2023 IEEE International Conference on Robotics and Automation (ICRA 2023)
May 29 - June 2, 2023. London, UK

Neural Contact Fields: Tracking Extrinsic Contact with Tactile Sensing

Carolina Higuera[1], Siyuan Dong[1], Byron Boots[1], and Mustafa Mukadam[2]
[1]University of Washington, [2]Meta AI

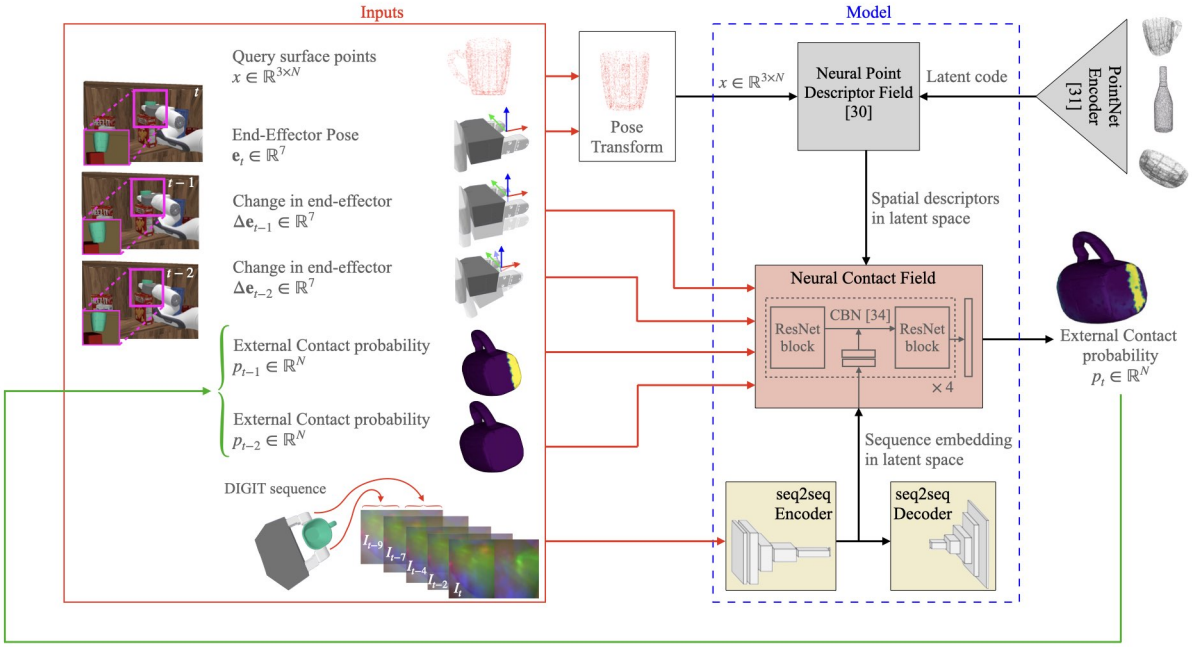

Fig. 1. Neural Contact Fields (NCF) is an implicit representation for tracking extrinsic contact on an object surface (between object and environment) with vision-based tactile sensing (between robot hand and object). Our model outputs extrinsic contact probability at the current timestep for a set of query points on the 3D surface of the object. This is done by taking a history of end-effector poses, tactile images and extrinsic contact probabilities as inputs.

***Abstract*— We present Neural Contact Fields, 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 methods that can actively control it in facilitating downstream manipulation tasks. Prior work for localizing environmental contacts 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. Neural Contact Fields 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. In addition to Neural Contact Fields, we also release our YCB-Extrinsic-Contact dataset of simulated extrinsic contact interactions to enable further research in this area. Project page: https://github.com/carolinahiguera/NCF**

I. 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 when the fingers are firmly grasping the book. While the book is in free space, no shear forces are experienced in the fingers. However, once the book makes external contact, for example if a corner touches the shelf, the book rotates in accordance with the kinematic and frictional constraints of the contact. As a result, shear forces are perceived in the fingers. In such manipulation tasks tracking extrinsic contact with tactile sensing becomes critical for spatial understating, since vision is heavily occluded. It is also a first step towards building methods that can then leverage as well as actively control extrinsic contacts in downstream policies.

Despite the valuable information that an extrinsic contact tracker can provide about object-environment interaction, for example for extrinsic dexterity [1], [2], this problem is understudied in literature. Prior approaches for localizing environmental contacts are characterized by making hypothesis about the geometry of the contact. For example, Ma et al. [3] formulate the problem as constrained optimization, where different constraints must be defined to solve for parameters that localize external contact, such as position of a point, direction of a line, or normal direction of a plane. Kim et al. [4] uses factor graphs and tactile measurements to actively estimate the contact line between the object and its environment for a peg-in-hole insertion task, while

2023 IEEE International Conference on Robotics and Automation (ICRA) | 979-8-3503-2365-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICRA48891.2023.10160526

having a controller that enforces the line contact. Although these works have started to explore the problem of extrinsic contact localization, several challenges remain. For instance, it is not clear: (i) how to track multi-modal contact patches without assuming a contact type; (ii) how to track contact to no-contact interactions, and vice versa; and (iii) how to generalize to complex object shapes and new environments.

In this work, we address these challenges with 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, which are well-captured by vision-based tactile sensors [5]. The full architecture of Neural Contact Fields is illustrated in Fig. 1. Our method 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 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.

In addition to open-sourcing Neural Contact Fields, we also release the YCB-Extrinsic-Contact dataset of simulated extrinsic contact interactions for three categories of objects in different environments. The dataset compiles information about end-effector poses, DIGIT tactile sensor images [5] from the TACTO simulator [6], and extrinsic contact ground-truth per time step of each training/testing trajectory. This dataset provides a simulation benchmark that we hope will enable further research in extrinsic contact perception.

II. Related Work

A. External contact localization

Although localizing external contacts can provide rich information to respond correctly in downstream tasks, such as insertion, packing, or assembling, this important problem has received little attention. Ma et al. [3] discuss a theoretical basis for localizing environmental contacts. They formulate the problem as a constraint-based estimation subject to the kinematic constraints imposed by tactile measurements of the object local motion, as well as the kinematic and frictional constraints imposed by rigid-body mechanics. This work presents a method to estimate three possible hypothesis of contact: point, line and patch contact. For each mode, different constraints must be formulated to solve for the parameters that allow to localize the external contact on the object, such as position of a point, direction of a line, or normal direction of a plane. To solve the problem, some assumptions must be guaranteed such as the object remains in contact with the environment and that the grasp is stable. The work acknowledges that a key limitation lies in having to assume a particular contact type to formulate the constraints and that a stronger implementation should consider multiple hypothesis for contact formations. In addition, [4] follows a similar approach, using factor graphs and tactile measurements to actively estimate the contact line between the object and its environment for a peg-in-hole insertion task. This method allows to parameterize the contact line, assuming that the object is always making contact, and generalizes well for objects with basic shapes and flat surfaces.

A learning based approach is presented in [7], in which a collision model is learned from point cloud of the object and its environment. The model predicts the likelihood that the object collides with the scene but does not provide further information about where is the contact located on the object. This method generalizes across four objects categories (mugs, cylinders, boxes, and bowls) with different scenes.

Our Neural Contact Fields generalize across three object categories, can localize multi-modal external contacts without making any assumptions about the contact type, and can capture contact/no-contact transitions.

B. Vision-based tactile sensing in robotics

Vision-based tactile sensing has been actively used in manipulation tasks given their ability to capture local contact events with high accuracy. Generally, these sensors are located at the gripper fingers, capturing the small forces and object displacements that allow to detect and predict contact events. For example, [8] shows that tactile sensing is adequate to distinguish slippage, rolling, making/breaking contact, among others, from a data-driven approach. Regarding their usage in manipulation tasks, [9] uses tactile sensing for packing four basic objects shapes in a box, with the hypothesis that different error directions will result in distinguishable tactile imprints. In [10] tactile sensing is used as the state for learning an insertion policy with reinforcement learning. This work highlights the need of feedback mechanisms to interactively correct the misalignment between object and environment and is a potential application for Neural Contact Fields.

Vision-based tactile sensing has been used in many other applications for robot manipulation. For example, estimating object poses from touch measurements [11], [12], [13] and learning a mapping of contact shapes to object poses [14], [15]. A global localization on an object surface is presented in [16] from a vision-based touch sensor sliding. Additionally, 3D shape reconstruction from touch has been presented in [17], [18] and [19].

C. Neural representations of 3D geometries

Our method for extrinsic contact tracking leverages neural fields to encode complex 3D geometries. Neural fields enable representing a shape's surface by a continuous volumetric field. For instance, DeepSDF [20] learns a Signed Distance Function (SDF) across a class of shapes, implicitly encoding a shape's boundary as the zero-level-set of the function. Occupancy Networks [21] reason about the occupancy probability of a 3D point, conditioned on the input observation of the shape (e.g., image, point cloud, etc.). Other such examples include volume density [22], [23] and

neural radiance fields [24], [25], [26]. The most significant advantage of neural fields is that they enable representing objects and scenes with infinite resolution in an efficient parameterization. They allow operating on a latent space that encodes information about the object class, as well as its most salient geometric features. We specifically take advantage of this feature to generalize across a diverse of categories of objects and shapes.

With such advantages, neural fields are increasingly being explored and built on for robotics applications like learning policies [27], robot self-models for space occupancy queries [28], room scale online signed distance fields for navigation [29], and learning transformations from demonstrations of a pick-and-place task [30].

III. Tracking Extrinsic Contact

In this paper we focus on the problem of tracking extrinsic contact between object and environment. When the object is making external contact, it moves in accordance with the kinematic constraints of the contact. This motion can be well-captured by a sequence of images from vision-based tactile sensors. Assuming that the object is rigidly grasped, 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. To achieve these, we estimate the probability of contact for any 3D point on an object surface with the model illustrated in Fig. 1. This model consists of three modules:

- A *PointNet Encoder + Neural Point Descriptor Fields* that allow us to generalize across different object shapes. This generates a spatial descriptor in the latent space of shapes for every 3D point $\mathbf{x}$ on an object.
- A *Sequence-to-Sequence Autoencoder* to extract the embedding from a sequence of tactile images. This embedding encodes the motion of the object due to the contact interaction with the environment.
- *Neural Contact Fields* (ours), which can estimate the probability of contact for a spatial descriptor, given the embedding of the motion due to the external contact and the most recent history of end-effector poses and external contact probabilities.

A. Neural Point Descriptor Fields

In this subsection we briefly provide background on Neural Point Descriptor Fields by Simeonov et al. [30] and how we are using them.

Neural Point Descriptor Fields represent an object as a function that maps a 3D coordinate $\mathbf{x}$ to a spatial descriptor, based on the architecture of Occupancy Networks [21]. The motivation behind Neural Point Descriptor Fields is that occupancy networks can be viewed as a classifier Φ, where for 3D shape reconstruction the decision boundary implicitly represents the object's surface. By conditioning the model on different low-dimensional latent codes, occupancy networks can reconstruct different shapes. These latent codes can be obtained as the output of a PointNet encoder $\mathcal{E}$ [31] that takes as input the point cloud $\mathbf{P}$ of the shape. The final layer performs a coarse classification for $\mathbf{x}$, inside or outside the shape, whereas the inner layers encode an increasing degree of detail about the surface shape.

Based on these insights, 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, given that the model is conditioned on the latent code $\mathcal{E}(\mathbf{P})$, the model is forced to parameterize the spatial descriptors grounded in the latent code of the object's category:

$$\mathbf{z} = g(\mathbf{x}|\mathbf{P}) = \bigoplus_{i=1}^{L} \Phi^i(\mathbf{x}, \mathcal{E}(\mathbf{P})), \tag{1}$$

where Φ^i is the output after activation of the i-th layer with L total number of layers, and $\bigoplus$ as the concatenation operator.

For tracking extrinsic contact, for each $\mathbf{x}$ in the set of N query surface points we have spatial descriptors that encode their relationship with salient features of the object. However, given that the object is in motion (for example, a robot arm is placing the object on a shelf) it is important that these descriptors remain unchanged regardless of the pose of the object.

In our work, we assume that the object is rigidly grasped, so that the configuration 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 [32]. Translation equivariance is easily implemented by mean-centering the object's point cloud $\mathbf{P}$. In this way, our Neural Contact Fields works in the latent space of spatial descriptors $\mathbf{z} \in \mathbb{R}^n$:

$$\mathbf{z}_i = g(\mathbf{x}_i|\mathbf{P}) = g(\mathbf{R}\mathbf{x}_i + \mathbf{t}|\mathbf{R}\mathbf{P} + \mathbf{t}) \forall\, i \in \{0, N\} \tag{2}$$

B. 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 and no changes are observed in the tactile measurements. 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 with a sequence-to-sequence autoencoder. First, we learn an autoencoder that allow us to capture a low-dimensional representation of raw 320×240 RGB tactile images per finger. Then, our sequence-to-sequence autoencoder uses as tokens these low-dimensional representation from left and right fingers concatenated with the end-effector pose. Both encoder and decoder have a convolutional LSTM as recurrent network [33]. This setting allow us to learn $\mathcal{E}(\mathbf{G}) \in \mathbb{R}^m$, an implicit representation of tactile sequences in a self-supervised manner.

This implicit representation corresponds to the last hidden state of the encoder.

C. Neural Contact Fields

Our overall pipeline, shown in Fig. 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 $\mathbf{P}$:

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

Neural Contact Fields (NCF) makes 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 tactile images at timesteps $t, t-2, t-4, t-7, t-9$:

$$f(\mathbf{z}, \mathcal{E}(\mathbf{G})) : \mathbb{R}^n \times \mathbb{R}^m \rightarrow [0, 1] \quad (4)$$

In addition, 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.

Given the similarities between Occupancy Networks and NCF in the sense that for both models the output is a probability, we follow a similar architecture. We feed our set of inputs through four fully-connected ResNet blocks with Conditional Batch-Normalization [34] 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 $\mathbf{x}$.

IV. The YCB-Extrinsic-Contact Dataset

To evaluate our NCF and to encourage further research in extrinsic contact modeling through tactile sensing, we open source the YCB-Extrinsic-Contact dataset. It consist of six kitchen cabinet scenarios with YCB objects in it on different configurations, as presented in Fig. 2, four for collecting training data and two for testing the model. A Franka Panda arm with a parallel gripper and DIGIT sensors [5] mounted on both fingers is interacting with these simulated environments in PyBullet [35]. The trajectories that the robot follows are randomly generated. We use TACTO [6] for simulating the sensors. For collecting extrinsic contact data, we use three categories of objects: mugs, bottles and bowls. We use several shapes for collecting training trajectories, obtained from the 3D Warehouse dataset [36]. In order to evaluate the generalization of our model on unseen shapes for the grasped object, we collect the testing trajectories using new shapes for the three objects in testing scenarios. The initial grasp pose for each object is fixed and rigid and a example is shown in Fig. 3.

The dataset provides for every contact event, access to a history of DIGIT images for each finger, end-effector pose, a reference point cloud of the grasped object, the set of query surface points and their ground truth probabilities of external contact.

Fig. 2. For collecting our YCB-Extrinsic-Contact dataset we simulate six kitchen cabinet environments with different configurations of YCB objects in PyBullet. Four scenes are use for collecting training data and the other two for testing. Objects in these environments can move around as the robot interacts with the environment when collecting extrinsic contact data.

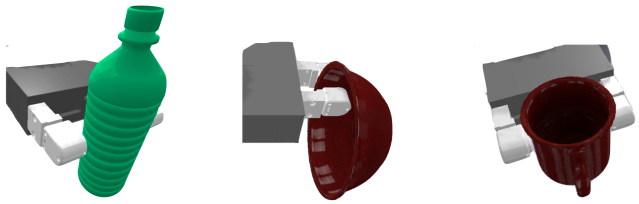

Fig. 3. For collecting our dataset, the grasp pose for all bottles (left), bowls (middle), and mugs (right) is fixed and rigid as shown.

V. 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 [37] with learning rate $1e^{-4}$.

We evaluate NCF offline with a GPU Nvidia RTX 3080, for which the forward pass takes ∼66 ms. In Fig. 4 we show qualitative results of our pipeline tracking extrinsic contact. We show key snapshots of a test trajectory collected with unseen mugs, bottles, bowls, and new scenarios 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. A video of these trajectories along with the

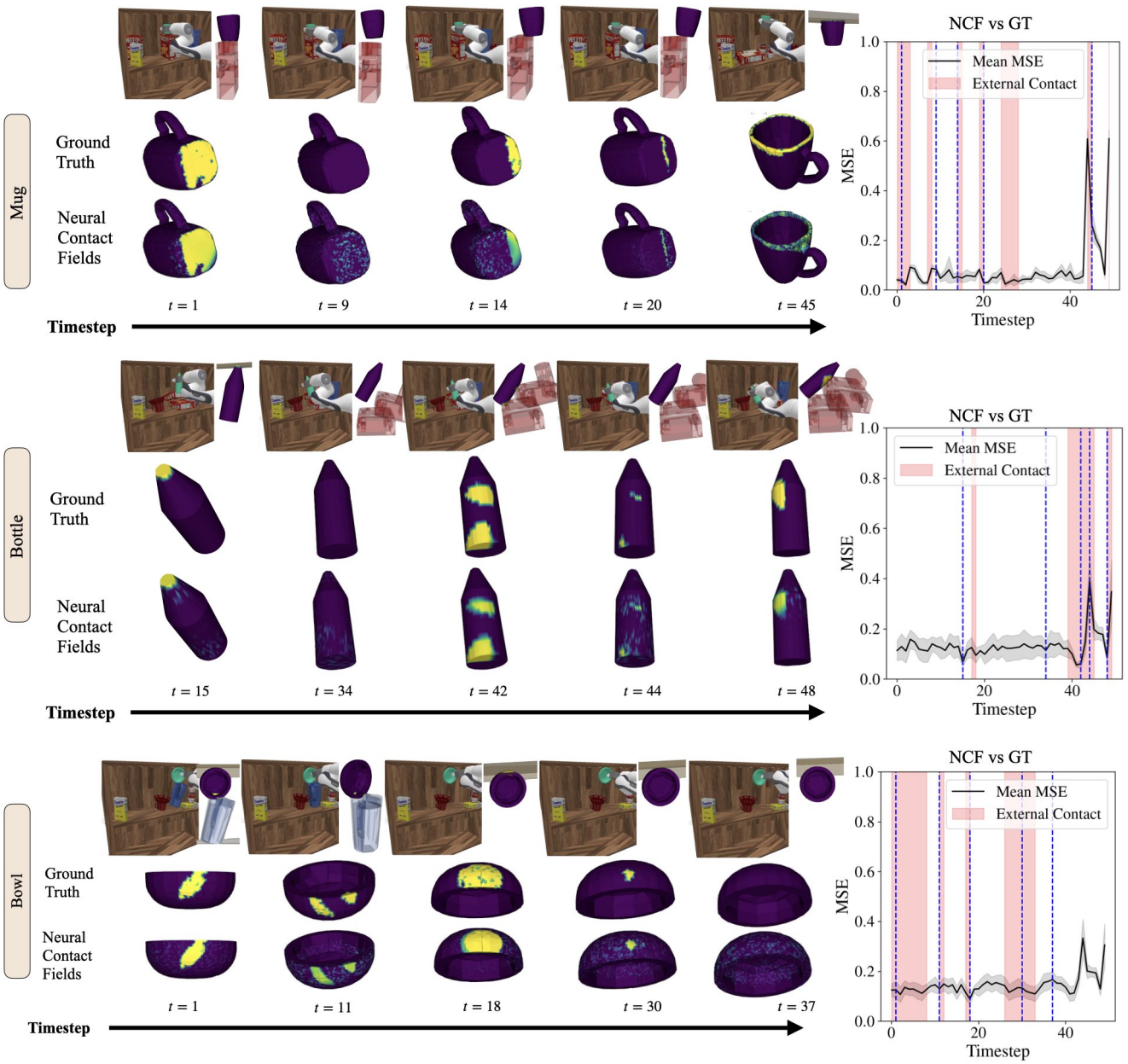

Fig. 4. Snapshots of extrinsic contact predictions on simulated testing trajectories at key timesteps (blue dotted lines) for mugs, bottles and bowls. For each row: **[top]** Pybullet frame and zoom-in view of the contact interaction, **[middle]** extrinsic contact ground truth probabilities, **[bottom]** Neural Contact Fields prediction, and **[right]** MSE between the ground truth and estimated extrinsic contact probabilities for the trajectory over time.

extrinsic contact tracking made by our NCF is available in the supplementary material.

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 10 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.

In addition, we perform ablations over the inputs of our NCF to quantify their contribution on the performance of the model. We compare four models: 1) NCF; 2) NCF without a history of contact probabilities and end-effector poses; 3) NCF considering only the current tactile frame rather than a history of images; and 4) NCF without a history of tactile images, contact probabilities and end-effector poses. In Fig. 5 we plot the MSE over time for each model and we highlight the timesteps with external contact events. The contribution of a history of tactile images for tracking extrinsic contact is noticeable. This is expected, given the intuition that the sequences of tactile images contain indirect information about the external contact. This results also suggest that considering only a history of contact probabilities instead of the sequence of tactile images generates a loss of information about how the external contact has been evolving over time.

We illustrate in Fig. 6 a failure case, specifically what

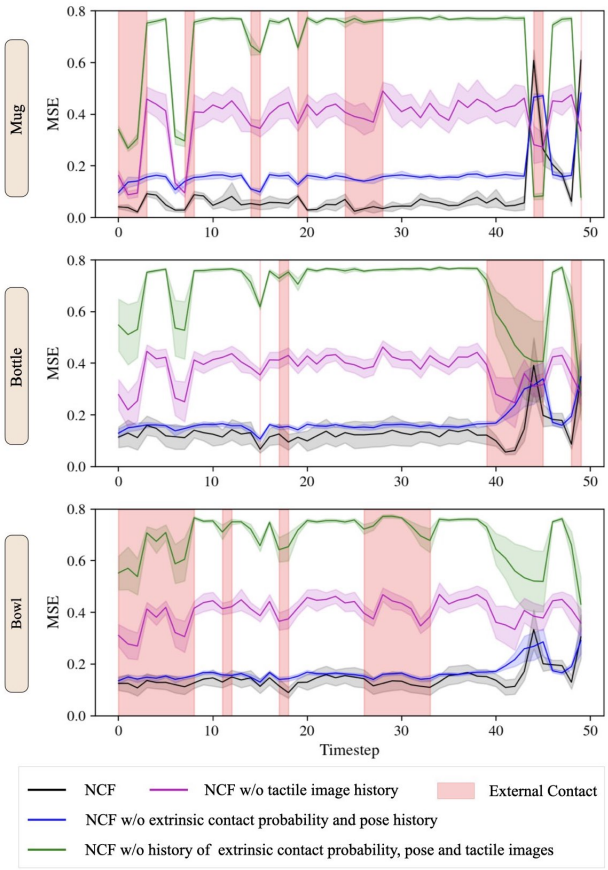

Fig. 5. Ablations over inputs for tracking extrinsic contact with NCF. We compare MSE with respect to ground-truth external contact probabilities over representative trajectories. We evaluate four models: NCF, without extrinsic contact probability and end-effector pose history, without tactile images history (considering only the current frame), and without history of tactile images, extrinsic contact probability and end-effector pose. From previous timesteps considering extrinsic contact and sequences of tactile measurements significantly improves performance of the model.

is happening around the timestep where NCF presents the highest error for the mug's test trajectory. We plot the extrinsic contact location prediction made by each of the four variants in the ablation and the ground truth for reference. In general, when transitioning from a no-contact state (in the figure, from timestep 43 to 44), all models have a high error while the contact information starts propagating. Note for example that the prediction made by the NCF model at timestep 44 does not capture the contact interaction yet, but starts being noticeable at the next timestep. Additionally, by looking at the predictions of all models at timestep 43, we can notice that the models without tactile image history suffer from higher uncertainty when the object is not making contact. This suggest the advantage of using a sequence of tactile images instead of only the current frame.

VI. Discussion

Summary. We presented 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 allow us to generalize within

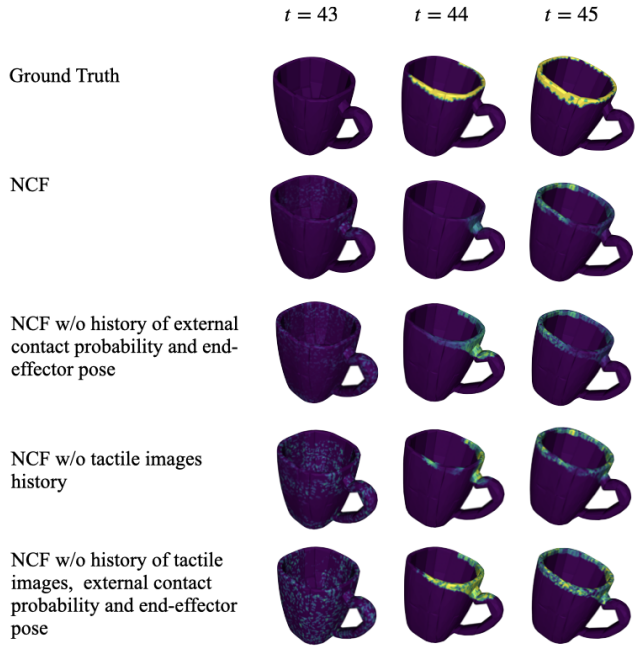

Fig. 6. Illustration of failure case tracking extrinsic contact in the mug's test trajectory. We show the context around time step 44, where NCF presents the highest error. In general, when transitioning from a no-contact state all models have a high error while the contact information starts propagating.

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 of contact. We believe NCF can be used in downstream robot manipulation tasks and leave its applications for future work.

Limitations. NCF has been tested in simulation only and we anticipate sim2real will be nontrivial since simulation isn't perfect and acquiring ground-truth extrinsic contact labels on real data at scale will be intractable. To close the gap in simulation we can leverage for example, domain randomization over background and lighting of the DIGIT sensor. Another alternative can be fine-tuning the sim-trained NCF on the real system with a downstream task policy in the loop. The current implementation of NCF is limited to tracking extrinsic contacts for three classes of objects that are grasped with a fixed relative pose and a rigid grasp (no slipping). Given initial evidence it should be possible to scale training to larger datasets with diverse objects, which will additionally aid in addressing challenges to transferring to the real world. However, for relaxing the rigid grasp assumption, we will need to estimate the relative hand-object pose. This will allow us to study the ambiguity on where the contact might happen based on the grasp pose. Our model also requires known class of object in order to input the correct reference point cloud. Although this can be addressed using pre-trained object detection models, the correctness of NCF will be compromised if the object is miss-classified. Finally, we are currently using the ground truths for establishing the prior information about the contact at the first timestep. This can be relaxed if during training we assume no-contact as prior for the first timestep of the trajectories.

ACKNOWLEDGMENT

The authors thank Sudharshan Suresh for feedback on the paper draft.

REFERENCES

[1] W. Zhou and D. Held, "Learning to grasp the ungraspable with emergent extrinsic dexterity," in *ICRA 2022 Workshop: Reinforcement Learning for Contact-Rich Manipulation*, 2022.

[2] N. C. Dafle, A. Rodriguez, R. Paolini, B. Tang, S. S. Srinivasa, M. Erdmann, M. T. Mason, I. Lundberg, H. Staab, and T. Fuhlbrigge, "Extrinsic dexterity: In-hand manipulation with external forces," in *2014 IEEE International Conference on Robotics and Automation (ICRA)*, 2014, pp. 1578–1585.

[3] 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)*. IEEE, 2021, pp. 11 262–11 268.

[4] 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)*, 2022, pp. 10 241–10 247.

[5] 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*, vol. 5, no. 3, pp. 3838–3845, 2020.

[6] S. Wang, M. Lambeta, P.-W. Chou, and R. Calandra, "Tacto: A fast, flexible, and open-source simulator for high-resolution vision-based tactile sensors," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3930–3937, 2022.

[7] M. Danielczuk, A. Mousavian, C. Eppner, and D. Fox, "Object rearrangement using learned implicit collision functions," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 6010–6017.

[8] Y. Zhang, W. Yuan, Z. Kan, and M. Y. Wang, "Towards learning to detect and predict contact events on vision-based tactile sensors," in *Conference on Robot Learning*. PMLR, 2020, pp. 1395–1404.

[9] S. Dong and A. Rodriguez, "Tactile-based insertion for dense box-packing," *IEEE International Conference on Intelligent Robots and Systems*, pp. 7953–7960, 9 2019.

[10] S. Dong, D. K. Jha, D. Romeres, S. Kim, D. Nikovski, and A. Rodriguez, "Tactile-rl for insertion: Generalization to objects of unknown geometry," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2021-May, pp. 6437–6443, 2021.

[11] P. Sodhi, M. Kaess, M. Mukadam, and S. Anderson, "Learning tactile models for factor graph-based estimation," in *IEEE Intl. Conf. on Robotics and Automation (ICRA)*, 2021.

[12] ——, "Patchgraph: In-hand tactile tracking with learned surface normals," in *2022 International Conference on Robotics and Automation (ICRA)*. IEEE, 2022, pp. 2164–2170.

[13] S. Suresh, M. Bauza, K.-T. Yu, J. G. Mangelson, A. Rodriguez, and M. Kaess, "Tactile slam: Real-time inference of shape and pose from planar pushing," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 11 322–11 328.

[14] M. Bauza, A. Bronars, and A. Rodriguez, "Tac2pose: Tactile object pose estimation from the first touch," 2022.

[15] M. B. Villalonga, A. Rodriguez, B. Lim, E. Valls, and T. Sechopoulos, "Tactile object pose estimation from the first touch with geometric contact rendering," in *Conference on Robot Learning*. PMLR, 2021, pp. 1015–1029.

[16] S. Suresh, Z. Si, S. Anderson, M. Kaess, and M. Mukadam, "MidasTouch: Monte-Carlo inference over distributions across sliding touch," in *Conference on Robot Learning (CoRL)*, 2022.

[17] S. Suresh, Z. Si, J. G. Mangelson, W. Yuan, and M. Kaess, "Shapemap 3-d: Efficient shape mapping through dense touch and vision," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 7073–7080.

[18] 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)*, 2018, pp. 1606–1613.

[19] E. Smith, R. Calandra, A. Romero, G. Gkioxari, D. Meger, J. Malik, and M. Drozdzal, "3d shape reconstruction from vision and touch," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 14 193–14 206.

[20] J. J. Park, P. Florence, J. Straub, R. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 165–174.

[21] 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.

[22] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," in *Proceedings of Machine Learning and Systems 2020*, 2020, pp. 3569–3579.

[23] R. Chabra, J. E. Lenssen, E. Ilg, T. Schmidt, J. Straub, S. Lovegrove, and R. Newcombe, "Deep local shapes: Learning local sdf priors for detailed 3d reconstruction," in *European Conference on Computer Vision*. Springer, 2020, pp. 608–625.

[24] 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," in *ECCV*, 2020.

[25] A. Pumarola, E. Corona, G. Pons-Moll, and F. Moreno-Noguer, "D-nerf: Neural radiance fields for dynamic scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 318–10 327.

[26] K. Park, U. Sinha, J. T. Barron, S. Bouaziz, D. B. Goldman, S. M. Seitz, and R. Martin-Brualla, "Nerfies: Deformable neural radiance fields," *ICCV*, 2021.

[27] Y. Li, S. Li, V. Sitzmann, P. Agrawal, and A. Torralba, "3d neural scene representations for visuomotor control," in *Conference on Robot Learning*. PMLR, 2022, pp. 112–123.

[28] B. Chen, R. Kwiatkowski, C. Vondrick, and H. Lipson, "Fully body visual self-modeling of robot morphologies," *Science Robotics*, vol. 7, no. 68, p. eabn1944, 2022.

[29] J. Ortiz, A. Clegg, J. Dong, E. Sucar, D. Novotny, M. Zollhoefer, and M. Mukadam, "iSDF: Real-time neural signed distance fields for robot perception," *Robotics: Science and Systems (RSS)*, 2022.

[30] 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)*. IEEE, 2022, pp. 6394–6400.

[31] 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*, 2017, pp. 652–660.

[32] 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*, 2021, pp. 12 200–12 209.

[33] 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*, vol. 28, 2015.

[34] 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*, vol. 30, 2017.

[35] E. Coumans and Y. Bai, "Pybullet, a python module for physics simulation for games, robotics and machine learning," 2016–2021.

[36] 3d warehouse. [Online]. Available: https://3dwarehouse.sketchup.com/?hl=en

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