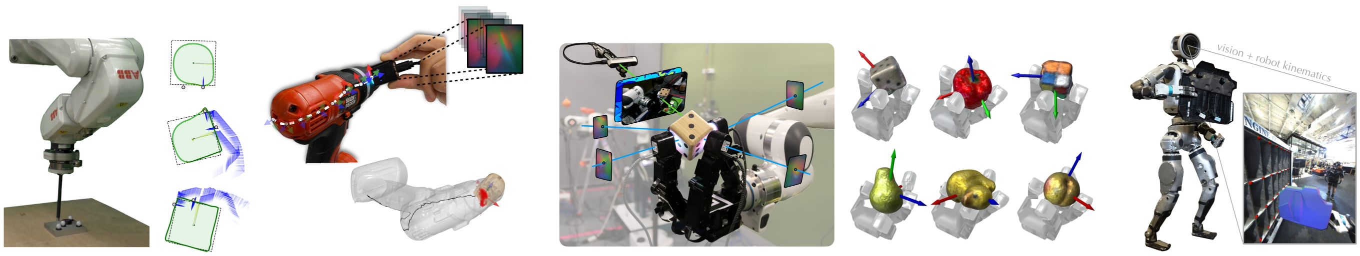
Perception amidst interaction: vision and touch enable useful manipulation

Sudharshan Suresh*
Boston Dynamics



(a) Tactile SLAM (b) MidasTouch (c) NeuralFeels (d) Humanoid perception

Fig. 1: My research is towards developing the core competencies of embodied perception in robots. These include methods that: **(a)** fuse localized touch information with physics constraints [1], **(b)** predict finger pose from touch [2], **(c)** create a learned representation of objects [3], and **(d)** combine vision with robot kinematics for humanoid manipulation [4].

I. Introduction

Robots currently lack the cognition to replicate even a fraction of the tasks humans do, a trend summarized by Moravec's Paradox [7]. Humans effortlessly combine their senses for everyday interactions—we can rummage through our pockets in search of our keys, and deftly insert them to unlock our front door. Before robots can demonstrate such dexterity, they must first be aware of the objects they manipulate. Unstructured environments with novel objects present challenges across robot perception, learning, and control.

In perception, knowledge of object pose and shape is crucial for downstream policy learning [6, 8]. The status quo for in-hand perception is restricted to tracking known objects with vision as the dominant modality [8], or circumventing the problem via fiducials [9, 10]. Moreover, vision fails in regimes where occlusion is imminent—like rotating [3, 11], re-orienting [8, 12], and sliding [2, 13]. Touch provides a local window into these interactions, but a general technique for visuo-tactile estimation remains an open question [14].

Alongside these challenges, advances in tactile sensing, rendering, and computer vision make this an opportune time to pursue this direction. First, vision-based touch sensors—like the GelSight and DIGIT [15–18]—provide spatial acuity at an affordable price. When chained with robot kinematics, they give dense, situated contact that can be processed similar to natural camera images. Second, touch simulation with realistic rendering [19, 20] enables practitioners to learn tactile observation models. Third, the progress in computer vision [21, 22] sets us up to transfer these ideas towards multimodal problems.

Research goals: In my research, I look at how we can leverage multimodal data—vision, touch, and proprioception—to unlock object manipulation capabilities. I operate under the constraints of robots in the wild: **(i)** causal perception *i.e.*, no access to future information, **(ii)** lack of fiducials and motion capture, **(iii)** noisy, occluded multimodal sensing, and **(iv)** apriori unknown objects. This is towards the long-term goal of robot dexterity: such that vision can locate a mug on the cluttered counter-top while touch can singulate the contours of the handle for a firm grasp.

*Email: suddhus@gmail.com; [1–3, 5, 6] is work done by the author while affiliated with Carnegie Mellon University and/or FAIR, Meta

II. Thesis research

My research studies **(i)** spatial representations for object-centric SLAM, **(ii)** tactile perception and simulation, and **(iii)** combining learned models with online optimization. I began with exploring how to fuse localized touch information with physics to reason about objects (Sec. II-A). Subsequently, I worked with high-dimensional touch sensors [17], focusing on developing a learned representation for pose estimation (Sec. II-B). Drawing upon these efforts, I developed a neural representation for in-hand perception, which unified vision, touch, and proprioception data (Sec. II-C). Finally, in Sec. II-D, I discuss my research in industry, that extends these ideas to drive humanoid manipulation problems.

A. Tactile SLAM: shape and pose from pushing [1]

When humans rummage a bag blindfolded, we can delineate objects just through tactile cues [23]. This is challenging for robots as, unlike vision, touch cannot provide global context about an object, but only localized information. This is analogous to the simultaneous localization and mapping (SLAM) problem for mobile robots, but rather by fusing force and contact measurements over time. To demonstrate this, I present a method that predicts both object shape and pose through a stream of tactile data from pushing [1]. Prior methods are restricted to constrained settings [24] or simple batch optimizations that have access to future data [25]. Our method combines surface contact information from the robot,

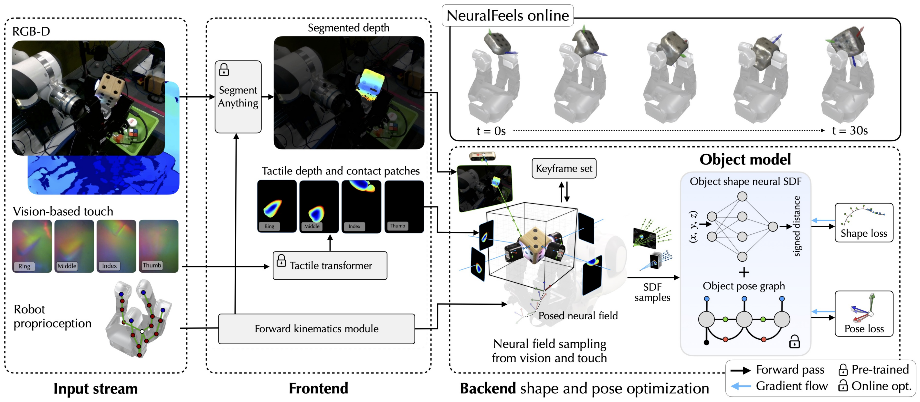


Fig. 2: **Perception stack with vision and touch.** In Neural Feels [3], an object-centric representation is learned from vision, touch, and proprioception. Sensor data is first fed into the *frontend*, which extracts visuo-tactile depth with pre-trained models. The *backend* samples from this depth to train a neural signed distance field (SDF), while the pose graph tracks the object's pose online. Through this combination of vision and touch, we accurately infer and track a novel object through in-hand rotation.

with quasi-static pushing constraints [26] with force-torque measurements (Fig. 1a). In a follow-up work, I extend this to perform 3D shape reconstruction of objects through touch [5].

B. *MidasTouch: learning pose estimation from touch [2]*

Along with shape reconstruction, robots can reason through touch *where* they make contact with the objects. Consider grasping a mug: with a curved body, flat base, rounded handle, and sharp lip. Without global context, a single-touch is ambiguous: a detected sharp edge could lie anywhere along the lip of the mug. Such a likelihood distribution is spread across the object's surface and not unimodal, but interaction over time can disambiguate it.

To demonstrate this idea, I worked on MidasTouch, that accurately predicts where we make contact with a known object through touch (Fig. 1b). Alongside, I open-sourced *YCB-Slide*, the largest tactile perception dataset with annotated poses. Given the small form-factor of vision-based touch sensors, prior methods have been restricted to small parts [27] or local tracking [28]. Our experiments demonstrate the surprising effectiveness of pairing learned tactile embeddings with Monte-Carlo methods to resolve any pose distribution ambiguities. This mirrors haptic *apprehension*, or the exploration humans perform when presented with a familiar object [29].

C. *NeuralFeels: Multimodal in-hand perception [3]*

With NeuralFeels, I put together these threads of work to build a multimodal perception system for in-hand manipulation. My goal was to present the robot with a novel object, and for it to infer and tracks its geometry through just interaction. In the work, I unify vision, touch, and proprioception into a neural representation and demonstrate SLAM for novel objects, and robust tracking of known objects (Fig. 1c). The algorithm is built on a dexterous hand [30] retrofit with vision-based touch sensors [17] and a fixed RGB-D camera. To explore the objects, we train a proprioception-driven policy in simulation for stable, in-hand rotation [11].

Over 70 rotation experiments, we show high-accuracy reconstructions and average pose drifts of 4.7mm, further reduced to 2.3mm with known object models. The chosen objects are sized between 6-18 cm in diagonal length. Additionally, under heavy visual occlusion we can achieve up to 94% improvements in tracking compared to vision-only methods. These results demonstrate that touch, at the very least, refines and, at the very best, disambiguates visual estimates during in-hand manipulation. NeuralFeels requires fewer sensors for pose estimation than prior work [10]—the entire online learning pipeline is illustrated in Fig. 2.

D. *Humanoid perception: enabling useful manipulation [4]*

I build upon this research direction as an industry research scientist at Boston Dynamics. Here, I led machine learning (ML) research focused on creating perception models for humanoid manipulation. For a robot to autonomously execute the wide-array of tasks in the factory, it requires accurate knowledge of both its environment and the objects it manipulates. One of the tasks I worked on is known as part sequencing, where the robot must autonomously grasp, carry, and insert objects from one receptacle to another.

For successful grasp and insertion policies, the margins are just a few centimeters, so accurate knowledge of the hand-object interaction is crucial. However these objects are often occluded, the environment constantly evolves, and lighting conditions are challenging. Our research developed vision-based ML models to accurately estimate the 3D pose of these objects from egocentric camera data. We combine these low-rate predictions, with force and proprioceptive robot data for consistent, real-time tracks for our manipulation policy [4].

III. FUTURE RESEARCH

While my thesis focused on object geometry, this work only scratches the surface of the broader unsolved challenges in visuo-tactile perception. As sensing standardizes [18, 31], the sim-to-real gap is narrowing, enabling the development of reliable simulators for touch-based policy learning [6, 32]. Moreover, interaction reveals properties like texture [33], friction [34], and object dynamics [35] that are imperceptible to cameras. With growing real-world touch datasets [2, 3, 36], researchers can train tactile representations [37] that have the potential to predict these latent properties.

Further, neural fields show promise in conjunction with multimodal sensing, as researchers explore high-fidelity, sample-efficient representations [38, 39]. The scope of multimodal sensing is growing with contact microphones [36, 40], heat, and even vibrations [31] actively being used by the robot learning community. Progress in learning from egocentric vision combined with touch sensing will drive the dexterous manipulation policies of the future.

References

[1] **Sudharshan Suresh**, Maria Bauza, Kuan-Ting Yu, Joshua G. Mangelson, Alberto Rodriguez, and Michael Kaess. Tactile SLAM: Real-time inference of shape and pose from planar pushing. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, May 2021.

[2] **Sudharshan Suresh**, Zilin Si, Stuart Anderson, Michael Kaess, and Mustafa Mukadam. MidasTouch: Monte-Carlo inference over distributions across sliding touch. In *Proc. Conf. on Robot Learning (CoRL)*. PMLR, 2022.

[3] **Sudharshan Suresh**, Haozhi Qi, Tingfan Wu, Taosha Fan, Luis Pineda, Mike Lambeta, Jitendra Malik, Mrinal Kalakrishnan, Roberto Calandra, Michael Kaess, Joseph Ortiz, and Mustafa Mukadam. Neural feels with neural fields: Visuo-tactile perception for in-hand manipulation. In *Science Robotics*, volume 9, page eadl0628, 2024.

[4] **Boston Dynamics**. Getting real with humanoids, 2025. URL bostondynamics.com/blog/getting-real-with-humanoids.

[5] **Sudharshan Suresh**, Zilin Si, Joshua G. Mangelson, Wenzhen Yuan, and Michael Kaess. ShapeMap 3-D: Efficient shape mapping through dense touch and vision. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, Philadelphia, PA, USA, May 2022.

[6] Haozhi Qi, Brent Yi, **Sudharshan Suresh**, Mike Lambeta, Yi Ma, Roberto Calandra, and Jitendra Malik. General in-hand object rotation with vision and touch. In *Proc. Conf. on Robot Learning (CoRL)*. PMLR, 2023.

[7] Hans Moravec. *Mind children: The future of robot and human intelligence*. Harvard University Press, 1988.

[8] Ankur Handa, Arthur Allshire, Viktor Makoviychuk, Aleksei Petrenko, Ritvik Singh, Jingzhou Liu, Denys Makoviichuk, Karl Van Wyk, Alexander Zhurkevich, Balakumar Sundaralingam, Yashraj Narang, Jean-Francois Lafleche, Dieter Fox, and Gavriel State. DeXtreme: Transfer of agile in-hand manipulation from simulation to reality. *arXiv*, 2022.

[9] OpenAI, Marcin Andrychowicz, Bowen Baker, Maciek Chociej, Rafał Józefowicz, Bob McGrew, Jakub Pachocki, Arthur Petron, Matthias Plappert, Glenn Powell, Alex Ray, Jonas Schneider, Szymon Sidor, Josh Tobin, Peter Welinder, Lilian Weng, and Wojciech Zaremba. Learning dexterous in-hand manipulation. *CoRR*, 2018.

[10] OpenAI, Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving Rubik's Cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

[11] Haozhi Qi, Ashish Kumar, Roberto Calandra, Yi Ma, and Jitendra Malik. In-hand object rotation via rapid motor adaptation. In *Conference on Robot Learning*, pages 1722–1732. PMLR, 2022.

[12] Tao Chen, Megha Tippur, Siyang Wu, Vikash Kumar, Edward Adelson, and Pulkit Agrawal. Visual dexterity: In-hand dexterous manipulation from depth. In *ICML Workshop on New Frontiers in Learning, Control, and Dynamical Systems*, 2023.

[13] Yu She, Shaoxiong Wang, Siyuan Dong, Neha Sunil, Alberto Rodriguez, and Edward Adelson. Cable manipulation with a tactile-reactive gripper. *Intl. J. of Robotics Research (IJRR)*, 40 (12-14):1385–1401, 2021.

[14] Shan Luo, Joao Bimbo, Ravinder Dahiya, and Hongbin Liu. Robotic tactile perception of object properties: A review. *Mechatronics*, 48:54–67, 2017.

[15] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. GelSight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

[16] Akhil Padmanabha, Frederik Ebert, Stephen Tian, Roberto Calandra, Chelsea Finn, and Sergey Levine. OmniTact: A multi-directional high-resolution touch sensor. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 618–624. IEEE, 2020.

[17] Mike Lambeta, Po-Wei Chou, Stephen Tian, Brian Yang, Benjamin Maloon, Victoria Rose Most, Dave Stroud, Raymond Santos, Ahmad Byagowi, Gregg Kammerer, 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 (RA-L)*, 5(3):3838–3845, 2020.

[18] Gelsight. Gelsight mini, 2025. URL www.gelsight.com/gelsightmini.

[19] Shaoxiong Wang, Mike Maroje Lambeta, Po-Wei Chou, and Roberto Calandra. TACTO: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors. *IEEE Robotics and Automation Letters (RA-L)*, 2022.

[20] Zilin Si and Wenzhen Yuan. Taxim: An example-based simulation model for gelsight tactile sensors. *IEEE Robotics and Automation Letters (RA-L)*, 2022.

[21] Joseph Ortiz, Alexander Clegg, Jing Dong, Edgar Sucar, David Novotny, Michael Zollhoefer, and Mustafa Mukadam. iSDF: Real-time neural signed distance fields for robot perception. *arXiv preprint arXiv:2204.02296*, 2022.

[22] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (ToG)*, 42(4): 1–14, 2023.

[23] Roberta L Klatzky, Susan J Lederman, and Victoria A Metzger. Identifying objects by touch: An expert system. *Perception & Psychophysics*, 37(4):299–302, 1985.

[24] Mark Moll and Michael A Erdmann. Reconstructing the shape and motion of unknown objects with active tactile sensors. In *Algorithmic Foundations of Robotics V*, pages 293–309. Springer, 2004.

[25] Kuan-Ting Yu, John Leonard, and Alberto Rodriguez. Shape and pose recovery from planar pushing. In *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, pages 1208–1215. IEEE, 2015.

[26] Kevin M Lynch, Hitoshi Maekawa, and Kazuo Tanie. Manipulation and active sensing by pushing using tactile feedback. In *Proc. IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, volume 1, 1992.

[27] Maria Bauza, Antonia Bronars, and Alberto Rodriguez. Tac2Pose: Tactile object pose estimation from the first touch. *arXiv preprint arXiv:2204.11701*, 2022.

[28] Paloma Sodhi, Michael Kaess, Mustafa Mukadam, and Stuart Anderson. Learning tactile models for factor graph-based estimation. In *Proc. IEEE Intl. Conf. on Robotics and Automation (ICRA)*, pages 13686–13692. IEEE, 2021.

[29] Susan J Lederman and Roberta L Klatzky. Hand movements: A window into haptic object recognition. *Cognitive psychology*, 19(3):342–368, 1987.

[30] Wonik Robotics. Allegro Hand, 2023. URL http://wiki.wonikrobotics.com/AllegroHandWiki/index.php/Allegro_Hand.

[31] Mike Lambeta, Tingfan Wu, Ali Sengul, Victoria Rose Most, Nolan Black, Kevin Sawyer, Romeo Mercado, Haozhi Qi, Alexander Sohn, Byron Taylor, et al. Digitizing touch with an artificial multimodal fingertip. *arXiv preprint arXiv:2411.02479*, 2024.

[32] Zhao-Heng Yin, Binghao Huang, Yuzhe Qin, Qifeng Chen, and Xiaolong Wang. Rotating without seeing: Towards in-hand dexterity through touch. *arXiv preprint arXiv:2303.10880*, 2023.

[33] Justin Kerr, Huang Huang, Albert Wilcox, Ryan Hoque, Jeffrey Ichnowski, Roberto Calandra, and Ken Goldberg. Learning self-supervised representations from vision and touch for active sliding perception of deformable surfaces. *arXiv preprint arXiv:2209.13042*, 2022.

[34] Simon Le Cleac'h, Hong-Xing Yu, Michelle Guo, Taylor Howell, Ruohan Gao, Jiajun Wu, Zachary Manchester, and Mac Schwager. Differentiable physics simulation of dynamics-augmented neural objects. *IEEE Robotics and Automation Letters*, 8(5):2780–2787, 2023.

[35] Balakumar Sundaralingam and Tucker Hermans. In-hand object-dynamics inference using tactile fingertips. *IEEE Transactions on Robotics*, 37(4):1115–1126, 2021.

[36] Ruohan Gao, Zilin Si, Yen-Yu Chang, Samuel Clarke, Jeannette Bohg, Li Fei-Fei, Wenzhen Yuan, and Jiajun Wu. ObjectFolder 2.0: A multisensory object dataset for Sim2Real transfer. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[37] Carolina Higuera, Akash Sharma, Chaithanya Krishna Bodduluri, Taosha Fan, Patrick Lancaster, Mrinal Kalakrishnan, Michael Kaess, Byron Boots, Mike Lambeta, Tingfan Wu, et al. Sparsh: Self-supervised touch representations for vision-based tactile sensing. *arXiv preprint arXiv:2410.24090*, 2024.

[38] Mauro Comi, Alessio Tonioni, Max Yang, Jonathan Tremblay, Valts Blukis, Yijiong Lin, Nathan F Lepora, and Laurence Aitchison. Snap-it, tap-it, splat-it: Tactile-informed 3D Gaussian splatting for reconstructing challenging surfaces. *arXiv preprint arXiv:2403.20275*, 2024.

[39] Aiden Swann, Matthew Strong, Won Kyung Do, Gadiel Sznaier Camps, Mac Schwager, and Monroe Kennedy. Touch-GS: Visual-tactile supervised 3D Gaussian splatting. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 10511–10518. IEEE, 2024.

[40] Samuel Clarke, Suzannah Wistreich, Yanjie Ze, and Jiajun Wu. X-capture: An open-source portable device for multi-sensory learning. *arXiv preprint arXiv:2504.02318*, 2025.