rate of 10 Hz, a maximum linear velocity of $10\,\mathrm{mm\,s^{-1}}$ and a maximum angular velocity of $5\,^{\circ}\,\mathrm{s}^{-1}$. When undergoing a series of predefined and random actions we notice no significant difference in Tool Centre Point (TCP) pose between simulation and the real robot.

Work-space coordinate frames are set in both simulation and reality, specific to each environment, with each action sent to the robot consisting of a Cartesian move relative to the work frame. Therefore, the learned policies can be transferred from sim-to-real without exactly replicating the simulated task; for example edges and surfaces can be placed in alternative locations provided the work frame is set correctly. As a consequence, the policy transfers even when there are notable differences in the simulation, such as mirrored arm configurations used in this work. Thus, in principle the policies could also be transferred to other robot arms, providing the same speed and frequency of control can be achieved.

B Reinforcement Learning Parameters

Near-default hyper-parameters are used in all training (full list in Table 4). Image-based observations use the Atari Nature [41] convolutional layers followed by two 256-node fully connected (FC) layers. State observations use only the FC layers. For tasks that require both image and state data, the state data is passed through two 64-node FC layers and the output concatenated with the flattened output of the convolutional layers, which is then passed through the final FC layers for action and value prediction. The convolutional weights are shared for all policy and value networks. Small random image translation augmentations help to improve performance and stabilise training, as proposed in [42, 43].

Table 4: RL and network hyperparameters.

	Param	Value
Feature Extractor	Input dim	[128, 128, C]
	Conv filters	[32, 64, 64]
	Kernel widths	[8, 4, 3]
	Strides	[4, 2, 1]
	Pooling	None
	Output dim	512
	State Encoder	[64, 64]
	Activation	*ReLU*
	Initialiser	*Orthoganal*
RL Net	Policy	[256, 256]
	Value	[256, 256]
	Activation	*Tanh*
PPO parameters	Learning Rate	3×10^{-4}
	n/ Envs	10
	Epoch steps	2048
	Batch size	64
	n/ Epochs	10
	Discount (γ)	0.95
	GAE lambda	0.9
	Clip range	0.2
	Entropy coeff	0.0
	VF coeff	0.5
	Max grad norm	0.5
	KL limit	0.1
	Optimiser	*Adam*

C Reinforcement Learning Environment Details

Table 5: Edge Follow environment description.

Observation	• **Env State:** {TCP pos, TCP lin vel, Goal pos, Edge ang} • **Tactile:** {Tactile Image} • **Vision:** {RGB Image} • **Vision + Tactile:** {RGB Image, Tactile Image}
Action Space	$\{x, y\}$
Reward	- (Euclidean distance from TCP to goal + perpendicular distance from TCP to edge)
Termination	• max episode length reached • euclidean distance from TCP to goal $< 1\text{cm}$
History	1 Frame.
Randomisation	• Edge randomly orientated through 360°. • Distance tip is embedded onto edge is randomly selected between 1.5mm and 3.5mm.

Table 6: Surface Follow environment description.

Observation	• **Env State:** {TCP pos, TCP orn, TCP lin vel, TCP ang vel, Goal pos, Target surface height, Target surface normal} • **Tactile:** {Tactile Image} • **Vision:** {RGB Image} • **Vision + Tactile:** {RGB Image, Tactile Image}
Action Space	$\{z, Rx, Ry\}$
Reward	-(z difference between TCP and local surface index + cosine difference between TCP normal and local surface normal)
Termination	• max episode length reached • euclidean distance from TCP to goal $< 1\text{cm}$
History	1 Frame.
Randomisation	• Surface randomly generated w/ OpenSimplex noise. • Direction of goal randomly selected from $[0°, 360°]$.

Table 7: Object Roll environment description.

Observation	• **Env State:** {TCP pos, TCP orn, TCP lin vel, TCP ang vel, Obj pos, Obj orn, Obj lin vel, Obj ang vel, Goal pos, Obj radius} • **Tactile:** {Tactile Image, Goal pos} • **Vision:** {RGB Image, Goal pos} • **Vision + Tactile:** {RGB Image, Tactile Image, Goal pos}
Action Space	$\{x, y\}$
Reward	-(euclidean distance from object to goal)
Termination	• max episode length reached • euclidean distance from object to goal $< 1\text{mm}$
History	1 Frame.
Randomisation	• Random starting position of object in TCP frame. • Random marble size between 5mm and 10mm diameter. • Random distance embedded into the sensor.

Table 8: Object Push environment description.

Observation	• **Env State:** {TCP pos, TCP orn, TCP lin vel, TCP ang vel, Obj pos, Obj orn, Obj lin vel, Obj ang vel, Goal pos, Goal orn} • **Tactile:** {Tactile Image, TCP pos, TCP orn, Goal pos, Goal orn} • **Vision:** {RGB Image, TCP pos, TCP orn, Goal pos, Goal orn} • **Vision + Tactile:** {RGB Image, Tactile Image, TCP pos, TCP orn, Goal pos, Goal orn}
Action Space	$\{y, Rz\}$
Reward	-(Euclidean distance from object to goal + cosine distance from object orn to goal orn + cosine distance from TCP normal to object normal)
Termination	• max episode length reached • Euclidean distance from object to final goal $< 2.5\text{cm}$
History	1 Frame.
Randomisation	• Random trajectory of goals generated with OpenSimplex Noise.

Table 9: Object Balance environment description.

Observation	• **Env State:** {TCP pos, TCP orn, TCP lin vel, TCP ang vel, Obj pos, Obj orn, Obj lin vel, Obj ang vel} • **Tactile:** {Tactile Image} • **Vision:** {RGB Image} • **Vision + Tactile:** {RGB Image, Tactile Image}
Action Space	$\{x, y\}$
Reward	+1 per step
Termination	• max episode length reached • object tilts passed set angle (35°)
History	2 Frames.
Randomisation	• Random external force perturbation applied at start of episode.

D Full Reinforcement Learning Results

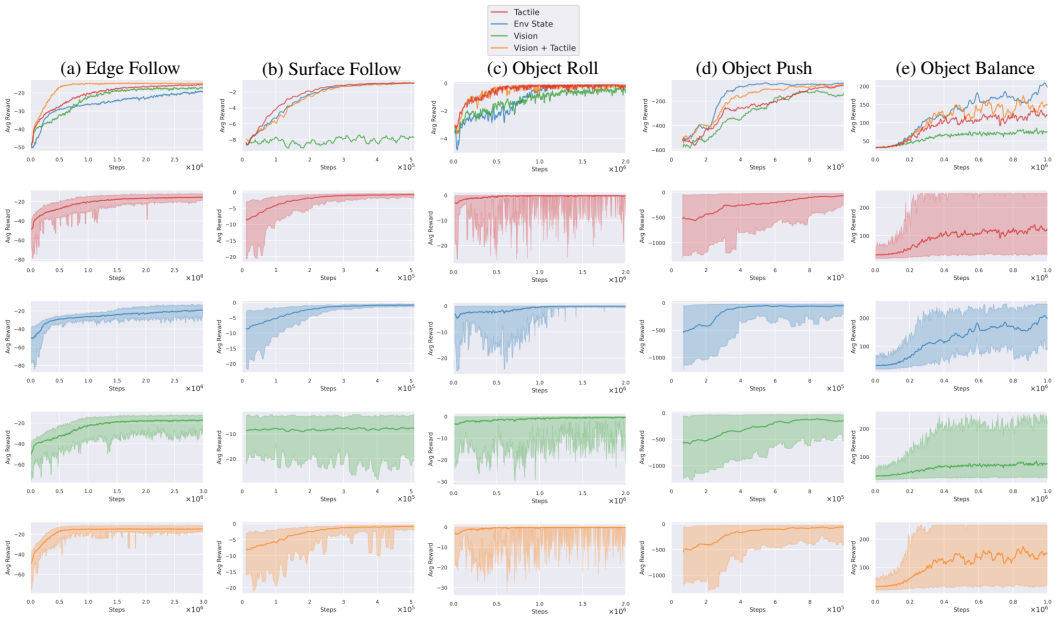

Figure 8: Full training results of reinforcement learning agents. Results are smoothed with a window size of 50 followed by averaging over 3 seeds. Shaded regions indicate maximum and minimum reward achieved over the 3 seeds (after smoothing).

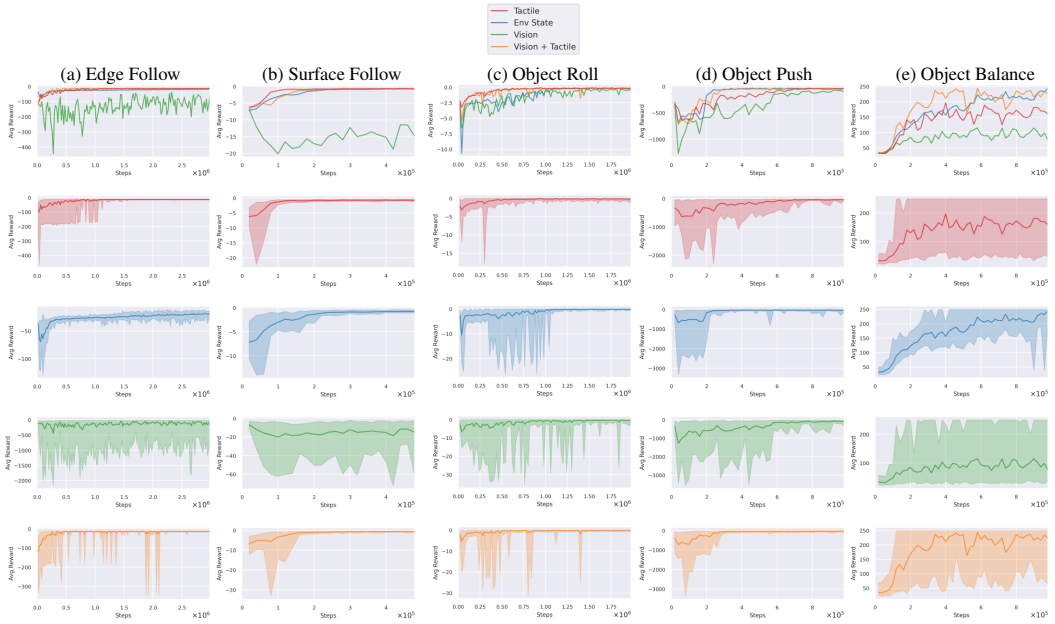

Figure 9: Full evaluation of reinforcement learning agents throughout training. 10 evaluation episodes occur every 20,000 steps, using deterministic actions. Results averaged over 3 seeds. Shaded regions indicate maximum and minimum reward achieved over the 3 seeds.

E Pix2Pix Architecture

Table 10: Pix2Pix architecture and parameters.

	Layer	Details			
Params	Batch Size	64			
	Learning Rate	0.0002			
	Image Norm	True			
	Image Trans	[2.5%, 2.5%]			
	Loss Weights	Wgan: 1.0, Wpix: 100.0]			
Generator	Input dim	[128, 128, 1]			
	Output dim	[128, 128, 1]			
		Input	**Output**	**Dropout**	**Norm**
	Down 1	1	64	None	False
	Down 2	64	128	None	True
	Down 3	128	256	None	True
	Down 4	256	512	0.5	True
	Down 5	512	512	0.5	True
	Down 6	512	512	0.5	True
	Down 7	512	512	0.5	True
	Up 1	512	512	0.5	True
	Up 2	1024	512	0.5	True
	Up 3	1024	512	0.5	True
	Up 4	1024	256	0.5	True
	Up 5	512	128	None	True
	Up 6	256	64	None	True
Discriminator	Input dim	[128, 128, 2]			
	Output dim	[16, 16, 1]			
		Input	**Output**	**Norm**	
	Disc 1	2	64	False	
	Disc 2	64	128	True	
	Disc 3	128	256	True	
	Disc 4	256	512	True	

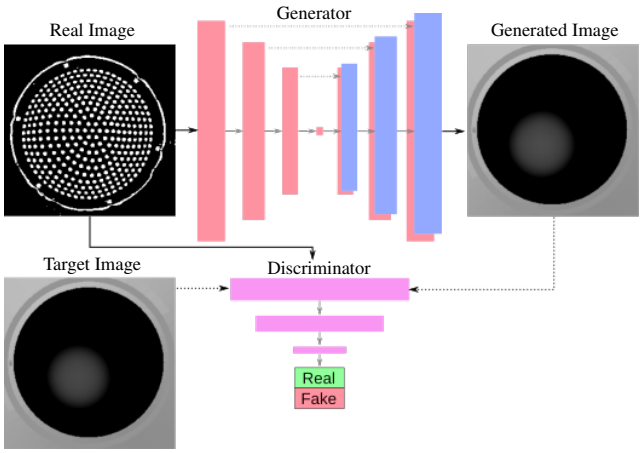

Figure 10: Real-to-sim translation of the tactile images uses a pix2pix-trained GAN. Real tactile images are processed by the generator to produce images that match the target simulated tactile images. The Discriminator is tasked with detecting whether an input tactile image pair is real or fake.

F Image Translation Data Collection

For the edge-following environment, we collect tactile images pressed onto a straight edge, varying the orientations, radial displacements and penetration of the sensor. A hemispherical sensor tip is used with the tool center point (TCP) located centrally at the end of the sensor. Relative to the TCP, the data is gathered over ranges: orientation $Rz \in [-179°, 180°]$, radial displacement $y \in [-6, 6]\text{mm}$, and penetration $z \in [3.5, 5.5]\text{mm}$.

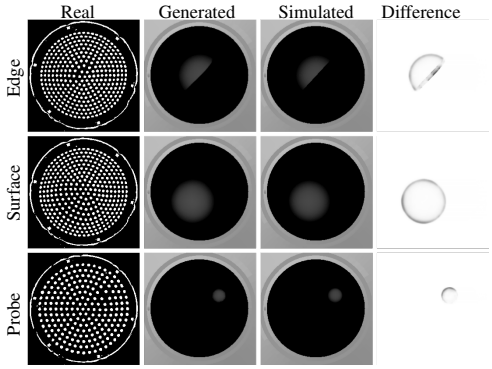

Figure 11: Image comparison between pairs of generated and simulated tactile images. Images are from validation sets for Real-to-Sim image translation. SSIM is used to create difference images.

For the surface-following and object-pushing environments, we collect tactile images pressed onto a flat surface, varying the orientations and penetration, also with a hemispherical tip. Relative to the TCP, data is gathered over ranges: orientation $\{Rx, Ry\} \in [-15°, 15°]$, and penetration $z \in [2, 5]\text{mm}$.

For the object-rolling environment, we collect tactile images pressed onto a spherical probe stimulus, using a flat sensor tip appropriate to this environment. 9 spherical probe stimuli are used, ranging over $[2, 6]\text{mm}$ radius in 0.5mm increments. The sensor is positioned to contact the probe at random placements within a 15mm disk surrounding the centre of the tip. In this case, shear is not introduced into the data collection because rolling objects induce negligible motion-dependent shear.

F.1 Full SSIM Scores

We measure the SSIM scores across the validation sets collected for each task, each consisting of 2000 image pairs. For the Edge, Surface and Probe datasets respectively, we found mean scores of $[0.995, 0.992, 0.994]$, min scores of $[0.985, 0.982, 0.974]$, and max scores of $[0.999, 0.999, 0.999]$. Whilst high scores are expected due to the relatively sparse target images, this indicates strong performance in all cases.

G Supervised Learning Comparison

Table 11: Mean Absolute Error (MAE) for Task I: predicting polar angle (radians), Task II: predicting radial displacement (mm) and Task III*: predicting position of a probe (mm).

Approach	Task I	Task II	Task III*
Ours	**0.079**	**0.119**	**0.059**
Ding et al. [11]	0.254	0.45	0.73

*Indicates some difference between tasks as discussed below.

Whilst we were unable to find a sim-to-real reinforcement learning baseline that does not require specific hardware, we can compare to previous work simulating the TacTip sensor for supervised learning tasks. Ding et al. [11] used an elastic deformation approach to simulate the TacTip sensor. They focussed on supervised learning from simulated pin positions rather than tactile images, and could accurately predict edge position and orientation, and the (x, y) position of a pole pressed into the sensor. In this section we draw a comparison between the presented method on their supervised learning tasks. The tasks are defined as follows: Task I predicts polar angle θ of an edge pressed into the sensor; Task II predicts radial displacement r of an edge pressed into the sensor; and Task III predicts the (x, y) location of the centre of a pole pressed into the sensor. More details are provided in [11] .

As this is not presented as a standardised benchmark, and we do not have access to the exact setup used, there are some differences in how these tasks have been carried out between the present work and [11]. In particular, we differ notably in Task III, where we use a flat TacTip sensor tip instead of the original hemispherical tip, because we are using the data, GANs and objects acquired for the object rolling task which needed a flat tip. That said, we expect there will only be small differences in accuracy between flat and curved TacTip tips.

For each task we first train a convolutional neural network to predict r, θ, x and y using a dataset of 10,000 simulated tactile images per task (taking $\sim$145 seconds to collect). We then collect a dataset of 2000 real sensor images and using GANs trained on edge or probe data, we translate from real-to-sim images to create the test datasets of generated images. The performance of the trained networks over the full generated datasets is then measured and compared with the results from [11].

As our simulated tactile data is comprised of images, we use a convolutional neural network (CNN). To avoid boundary issues over full rotations, we predict a sine/cosine encoding of angle. The CNN has 4 convolutional layers (each with kernel: 5, stride: 1, padding: 2, maxpool: 2) and 3 fully connected layers (dimensions = 1024, 512, output_dim). Batch normalization is applied only to the convolutional layers before the ELU [44] activation. Images of resolution 256×256-pixels were used.

Table 11 shows a large reduction in Mean Absolute Error (MAE) for both radial displacement and angle prediction in comparison with previous work simulating the TacTip [11]. A 3-fold improvement is found in predicting the angle θ (Task I) and a near 4-fold improvement when predicting radial distance (Task II). We also find a large decrease in MAE from $0.73\mathrm{mm}$ to $0.059\mathrm{mm}$ for position prediction (Task III), although as mentioned above there are some differences in setup.