Neural Contact Fields: Tracking Extrinsic Contact with Tactile Sensing

Carolina Higuera
University of Washington

Siyuan Dong
University of Washington

Byron Boots
University of Washington

Mustafa Mukadam
Meta AI

Abstract: We present Neural Contact Fields (NCF), a method that brings together neural fields and tactile sensing to address the problem of tracking extrinsic contact between object and environment. Knowing where the external contact occurs is a first step towards agents that can plan trajectories over contacts for long-horizon manipulation tasks. Prior work typically assume a contact type (e.g. point or line), does not capture contact/no-contact transitions, and only works with basic geometric-shaped objects. NCF are the first method that can track arbitrary multi-modal extrinsic contacts without making any assumptions about the contact type. Our key insight is to estimate the probability of contact for any 3D point in the latent space of object's shapes, given vision-based tactile inputs that sense the local motion resulting from the external contact. In experiments, we find that Neural Contact Fields are able to localize multiple contact patches without making any assumptions about the geometry of the contact, and capture contact/no-contact transitions for known categories of objects with unseen shapes in unseen environment configurations.

Keywords: Neural Fields, Tactile Sensing, Extrinsic Contact

1 Introduction

We investigate the problem of tracking extrinsic contact between object and environment using tactile perception between robot hand and object. Consider the task of placing a book on a bookcase. While the book is in free space, no forces are experienced in the fingers. However, once the book makes external contact, it moves in accordance with the constraints of the contact and shear forces are perceived in the fingers. For a robot to accomplish this task, it could be beneficial to plan the sequences of external contacts over time. In such manipulation tasks tracking extrinsic contact with tactile sensing becomes critical for spatial understating. It is also a first step towards building methods that can then leverage as well as actively control extrinsic contacts in downstream policies.

Tracking extrinsic contacts is an understudied problem in literature and it is still an open question what is their best abstraction. Prior approaches are characterized by making hypothesis about the geometry of the contact [2] and requires the contact to be enforced [3]. We present Neural Contact Fields (NCF), the first method that can track arbitrary multi-modal extrinsic contact from tactile perception. Our key insight is to leverage neural fields to generalize across different object shapes, given the local motion produced by the contact and captured by vision-based tactile sensors, such as the DIGIT sensor [1]. Our method, illustrated in Figure 1 estimates the probability of external contact for any 3D point on an object surface given a sequence of tactile images and the most recent history of end-effector poses and external contact probabilities. We train NCF to track extrinsic contact on three categories of objects with simulated tactile data. In experiments, we find that NCF

6th Conference on Robot Learning (CoRL 2022), Auckland, New Zealand.

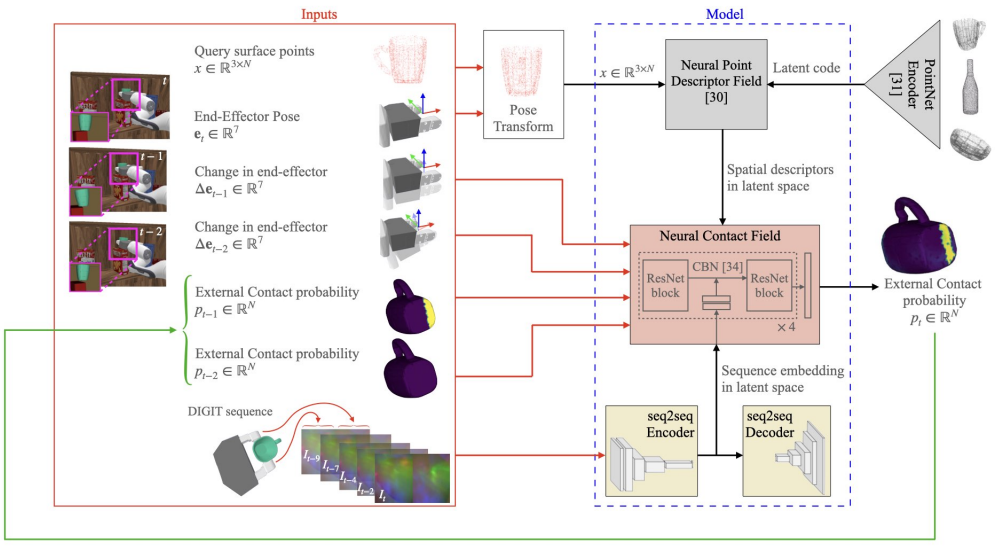

Figure 1: Neural Contact Fields (NCF) allows tracking extrinsic contact on an object surface (between object and environment) with vision-based tactile sensing with the DIGIT sensor [1] (between robot hand and object). Our model outputs extrinsic contact probability at the current timestep for a set of query points on the 3D surface of the object.

is able to localize multiple contact patches without making any assumptions about its geometry and can capture contact/no-contact transitions, on unseen shapes in unseen environments.

2 Tracking Extrinsic Contact

We focus on the problem of tracking extrinsic contact between object and environment. Assuming that the grasp pose is fixed (same for all objects of the same category), our goals are threefold: i) to track multiple contact patches without assuming the contact type, ii) capturing contact to no-contact and vice-versa transitions, and iii) to generalize contact tracking to complex and unseen object shapes and environments. Our proposed model consists of three modules:

- A *PointNet Encoder + Neural Point Descriptor Fields* that allow us to generalize across different object shapes.
- A *Sequence-to-Sequence Autoencoder* to extract the embedding from a sequence of tactile images.
- *Neural Contact Fields* (ours), which can estimate, in latent space, the probability of contact for any 3D point on an object surface.

2.1 Neural Point Descriptor Fields

Neural Point Descriptor Fields, proposed by Simeonov et al. [4], represent an object as a function that maps a 3D coordinate $\mathbf{x}$ to a spatial descriptor, based on the architecture of Occupancy Networks [5]. It uses the concatenation of all activations across layers of the occupancy network as the spatial descriptor $\mathbf{z}$ for a 3D point $\mathbf{x}$. Furthermore, the model is forced to parameterize the spatial descriptors grounded in the object's category. This is achieved by conditioning the model on different low-dimensional latent codes. These latent codes can be obtained as the output of a PointNet encoder $\mathcal{E}$ [6] that takes as input the point cloud $\mathbf{P}$ of the shape.

In our work, even though the contact constraints can cause motions of the object, for simplicity we consider that these are very small and in general, the object is rigidly grasped. Therefore, the con-

figuration of the object in world frame is subject to a rigid body transform $(\mathbf{R}, \mathbf{t}) \in SE(3)$, which in turn is determined by the end-effector pose. Neural point descriptor fields allow rotation equivariance by using an occupancy network equipped with Vector Neurons [7]. Translation equivariance is implemented by mean-centering the object's point cloud $\mathbf{P}$. In this way, our Neural Contact Fields work in the latent space of spatial descriptors $\mathbf{z} \in \mathbb{R}^n$:

$$\mathbf{z}_i = g(\mathbf{x}_i|\mathbf{P}) = g(\mathbf{R}\mathbf{x}_i + \mathbf{t}|\mathbf{R}\mathbf{P} + \mathbf{t}) \forall\ i \in \{0, N\} \tag{1}$$

2.2 Sequence-to-Sequence Autoencoder

In our setting, the object is grasped by a robot with a parallel gripper and tactile sensors in the fingers. If the grasp is rigid and no contact is happening, the pose of the object is completely determined by the end-effector pose. However, under external contact, the object might be pivoting or slipping slightly. This motion is well captured by vision-based tactile sensors.

We capture the information of the relative motion from a sequence $\mathbf{G}$ with k tactile images. First, we learn an autoencoder that allows us to capture a low-dimensional representation of raw 320×240 RGB tactile images per finger captured by a DIGIT sensor. Then, our sequence-to-sequence autoencoder uses as tokens this low-dimensional representation for both fingers concatenated with the end-effector pose. Encoder and decoder have a convolutional LSTM as recurrent network [8]. This setting allows us to learn $\mathcal{E}(\mathbf{G}) \in \mathbb{R}^m$, an implicit representation of tactile sequences in a self-supervised manner.

2.3 Neural Contact Fields

Our overall pipeline, shown in Figure 1 aims to represent the extrinsic contact of an object as a function that maps a 3D point $\mathbf{x}$ to a probability of contact, given the object's motion captured by a sequence $\mathbf{G}$ of tactile images from the gripper's fingers and a canonical point cloud description of the object $\mathbf{P}$:

$$f(\mathbf{x}, \mathbf{G}, \mathbf{P}) : \mathbb{R}^3 \times \mathbb{R}^{k \times 2 \times 320 \times 240 \times 3} \times \mathbb{R}^{3 \times N} \rightarrow [0, 1] \tag{2}$$

Neural Contact Fields (`NCF`) make it possible to track extrinsic contacts by estimating in latent space of spatial descriptors whether a 3D point $\mathbf{x}$ is making extrinsic contact or not. This is conditioned on the implicit representation of the motion of the object due to contact, which is captured by a history of tactile images:

$$f(\mathbf{z}, \mathcal{E}(\mathbf{G})) : \mathbb{R}^n \times \mathbb{R}^m \rightarrow [0, 1] \tag{3}$$

We also provide the `NCF` model with temporal information about the history of the object state, such as the probability for the queried points of being in contact $p_{t-1,t-2}$ and the change in end-effector pose with respect to the current one $\Delta\mathbf{e}_{t-1,t-2}$ for the last two timesteps. We feed our set of inputs through four fully-connected ResNet blocks with Conditional Batch-Normalization [9] to condition the network on the embedding of the object's motion sequence $\mathcal{E}(\mathbf{G})$. Finally, we use a fully-connected layer and apply the sigmoid as activation function to obtain contact probabilities for each 3D coordinate $\mathbf{x}$.

3 Evaluation

We use a pretrained implementation of Neural Point Descriptor Fields to get the spatial descriptors for 3D coordinates in the point cloud for mugs, bottles, and bowls. We trained separately the sequence-to-sequence autoencoder to learn the embedding of DIGIT image sequences. These sequences have length $k = 5$ and contain the images for both fingers at timesteps $t, t-2, t-4, t-7$ and $t-9$. For our `NCF` we trained four fully-connected ResNet blocks with Conditional Batch-Normalization on training trajectories that contain in total 4500 contact events. We use negative log-likelihood as loss function and Adam optimizer [10] with learning rate $1e^{-4}$.

In Figure 2 we show qualitative results of our pipeline tracking extrinsic contact. We show key snapshots of a test trajectory collected with unseen shapes of mugs, bottles, bowls, and scenarios

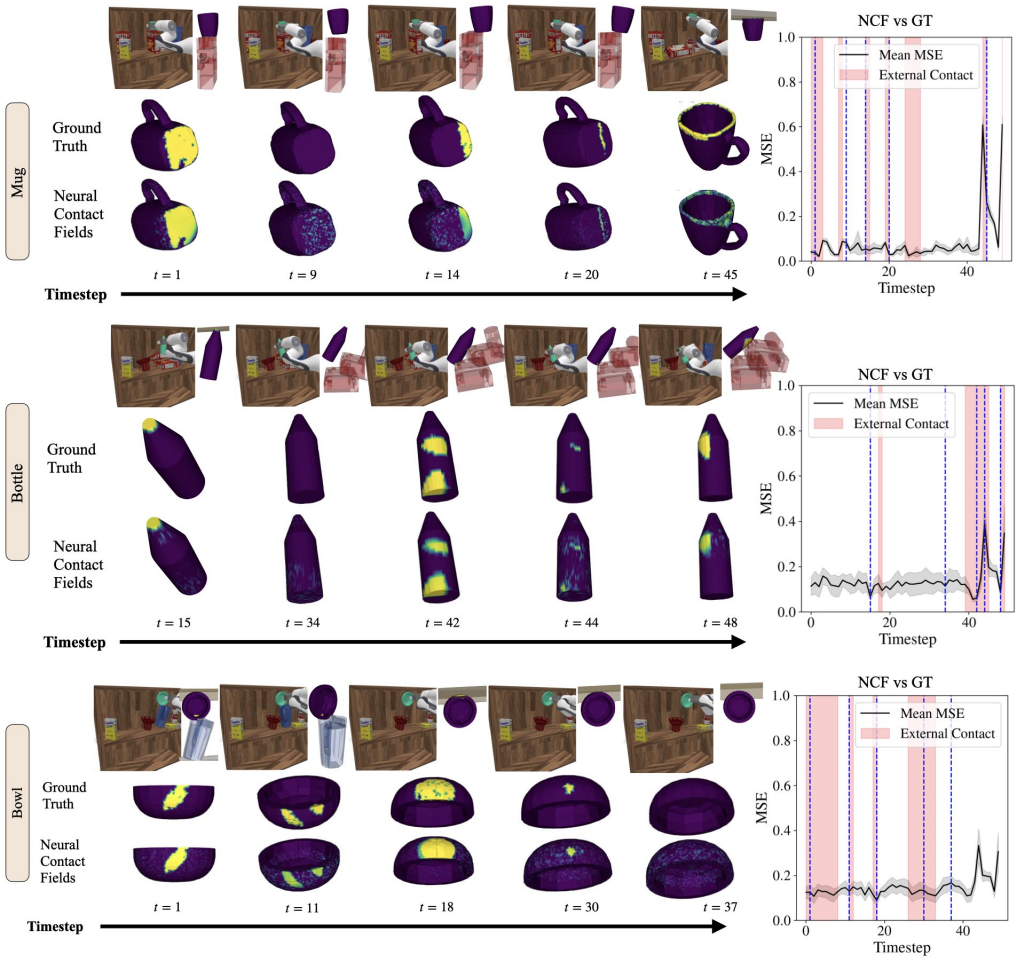

Figure 2: Snapshots of extrinsic contact predictions on simulated testing trajectories at key timesteps (blue dotted lines) for unseen shapes of mugs, bottles, and bowls. For each row: **[top]** Pybullet frame and zoom-in view of the contact interaction, **[middle]** extrinsic contact ground truth probabilities, **[bottom]** NCF prediction, and **[right]** MSE between the ground truth and estimated extrinsic contact probabilities for the trajectory over time.

with new configurations in simulation. From a qualitative point of view, we demonstrate the ability of NCF to track extrinsic contacts for the three novel objects. For example, with the mug trajectory, we show that the NCF can transition from patch to break contact to a new contact location. With the snapshots for the trajectories with the bottle and bowl, we show that NCF can localize multiple contact patches produced during complex contact interactions.

For a quantitative evaluation, we analyze the mean squared error (MSE) between the ground truth external contact probability and the predictions. At each timestep, we plot the MSE after running the model ten times and applying Monte Carlo Dropout to compute a 95% confidence interval. NCF is close to the ground truth during the majority of the trajectories. From these results, we identify cases that might induce a spike in error. For example, when the object transitions from no-contact to contact or when the shape of the contact patch changes drastically due to a non-smooth change in end-effector pose.

4 Conclusion

We introduced Neural Contact Fields (NCF), a novel method for tracking extrinsic contact. Our model outputs the probability of external contact for a set of query points on the 3D surface of an object based on vision-based tactile sensing. Our method works in the latent space of spatial

descriptors, which allows us to generalize within categories of objects with different shapes. Our experiments in simulation demonstrate the capability of NCF in localizing and tracking complex contact interactions, such as multiple contact patches, and breaking and making contact. We believe NCF can be used in downstream robot manipulation tasks and leave its applications for future work.

References

[1] 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. *IEEE Robotics and Automation Letters*, 5(3):3838–3845, 2020.

[2] D. Ma, S. Dong, and A. Rodriguez. Extrinsic contact sensing with relative-motion tracking from distributed tactile measurements. In *2021 IEEE international conference on robotics and automation (ICRA)*, pages 11262–11268. IEEE, 2021.

[3] S. Kim and A. Rodriguez. Active extrinsic contact sensing: Application to general peg-in-hole insertion. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 10241–10247, 2022. doi:10.1109/ICRA46639.2022.9812017.

[4] A. Simeonov, Y. Du, A. Tagliasacchi, J. B. Tenenbaum, A. Rodriguez, P. Agrawal, and V. Sitzmann. Neural descriptor fields: Se (3)-equivariant object representations for manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6394–6400. IEEE, 2022.

[5] L. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] C. R. Qi, H. Su, K. Mo, and L. J. Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 652–660, 2017.

[7] C. Deng, O. Litany, Y. Duan, A. Poulenard, A. Tagliasacchi, and L. J. Guibas. Vector neurons: A general framework for so (3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12200–12209, 2021.

[8] X. Shi, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. *Advances in neural information processing systems*, 28, 2015.

[9] H. De Vries, F. Strub, J. Mary, H. Larochelle, O. Pietquin, and A. C. Courville. Modulating early visual processing by language. *Advances in Neural Information Processing Systems*, 30, 2017.

[10] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.