Imagine2touch: Predictive Tactile Sensing for Robotic Manipulation using Efficient Low-Dimensional Signals.

Abdallah Ayad, Adrian Röfer, Nick Heppert, Abhinav Valada

***Abstract*— Humans seemingly incorporate potential touch signals in their perception. Our goal is to equip robots with a similar capability, which we term Imagine2touch. Imagine2touch aims to predict the expected touch signal based on a visual patch representing the area to be touched. We use ReSkin, an inexpensive and compact touch sensor to collect the required dataset through random touching of five basic geometric shapes, and one tool. We train Imagine2touch on two out of those shapes and validate it on the ood. tool. We demonstrate the efficacy of Imagine2touch through its application to the downstream task of object recognition. In this task, we evaluate Imagine2touch performance in two experiments, together comprising 5 out of training distribution objects. Imagine2touch achieves an object recognition accuracy of 58% after ten touches per object, surpassing a proprioception baseline.**

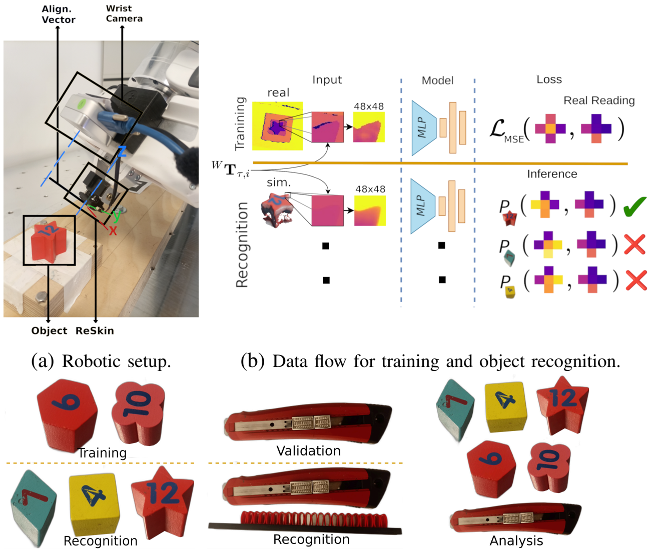
I. Introduction

Dexterous object manipulation requires the ability to identify an object and track its state throughout the task. Vision-based methods for object detection [1], pose estimation [2], and tracking [3] have become ubiquitous in robotics. However, vision, standalone, is an unreliable modality: the view of the object can be partially or fully occluded during the motion. Humans, on the other hand, are capable of manipulating objects under occlusion using their proprioceptive and tactile senses: we can find a pen in a backpack without looking, with the help of our ability to predict how possible objects feel. In this work, we seek to enable robots to perform similar visuo-tactile skills. For this purpose, we platform ReSkin [4], a magnetic-based tactile sensor, that is low cost and compact compared to its vision based counter parts such as [5]–[7].

To enable such skills, our approach fuses tactile and vision modalities using a common embedding. This approach has been widely researched across multiple modalities for various tasks [8]–[20]. Specifically, in this work, we contribute a cross-modal model for predicting tactile readings from corresponding depth-image patches and an object-recognition demonstrator in which we use our trained model in an ensemble of probabilistic models scheme. We show that despite the low dimensional sensor readings, our method is able to achieve competitive results on this task. We share code and model at https://github.com/AbdallahAyman/Imagine2touch.

II. Technical Approach

We first detail our Imagine2touch approach and then describe how we exploit it for the downstream task. **Imagine2touch-Model**: Our proposed model is a function $IT : z_d \Rightarrow \tilde{\tau}$, which takes a processed depth image $z_d \in \mathbb{R}^{48\times 48}$ of the object surface as input, and predicts the tactile reading that would be emitted touching the surface $\tilde{\tau} \in \mathbb{R}^{15}$. We implement this function as a neural network, consisting of a single 200-neurons-layer MLP encoder, followed by a 5-neurons-bottleneck, the output of which is fed through a single 500-neurons-layer MLP decoder, see Fig. 1b. We add an auxiliary *input-decoding* head with a separate 2000-neurons layer decoder to motivate modality fusion at the bottleneck stage. We choose a low-capacity model to prevent overfitting on our small dataset and due to the size imbalance between the input and output of our network. To train, validate, and analyze our model, we collect a dataset of tactile and visual pairs from 1630 samples from objects in Fig. 1e using the setup in Fig. 1a. The analysis shown in Fig. 2 indicates that mapping from one modality to the other should be possible.



(a) Robotic setup. (b) Data flow for training and object recognition.

(c) Objects: primitives. (d) Objects: tools. (e) Full dataset.

Fig. 1: *(a)*: Robotic setup for our approach. The alignment vector shows the direction on which the robot moves for collecting one data sample to pair the wrist camera and ReSkin readings. *(b)*: Data flow for training our model and using its inference in object recognition. The depth patch is cropped and processed from the full image using the end-effector pose $^W\mathbf{T}_{\tau,i}$ to match the touch area. It is then passed to the model, which we optimize using the MSE-loss between its output and the real touch reading. At recognition time, the robot has access only to possible 3D renderings. We use the probabilistic touch model in Sec. II for recognition. *(c)*: Objects set: primitives. First row: primitives used for training the model. Second row: primitives used for one instance of the object recognition experiment. *(d)*: Objects set: tools. First row: Tools used for validating the model. Second row: Tools used for the second instance of the object recognition experiment. *(e)*: Full objects dataset used for analysis.

Object Recognition: We use Imagine2touch to recognize objects from a possible set. We subsequently perform N touches, considering every touch as an independent proba-

Department of Computer Science, University of Freiburg, Germany.
This work was funded by the Carl Zeiss Foundation with the ReScaLe project and the BrainLinks-BrainTools center of the University of Freiburg.

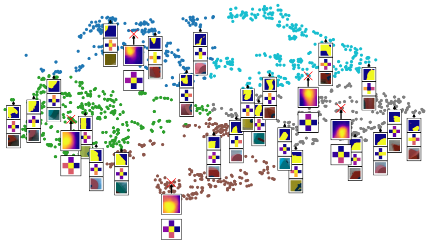


Fig. 2: t-SNE plot of our data distribution. Five-means clustering of our processed depth data points with the associated images, tactile visualizations, and RGB images for example points. The means of the clusters are projected and highlighted in red with associated mean processed depth images and mean tactile visualizations. The plot shows distributed sensor activation, and correspondence between the depth patches and the signals.

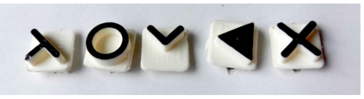


Fig. 3: Shape classification experiment setup. We use differently shaped *stamps* to indent the statically mounted sensor's gel pad in different locations. From left to right the shapes are: *T*, *circle*, *angle*, *triangle*, *cross*. All stamps are at most $10\,\mathrm{mm}$ wide and $3.5\,\mathrm{mm}$ deep.

bilistic model. We update our object hypothesis based on the touch measurement through aggregating these models.

Algorithmic Outline: Let O be a set of possible objects with known 3D representation (e.g. meshes/surfaces). Till we reach our maximum number of touches N, we calculate each object likelihood at step i given the current observations consisting of previous readings $T = \{\tau_1, \dots, \tau_i\}$. We sample from that distribution to get an object hypothesis $\tilde{o}$ on which we will sample a location $\tilde{l}$ based on a heuristic that aims to distinguish $\tilde{o}$. We don't detail this heuristic here.

Ensemble model: We assume that the probability of a possible object to be the true one after one touch follows a normal distribution with parameters (μ) and (σ^2). We assume those parameters are included in the delta between the standardized actual touch signal (τ) and the inferred touch signal ($\tilde{\tau}$). We define our probabilistic model as follows:

$$P(o = o'|\tau_i, \tilde{\tau_i}) = e^{-(\tau_i - \tilde{\tau_i})}, \tag{1}$$

where $o' \in O$ is a possible object and o is the real object. We assume independence between the touches to render the model more robust against noise and outliers. Additionally, we binarize the probabilities among possible objects by selecting a mutually exclusive winner to balance the weight of each model (i.e. touch). To finally recognize an object, akin to ensembling of weak models, the probability of an object can be calculated as the average among them:

$$P(o = o') = \frac{1}{N}(P(o'|\tau_i, \tilde{\tau_i}) + \ldots + P(o'|\tau_N, \tilde{\tau_N})) \tag{2}$$

III. Experiments

We define two experiments. The first is shape classification. It is a feasibility check for the second. It demonstrates that higher concepts can be extracted from ReSkin sensor.

Shape Classification: In the original work [4], Bhirangi *et al.* demonstrates that it is feasible to identify exact touch locations and interaction forces. To verify that the sensor can additionally be used for recognizing shapes (i.e. multi-touch contacts), we performed a preliminary experiment shown in Fig. 3. We conclude that the sensor data can be used to distinguish between different contact shapes, see Tab. I.

Shape	Letter”T”	Circle	Angle	Triangle	Cross	Total
Acc.	0.91	0.90	0.90	0.95	0.94	0.92

TABLE I: Shape experiment results. Shapes are shown respectively in Fig. 3. Classification Accuracy Results for each shape and their average are produced using an MLP model with a hidden layer of size 500, a bottleneck of size 10, and a 5-class classification head. The model is trained on 80% of 1280 datapoints using cross-entropy loss.

Object set	Primitives		Tools		*Mean*	
Touch Model	prop.	I2T	prop.	I2T	prop.	I2T
	23%	**50%**	**80%**	70%	46%	**58%**

TABLE II: Results from the object recognition experiment. We find that both I2T (Imagine2touch) and prop. (proprioception) work for identifying objects from the tools set. The primitive objects are more difficult for both methods due to their similarity in extent and tactile features. Despite the inherent difficulty of distinguishing similar objects without a dense measure such as vision, Imagine2touch exceeds random chance, and the proprioception baseline across the objects sets.

Object Recognition: To evaluate Imagine2touch performance against it, we define the following experiment. Akin to reaching into a backpack for a pen, in this task, the agent needs to identify the correct object out of a possible set, for which it has 3D models, on which Imagine2touch predicts hypothetical touches. We use the ensembling scheme from Sec. II and compare Imagine2touch performance to proprioception, which we define here as the delta bet. the real contact location and the nearest point in a possible 3D model. This baseline is the minimal tactile sense that could be implemented on any robot.

We conduct one instance for each out-of-training distribution primitive in Fig. 1c and one for each tool in 1d: The set O in Sec. II is adapted for each instance. As we do not focus on pose estimation, we fix the objects to wooden bases to immobilize them. The robot touches each object in each instance 10 times in locations sampled according to our heuristic mentioned in Sec. II. We report the success rate per touch of recognizing the object in Tab. II. In conclusion, we find that Imagine2touch improves over the proprioception performance in this task and generalizes to predicting touch signals outside of its training distribution.

IV. Conclusion

In this work, we investigated the use of the novel, low-cost, and compact ReSkin sensor as a platform to learn a predictive touch sense for general robotic tasks. We proposed Imagine2touch, a novel approach that infers expected tactile readings from small depth images of surfaces. We additionally introduce a procedure for collecting data to train the model. We leveraged the model for a downstream task involving five OOD objects and demonstrated that our model is able to generalize. We view our results as an encouraging step towards using inexpensive tactile sensors such as ReSkin more often in robotics. For future work, we see an opportunity for building the inverse of our approach: a model predicting depth images from tactile signals to obtain 3D object features. This would enable full tactile 3D reconstruction.

References

[1] C. Lang, A. Braun, and A. Valada, "Robust object detection using knowledge graph embeddings," in *DAGM German Conference on Pattern Recognition*, 2022, pp. 445–461.

[2] E. Chisari, N. Heppert, T. Welschehold, W. Burgard, and A. Valada, "Centergrasp: Object-aware implicit representation learning for simultaneous shape reconstruction and 6-dof grasp estimation," *arXiv preprint arXiv:2312.08240*, 2023.

[3] J. O. von Hartz, E. Chisari, T. Welschehold, W. Burgard, J. Boedecker, and A. Valada, "The treachery of images: Bayesian scene keypoints for deep policy learning in robotic manipulation," *IEEE Robotics and Automation Letters*, 2023.

[4] R. Bhirangi, T. Hellebrekers, C. Majidi, and A. Gupta, "Reskin: versatile, replaceable, lasting tactile skins," in *Proc. Conf. on Rob. Learn.*, 2021.

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

[6] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon *et al.*, "Digit: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 3, pp. 3838–3845, 2020.

[7] N. F. Lepora, Y. Lin, B. Money-Coomes, and J. Lloyd, "Digitac: A digit-tactip hybrid tactile sensor for comparing low-cost high-resolution robot touch," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, pp. 9382–9388, 2022.

[8] A. Younes, D. Honerkamp, T. Welschehold, and A. Valada, "Catch me if you hear me: Audio-visual navigation in complex unmapped environments with moving sounds," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 928–935, 2023.

[9] J. V. Hurtado and A. Valada, "Semantic scene segmentation for robotics," in *Deep learning for robot perception and cognition*, 2022, pp. 279–311.

[10] L. Rustler, J. Lundell, J. K. Behrens, V. Kyrki, and M. Hoffmann, "Active Visuo-Haptic Object Shape Completion," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 5254–5261, 2022.

[11] E. Smith, R. Calandra, A. Romero, G. Gkioxari, D. Meger, J. Malik, and M. Drozdzal, "3D Shape Reconstruction from Vision and Touch," in *Proc. Adv. Neural Inform. Process. Syst.*, 2020.

[12] P. K. Murali, B. Porr, and M. Kaboli, "Touch if it's transparent! ACTOR: Active Tactile-based Category-Level Transparent Object Reconstruction," in *Proc. IEEE Int. Conf. on Intel. Rob. and Syst.*, 2023.

[13] P. Falco, S. Lu, A. Cirillo, C. Natale, S. Pirozzi, and D. Lee, "Cross-modal visuo-tactile object recognition using robotic active exploration," in *Proc. IEEE Int. Conf. on Rob. and Auto.*, 2017.

[14] P. Falco, S. Lu, C. Natale, S. Pirozzi, and D. Lee, "A transfer learning approach to cross-modal object recognition: from visual observation to robotic haptic exploration," *IEEE Trans. Robot.*, 2019.

[15] M. A. Lee, Y. Zhu, P. Zachares, M. Tan, K. Srinivasan, S. Savarese, L. Fei-Fei, A. Garg, and J. Bohg, "Making Sense of Vision and Touch: Learning Multimodal Representations for Contact-Rich Tasks," *IEEE Trans. on Robotics*, vol. 36, no. 3, 2020.

[16] S. Zhong, A. Albini, O. P. Jones, P. Maiolino, and I. Posner, "Touching a NeRF: Leveraging Neural Radiance Fields for Tactile Sensory Data Generation," in *Proc. Conf. on Rob. Learn.*, 2023, pp. 1618–1628.

[17] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[18] J.-T. Lee, D. Bollegala, and S. Luo, ""Touching to See" and "Seeing to Feel": Robotic Cross-modal Sensory Data Generation for Visual-Tactile Perception," in *Proc. IEEE Int. Conf. on Rob. and Auto.*, 2019.

[19] F. Yang, J. Zhang, and A. Owens, "Generating visual scenes from touch," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2023, pp. 22 070–22 080.

[20] F. Yang, C. Feng, Z. Chen, H. Park, D. Wang, Y. Dou, Z. Zeng, X. Chen, R. Gangopadhyay, A. Owens, and A. Wong, "Binding Touch to Everything: Learning Unified Multimodal Tactile Representations," *arXiv preprint arXiv:2401.18084*, 2024.