Implicit Neural Representation for 3D Shape Reconstruction Using Vision-Based Tactile Sensing

Mauro Comi[1,2], Alex Church[1,2], Kejie Li[3], Laurence Aitchison[1], Nathan F. Lepora[1,2]
[1]University of Bristol, [2]Bristol Robotics Laboratory, [3]University of Oxford

***Abstract*—Humans rely on their senses of vision and touch to build a 3D understanding of their physical surroundings. This understanding is critical in various fields of robotics, including those related to dexterity. Recently, there has been a growing interest in exploring and manipulating objects using data-driven approaches that utilise high-resolution vision-based tactile sensors. However, existing techniques present some limitations, such as the inability to reconstruct concave surfaces, difficulty in generalising over unseen shapes, and lack of flexibility due to the use of discrete data structures. To address these issues, we propose a Deep Learning approach for 3D shape reconstruction that leverages the rich information provided by the open sourced vision-based tactile sensor TacTip and the expressivity of the continuous implicit neural representation DeepSDF. Our technique consists of two components: (1) a Convolutional Neural Network that maps tactile images into local meshes representing the surface at the touch location, and (2) an implicit neural function based on DeepSDF that predicts a signed distance function to extract the desired 3D shape. Our approach demonstrates promising capacity in reconstructing a 3D shape from touch inputs. We believe that shape understanding will contribute to a larger effort to the development of safe and robust robot learning algorithms for physical world interaction using tactile sensing.**

***Index Terms*—tactile sensing, robot perception, 3D reconstruction, neural fields**

I. Introduction

The current state of 3D shape reconstruction research is primarily concerned with the sense of vision [1] [2]. Recently, data-driven methodologies that utilise vision-based tactile sensors have been proposed as a way to improve object exploration and manipulation tasks [3]. Compared to solely camera-based sensing, these tactile sensors provide a range of benefits, such as the ability to capture detailed contact information, and being effective even when an object is occluded. Additionally, they simplify the translation of simulated tasks to real-world scenarios as they require a smaller observation space [4].

Smith et al. [5] proposed a technique that leverages the rich information provided by vision-based tactile sensors for 3D shape reconstruction. The DIGIT tactile sensor [6] is employed in this method, but its flat design restricts its effectiveness in handling objects with concave surfaces. Additionally, the Graph Convolutional Network (GCN) [7] utilised for mesh refinement often results in suboptimal reconstructions due to its reliance on discrete data structures.

M. Comi was supported by the UK Research and Innovation (UKRI). N. Lepora was supported by a Leadership Award from the Leverhulme Trust on 'A biomimetic forebrain for robot touch' (RL-2016-39). Corresponding author: mauro.comi@bristol.ac.uk

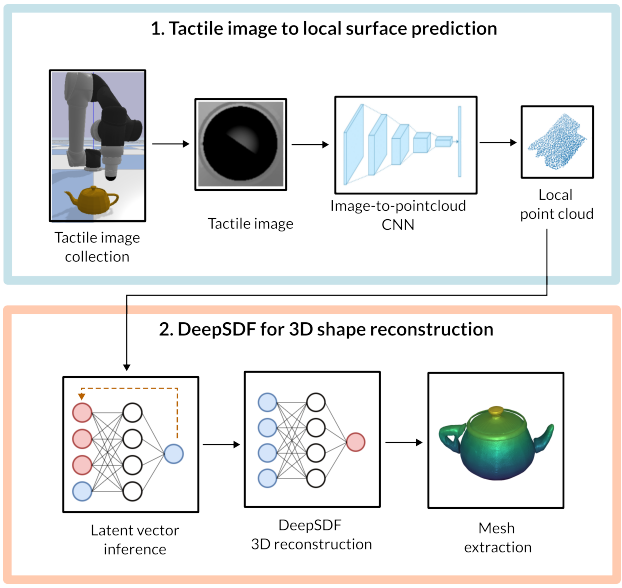

Fig. 1. An overview of our DeepSDF-based two-step process for 3D object reconstruction. First, a manipulation robot samples the surface of the object to acquire a 2D tactile image. Next, a convolutional neural network maps the 2D image into a set of 3D points corresponding to the local surface of the object where the touch occurred. A pre-trained DeepSDF algorithm uses these point clouds to predict a continuous signed distance function that describes the shape of the object.

Continuous implicit representations have gained attention due to their ability to encode extensive prior information about the continuous space of 3D shapes. DeepSDF [8] is an implicit neural function that maps a point in 3D space to the signed Euclidean distance between the point and the boundary of an object. The surface where the signed distance function is zero is a continuous and smooth surface that is representative of the shape of the object. This model can be optionally conditioned on a latent vector to encode an arbitrary number of shapes. DeepSDF has also demonstrated the ability to reconstruct 3D geometry from partial point clouds, which is useful in our context as tactile sensors provide partial observations of an object.

Main contribution: In this work, we present a 3D shape reconstruction approach that uses solely vision-based tactile sensing, specifically the 3D printed biomimetic vision-based tactile sensor TacTip [9]. Compared to previous methods, we employ an implicit neural representations that encodes a smooth and continuous surface, rather than predicting a set of

independent meshes. Specifically, our framework combines a Convolutional Neural Network (CNN) for local surface reconstruction [5] with a DeepSDF model for mesh reconstruction. The objects used for training these two models are a subset of the ShapeNetCorev2 dataset [10]. The TacTip's soft domed structure can correctly capture geometries that may not be accurately handled by flat sensors. This approach is fully integrated into Tactile Gym [4] [3], a robot learning suite for manipulation tasks designed to bridge the sim-to-real gap.

II. Related work

A. Vision-based tactile sensing for 3D shape reconstruction

The development of low-cost, open-source, and robust vision-based tactile sensors has opened new avenues of study in 3D reconstruction over the past few years. To the best of our knowledge, [11] proposed the first work on active touch exploration for object reconstruction that employs vision-based tactile sensing (DIGIT sensor). However, this technique heavily relies on shape priors learned by the vision model, and predicts a voxel-based reconstruction that is not suitable for retrieving fine details. In contrast, Smith et al. [12] [5] proposed a different approach that decouples vision and tactile sensing. Their method combines two novel reconstruction models. The first model is a CNN that maps a tactile image into a mesh representing the local 3D surface at the touch location. The second model consists of a series of GCNs that deform an initial spherical mesh into the desired object. This is achieved by predicting the coordinates of a discrete number of independent meshes whose face ordering is fixed. The resulting reconstruction is a collection of independent meshes instead of a single object, which displays sharp features due to the use of fixed face ordering during the mesh refinement process. In contrast, our approach results in a smooth and continuous surface representing the object's shape.

B. Implicit neural representations in robotic manipulation

Neural radiance fields (NeRFs) [13] have become increasingly popular as an implicit representation in vision, graphics, and robotics. NeRF is a function parameterised by a neural network that encodes the occupancy of a 3D scene, similar to DeepSDF. However, NeRF also encodes the radiance field of a scene, enabling the generation of novel views from any viewpoint.

An interesting application of NeRFs in vision-based tactile sensing is the generation of synthetic tactile images. Gao et al. [14] [15] developed a NeRF-like model to generate tactile images of single objects, while Zhong et al. [16] proposed a similar approach that can generate tactile images for previously unseen objects. These approaches rely on NeRF's ability to encode 3D geometry but are not explicitly used for 3D reconstruction.

Another area of research involving NeRFs is object-centric 3D reconstruction. In contrast to previous methods that require multiple views, CodeNeRF [17] and AutoRF [18] learn a radiance field from a single or few images. Additionally, these methods represent multiple objects by conditioning on per-object latent codes. Building on these approaches, [19] propose a unified representation for 3D reconstruction and grasp pose prediction from a single view.

NeRF-based approaches are promising for applications requiring high-quality scene appearance prediction. However, these methods are computationally expensive, although efforts have been directed to reduce the time and data required for training [20] [21]. For applications that require 3D reconstruction from tactile sensors, object appearance representation may not be necessary. Instead, an accurate representation of the object's 3D geometry is essential, as well as a relatively fast inference time. To achieve this, our proposed approach relies on DeepSDF, which is a more efficient method for encoding and reconstructing 3D surfaces from partial observations. Compared to NeRF-based approaches, our method is less computationally expensive and requires less training data, making it more practical for real-world applications.

III. Methodology

To ensure the reliability of our reconstruction approach, we utilised 600 objects that were randomly sampled from the ShapeNetCoreV2 dataset for both training and evaluation purposes. To ensure that DeepSDF can accurately extract and predict a consistent signed distance function, each object was processed to create watertight meshes, which are closed surfaces that don't have any gaps or holes. The reconstructing process through tactile images sampled on the object's surface involves a two-step procedure (Fig. 1):

1) Tactile images to point cloud prediction: A simulated manipulation robot, specifically a UR5 robot equipped with the tactile sensor TacTip, samples the surface of the object. The data collection procedure uses Tactile Gym, a robotic learning platform developed on top of the PyBullet simulator. The touch results in the acquisition of a single channel 2D tactile image that represents the surface depth map. To extract the local surface of the object at touch location, a convolutional neural network (CNN) is employed to map the image into a set of 3D points defined in the sensor's frame. The CNN is trained to minimise the Chamfer Distance [22] between the predicted and observed point clouds. This model is inspired by the touch chart prediction model proposed in [5].

2) 3D shape reconstruction using DeepSDF: The point cloud generated by the CNN is used as input to a DeepSDF-based shape prediction model. DeepSDF employs a deep neural network to learn a continuous signed distance function that describes the shape of an object. The chosen architecture is the one originally proposed by [8]. When conditioned on a partial point cloud of the object's surface, the pre-trained DeepSDF infers a 128-dimensional latent vector that best represents the partial point cloud. This allows the model to encode a wide variety of objects and to reconstruct unseen shapes through interpolation in the latent space. The inferred latent vector is then used by the DeepSDF model to predict the signed distance function of the volume surrounding the object. Finally, the Marching Cubes algorithm is applied to

extract the zero-level set of the signed distance function, which corresponds to the 3D shape of the object.

IV. Results

Fig. 2 displays the point clouds (on the right) predicted from the collected tactile images (on the left) and overlaid onto an object. The reconstruction model is trained only on the predicted point clouds, while the object point cloud is presented only for visualisation purposes. The colour coding in the point clouds represents the distance from the object surface, where blue indicates a positive signed distance function and red indicates a negative distance. By utilising both positive and negative samples, we increased the robustness of the latent code inference procedure.

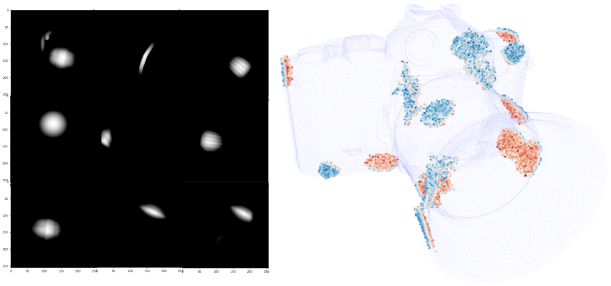

Fig. 2. On the left, collected tactile images on the surface of the object. On the right, the predicted point clouds used to reconstruct the entire object.

Fig. 3 shows multiple examples of reconstructed shapes at increasing numbers of touches. For each set of collected images, the corresponding point clouds are predicted. The DeepSDF model is conditioned on these predictions to infer the shape latent code, which is then used to extract the reconstructed mesh. Our results indicate that increasing the number of touches improves the quality of the reconstruction.

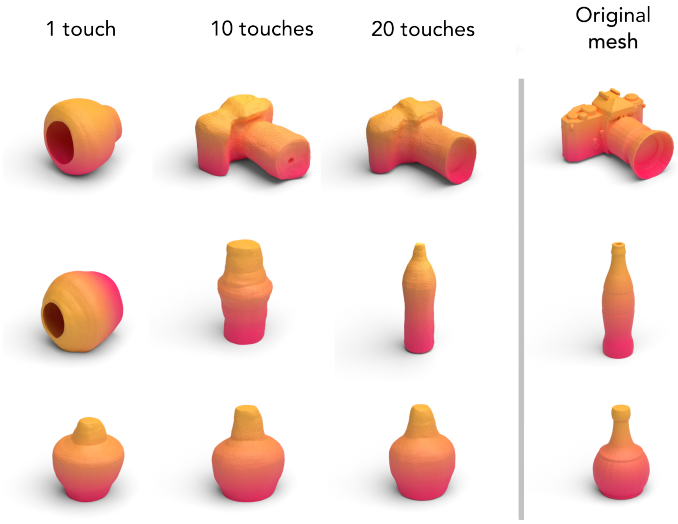

Fig. 3. Reconstruction results of our method across different number of touches.

V. Further Work and Conclusion

In this study, we investigated the potential of the implicit neural representation DeepSDF in encoding multiple shapes and reconstructing them from partial point clouds generated by tactile sensors. Our qualitative results demonstrate that DeepSDF can successfully reconstruct objects from partial and scattered point clouds, and that the reconstruction quality improves as the number of touch points increases.

In the next phase of our research, we plan to improve the DeepSDF algorithm to develop a full pipeline capable of reconstructing objects based on simulated touches. We will test this pipeline on a real robot using the sim-to-real approach described in [4], which has already been implemented in Tactile Gym. Additionally, support for multiple sensors, such as GelSight-based sensors, will be added.

Our proposed approach will contribute to the development of effective and reliable multi-modal robot learning algorithms that incorporate shape understanding to enhance safety and robustness when interacting with the physical world.

References

[1] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese, "3d-r2n2: A unified approach for single and multi-view 3d object reconstruction," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, pp. 628–644.

[2] S. Tulsiani, T. Zhou, A. A. Efros, and J. Malik, "Multi-view supervision for single-view reconstruction via differentiable ray consistency," pp. 2626–2634. [Online]. Available: http://arxiv.org/abs/1704.06254

[3] Y. Lin, J. Lloyd, A. Church, and N. F. Lepora, "Tactile gym 2.0: Sim-to-real deep reinforcement learning for comparing low-cost high-resolution robot touch," *IEEE Robotics and Automation Letters*, vol. 7, no. 4, p. 10754–10761, 2022.

[4] A. Church, J. Lloyd, N. F. Lepora, and others, "Tactile sim-to-real policy transfer via real-to-sim image translation," in *Conference on Robot Learning*. PMLR, pp. 1645–1654.

[5] E. Smith, D. Meger, L. Pineda, R. Calandra, J. Malik, A. Romero Soriano, and M. Drozdzal, "Active 3d shape reconstruction from vision and touch," vol. 34, pp. 16 064–16 078.

[6] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer, D. Jayaraman, and R. Calandra, "DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation," vol. 5, no. 3, pp. 3838–3845. [Online]. Available: http://arxiv.org/abs/2005.14679

[7] J. Bruna, W. Zaremba, A. Szlam, and Y. Lecun, "Spectral networks and locally connected networks on graphs international conference on learning representations (iclr2014)."

[8] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," pp. 165–174.

[9] N. F. Lepora, "Soft biomimetic optical tactile sensing with the TacTip: A review," vol. 21, no. 19, pp. 21 131–21 143. [Online]. Available: https://ieeexplore.ieee.org/document/9499032/

[10] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "ShapeNet: An Information-Rich 3D Model Repository," Stanford University — Princeton University — Toyota Technological Institute at Chicago, Tech. Rep. arXiv:1512.03012 [cs.GR], 2015.

[11] S. Wang, J. Wu, X. Sun, W. Yuan, W. T. Freeman, J. B. Tenenbaum, and E. H. Adelson, "3d shape perception from monocular vision, touch, and shape priors," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 1606–1613. [Online]. Available: https://ieeexplore.ieee.org/document/8593430/

[12] E. J. Smith, R. Calandra, A. Romero, G. Gkioxari, D. Meger, J. Malik, and M. Drozdzal, "3d shape reconstruction from vision and touch," vol. 33, pp. 14 193–14 206.

[13] 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.

[14] R. Gao, Y.-Y. Chang, S. Mall, L. Fei-Fei, and J. Wu, "Objectfolder: A dataset of objects with implicit visual, auditory, and tactile representations," *arXiv preprint arXiv:2109.07991*, 2021.

[15] R. Gao, Z. Si, Y.-Y. Chang, S. Clarke, J. Bohg, L. Fei-Fei, W. Yuan, and J. Wu, "Objectfolder 2.0: A multisensory object dataset for sim2real transfer," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 10 598–10 608.

[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 *6th Annual Conference on Robot Learning*, 2022.

[17] W. Jang and L. Agapito, "Codenerf: Disentangled neural radiance fields for object categories," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 12 949–12 958.

[18] N. Müller, A. Simonelli, L. Porzi, S. R. Bulò, M. Nießner, and P. Kontschieder, "Autorf: Learning 3d object radiance fields from single view observations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 3971–3980.

[19] V. Blukis, T. Lee, J. Tremblay, B. Wen, I. S. Kweon, K.-J. Yoon, D. Fox, and S. Birchfield, "Neural fields for robotic object manipulation from a single image," *arXiv preprint arXiv:2210.12126*, 2022.

[20] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Transactions on Graphics (ToG)*, vol. 41, no. 4, pp. 1–15, 2022.

[21] A. Yu, S. Fridovich-Keil, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, "Plenoxels: Radiance fields without neural networks," *arXiv preprint arXiv:2112.05131*, 2021.

[22] X. Sun, J. Wu, X. Zhang, Z. Zhang, C. Zhang, T. Xue, J. B. Tenenbaum, and W. T. Freeman, "Pix3d: Dataset and methods for single-image 3d shape modeling," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 2974–2983.