VIRDO: Visio-tactile Implicit Representations of Deformable Objects

Youngsun Wi[1], Pete Florence[2], Andy Zeng[2] and Nima Fazeli[1]

Abstract— **Deformable object manipulation requires computationally efficient representations that are compatible with robotic sensing modalities. In this paper, we present VIRDO: an implicit, multi-modal, and continuous representation for deformable-elastic objects. VIRDO operates directly on visual (point cloud) and tactile (reaction forces) modalities and learns rich latent embeddings of contact locations and forces to predict object deformations subject to external contacts. Here, we demonstrate VIRDOs ability to: i) produce high-fidelity cross-modal reconstructions with dense unsupervised correspondences, ii) generalize to unseen contact formations, and iii) state-estimation with partial visio-tactile feedback.**

I. Introduction

In this paper, we present VIRDO – an implicit, dense, cross modal, and continuous architecture that addresses these fundamental representation and perception challenges for the class of elastically deformable objects. The central feature of our method is learning deformation fields informed by cross modal visual and tactile cues of external contacts. We further contribute a dataset of elastically deformable objects with boundary conditions used to evaluate. This paper focuses on dense geometric representations (signed-distance function) because they can facilitate downstream tasks such as state-estimation from partial views and estimating dense correspondences, as we demonstrate, as well as bootstrapping keypoint/affordance learning.

Problem Statement & Assumptions : Our goal is to derive a computationally efficient and generative model that: 1) predicts object deformations subject to external forces; and 2) is compatible with common robotic sensors. We assume the object geometry is described by its point cloud: an unordered set $\boldsymbol{P} := \{\boldsymbol{p} \in \mathbb{R}^3 : \text{SDF}(\boldsymbol{p}) = 0\}$ where SDF denotes the signed distance w.r.t. the surface of the object. Point clouds are obtained from commodity depth sensors or 3D scanners commonly found in industry. Contact locations are also given as a set of points $\boldsymbol{Q}$ which can be given by an upstream perception algorithm such as [1]–[3]. For the tactile input, net reaction force is given by $\boldsymbol{u} \in \mathbb{R}^3$ at the wrist of the robot which can be measured by common industrial F/T sensors or recovered from joint torques. In this paper, we derive a continuous and implicit representation of the deformed object geometry ($f(\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{u}) = s$) subject to external forces and their locations. Here the object geometry is given by the zero-level set of the implicit function; i.e. $s = 0$.

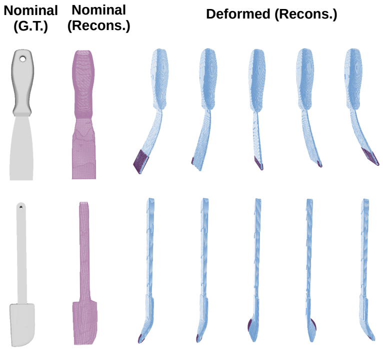

Fig. 1: **Reconstruction Results:** Example reconstructions of multiple nominal shapes and their deformations, learned simultaneously by VIRDO. Marching Cube algorithm is used for the reconstruction, where we highlighted the contact location as purple region.

II. Representing Deformable Objects using SDFs

At a high-level, VIRDO decomposes object representations into a nominal shape representation and a point-wise deformation field. The point-wise deformation field is produced using a summary of all boundary conditions (contact locations, reaction force, and fixed constraints) leveraging a permutation invariant set operator. The structure is fully differentiable and can be learned end-to-end.

A. Nominal Shape Representation

The nominal shape of an object is the geometry it takes in the absence of external contact forces. The nominal geometry is produced by the object module as $\boldsymbol{O}(\boldsymbol{x}|\boldsymbol{\Psi}_o(\boldsymbol{\alpha})) = s$ as a signed distance field, where $\boldsymbol{x} = (x, y, z)$ is a query point, and s is the signed-distance. The purpose of the object code (α) is to allow VIRDO to represent multiple objects. Here, we use a hyper-network $\boldsymbol{\Psi}_o(\boldsymbol{\alpha})$ to decode the object code into object module's weights and biases $\boldsymbol{\theta}_o$, similar to [4].

B. Deformed Object Representation

VIRDO uses two main components to model deformations: First, Force Module summarizes the contact formation (external contact locations and the reaction force). Second, Deformation Module uses this summary to predict a deformation field. Intuitively, the deformation field maps the deformed object back into its nominal shape.

The Force Module $\boldsymbol{F}$ is an encoder that summarizes the contact locations and reaction force $(\boldsymbol{Q}, \boldsymbol{u})$ into a force code $\boldsymbol{z} = \boldsymbol{F}(\boldsymbol{Q}, \boldsymbol{u})$. We assume that the contact location set $\boldsymbol{Q}$ is given as a subset of the nominal point cloud ($\boldsymbol{Q} \subset \boldsymbol{P}$) and the reaction force $\boldsymbol{u} \in \mathbb{R}^3$ is directly measured at the robot's wrist. Point clouds, including the contact location set $\boldsymbol{Q}$, are unordered and variable in length. Our contact location encoder utilizes the PointNet architecture [5].

[1] Youngsun Wi and Nima Fazeli are with the Robotics Institute at the University of Michigan, MI, USA <yswi,nfz>@umich.edu

[2] Pete Florence and Andy Zeng with Robotics at Google, Mountain View, CA, USA <peteflorence,andyzeng>@google.com

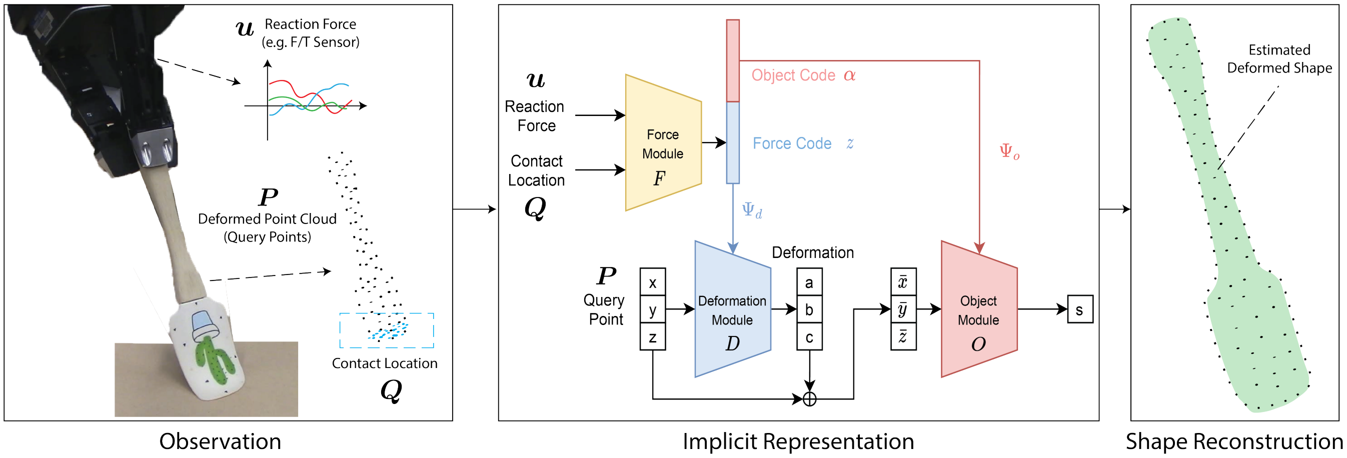

Fig. 2: **Representation Architecture:** The left panel depicts how visual data in the form of point clouds and tactile in the form of reaction forces may be collected in practical robotic settings. The middle panel depicts the network and how this information is processed to predict the implicit surface representation encoded as a signed-distance function. Finally, the right panel depicts the reconstruction of the estimated true surface given the external contacts and reaction force.

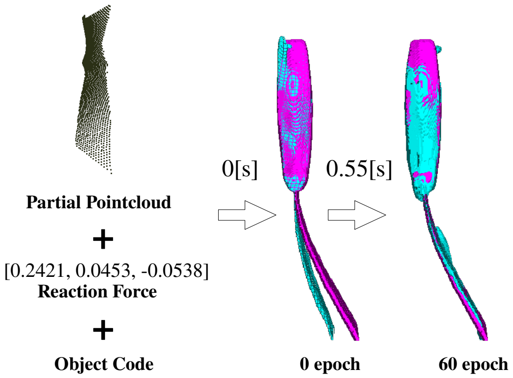

Fig. 3: **Inference:** Reconstructions with inferred deformation field (cyan), ground truth deformed object (magenta)

We define the deformation field as a 3D vector field that pushes a deformed object back to its original (nominal) shape. The deformation field is produced by the deformation module $\boldsymbol{D}$ with parameters $\boldsymbol{\theta}_d$ predicted by the hypernetwork $\boldsymbol{\Psi}_d$. We highlight that $\boldsymbol{\theta}_d$ is conditioned on the latent code pair $(z, \boldsymbol{\alpha}) \in \mathbb{R}^{l+m}$ to capture the underlying object-specific deformation behavior. This results in $\boldsymbol{D}$ predicting different deformation fields for different objects despite similar contact locations and reaction force measurements. This is desirable because objects may be geometrically similar but deform differently due to varying material properties. Details of loss formulations for training are well described in [6].

III. Experiments and Results

A. Data Preparation

In total, we generated 6 objects each with 24 unique boundary conditions using MATLAB PDE toolbox. The 3D meshes were collected from open-sourced 3D model repository. For the training, we normalized the point cloud with the geometric center at $[0,0,0]$.

B. Representing Known Deformable Shapes

The average reconstruction accuracy is 0.3474x10^3 in Chamfer Distance(CD). CD is measured between reconstructions and query points unseen during training in average, where we utilized Marching Cube algorithm [7] for the reconstruction. We emphasize that only one neural network model was used for the entire data-set.

C. Deformation Field Inference

We test the model's ability to infer a deformation field given reaction force, partial pointcloud, and object code, seen from the training. Here, we infer the contact feature in Fig. **??** to estimate deformation. First, we randomly initialize the contact feature from $N(0, 0.01^2)$. Then, we update the feature with an L1 loss which only consumes a partial zero-level set: $L_{infer} = \sum_{\boldsymbol{x}\in\boldsymbol{\Omega_o}} |clamp(\boldsymbol{O}(\boldsymbol{x}), \delta)|$. The loss encourages VIRDO to update the contact feature by minimizing the mismatch between the initial guess and the partial observation. Fig. 3 is a partial pointcloud where the handle and the tip are occluded, rendered in simulation with a single pinhole camera. At epoch 0, the model already makes deformation field fairly close to the ground truth. This shows VIRDO's ability to perform state estimation when the vision is missing. As the gradient descent progresses on the contact feature, the estimated deformation converges towards the ground truth. We note that only the on-surface points are used for this experiment, since an RGBD camera would feasibly only give on-surface points in real-world experiments; however, it is also possible boost the inference performance by collecting off-surface samples along the camera ray and augment the partial observation.

IV. Discussion & Limitations

The fundamental principle driving VIRDO is the ability to learn deformation fields informed by visio-tactile sensing. VIRDO is the first learned implicit method to integrate tactile and visual feedback while modeling object deformations subject to external contacts. The representation has arbitrary resolution and is cheap to evaluate for point-wise sampling. Additionally, the latent code is well-behaved and can be used for inference.

References

[1] T. Hermans, F. Li, J. M. Rehg, and A. F. Bobick, "Learning contact locations for pushing and orienting unknown objects," in *2013 13th IEEE-RAS international conference on humanoid robots (humanoids)*, IEEE, 2013, pp. 435–442.

[2] M. Sharma and O. Kroemer, "Relational learning for skill preconditions," *arXiv preprint arXiv:2012.01693*, 2020.

[3] Q. V. Le, D. Kamm, A. F. Kara, and A. Y. Ng, "Learning to grasp objects with multiple contact points," in *2010 IEEE International Conference on Robotics and Automation*, IEEE, 2010, pp. 5062–5069.

[4] Y. Deng, J. Yang, and X. Tong, "Deformed implicit field: Modeling 3d shapes with learned dense correspondence," *arXiv preprint arXiv:2011.13650*, 2020.

[5] 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.

[6] Y. Wi, P. Florence, A. Zeng, and N. Fazeli, "VIRDO: visio-tactile implicit representations of deformable objects," *CoRR*, vol. abs/2202.00868, 2022. arXiv: `2202.00868`. [Online]. Available: `https://arxiv.org/abs/2202.00868`.

[7] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," *ACM siggraph computer graphics*, vol. 21, no. 4, pp. 163–169, 1987.