RoPotter: Toward Robotic Pottery and Deformable Object Manipulation with Structural Priors

Uksang Yoo [1✉*], Adam Hung[2*], Jonathan Francis[1,3], Jean Oh[1], and Jeffrey Ichnowski[1]

***Abstract*— Humans are capable of continuously manipulating a wide variety of deformable objects into complex shapes. This is made possible by our intuitive understanding of material properties and mechanics of the object, for reasoning about object states even when visual perception is occluded. These capabilities allow us to perform diverse tasks ranging from cooking with dough to expressing ourselves with pottery-making. However, developing robotic systems to robustly perform similar tasks remains challenging, as current methods struggle to effectively model volumetric deformable objects and reason about the complex behavior they typically exhibit. To study the robotic systems and algorithms capable of deforming volumetric objects, we introduce a novel robotics task of continuously deforming clay on a pottery wheel. We propose a pipeline for perception and pottery skill-learning, called RoPotter, wherein we demonstrate that structural priors specific to the task of pottery-making can be exploited to simplify the pottery skill-learning process. Namely, we can project the cross-section of the clay to a plane to represent the state of the clay, reducing dimensionality. We also demonstrate a mesh-based method of occluded clay state recovery, toward robotic agents capable of continuously deforming clay. Our experiments show that by using the reduced representation with structural priors based on the deformation behaviors of the clay, RoPotter can perform the long-horizon pottery task with 44.4% lower final shape error compared to the state-of-the-art baselines. Supplemental materials, experiment data, and visualizations are available at ropotter.github.io.**

I. Introduction

When we perform long-horizon deformable object manipulation tasks such as making pottery, we can continuously manipulate the deformable objects into complex shapes, despite occlusions in the perception of the object's state. This ability to reason about a deformable object's occluded geometry allows us to perform diverse tasks robustly—from cooking with dough to expressing ourselves with pottery. However, developing robotic systems that can robustly perform these tasks remains challenging, due to the complex deformation behavior of volumetric deformable objects [1]. A common approach to manipulating volumetric deformable objects such as clay has been to learn a dynamics model of the object. Despite often remarkable results [2–4], such approaches presently have two drawbacks. First, state-of-the-art methods for learning an explicit dynamics model of the deformable objects in interaction with the environment suffer from a high sample-complexity. For 2-dimensional deformable objects such as cloth, researchers successfully trained the model in simulation and directly transferred it to the real world [5]. However, methods have struggled to similarly model high dimensional interactions due to the increased deformation complexity and the more notable sim-to-real gap in modeling contact dynamics. Secondly, and as a result of the first drawback, recent works on volumetric deformable object manipulation often relied on well-parameterized and task-dependent action sequences that allow fully unoccluded observation of the object in between each action to prevent drifting [2–4].

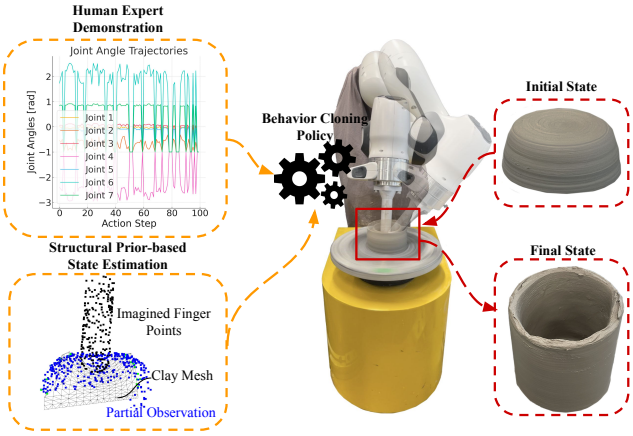

Fig. 1. RoPotter Pipeline. We train a behavior cloning policy with expert demonstration data from teleoperation and point cloud state estimates of the clay. We demonstrate that the structural priors of the clay deformations and geometry can assist with the recovery of occluded points during continuous deformable object manipulation.

Behavior cloning is receiving growing attention among robotic manipulation researchers as it requires fewer expert design choices, such as reward-shaping, as long as the precise actions during demonstrations are known [6–8]. Recently proposed approaches such as Implicit Behavior Cloning [9] and Diffusion Policy [10] have demonstrated an exceptional ability to learn complex manipulation skills given a reasonable number of human demonstrations, with some spatial generalizability and robustness to distractions [11]. However, despite their success in diverse manipulation tasks with cloth and clay-rolling—to the best of our knowledge—these works did not generalize these imitation learning pipelines to the manipulation of volumetric deformable objects such as a lump of clay. In this work, we use a pottery wheel with clay to study volumetric deformable object manipulation and demonstrate a pipeline for learning a bowl-making policy

*Equal contribution. ✉Corresponding author.
[1]Robotics Institute, Carnegie Mellon University, Pittsburgh, USA {uyoo,jmf1,hyaejino,jichnows}@andrew.cmu.edu
[2]University of Michigan, Ann Arbor, USA adamhung@umich.edu
[3]Bosch Center for Artificial Intelligence, Pittsburgh, USA

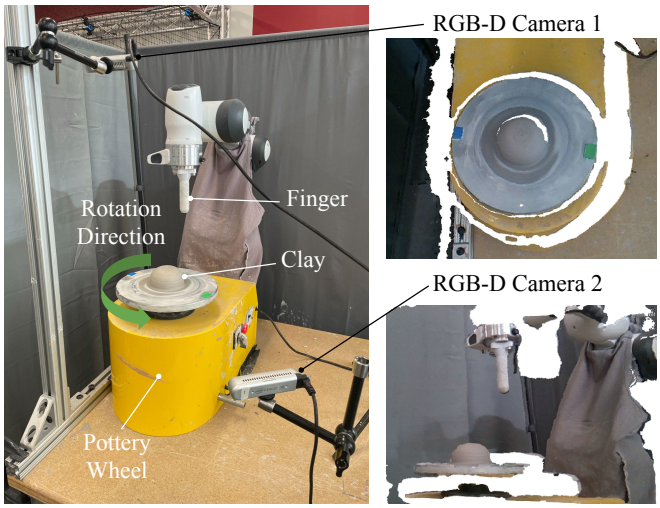

Fig. 2. Our setup for robotic pottery-making: two RGB-D cameras each provide a partial view of the clay, and multiple instances of the point clouds are combined to increase how much of the total surface of the clay can be observed.

with behavior cloning. Our task setup allows researchers to study the space of continuous volumetric deformable object manipulation with a simplified rigid 0 degree-of-freedom end-effector, as shown in Fig. 2. Because clay continuously deforms under contact, staged discrete-action approaches previously taken with volumetric deformable object manipulation tasks [4] are difficult to apply.

We present RoPotter, a pipeline for perception and behavior cloning based on diffusion policy [10, 11] to deform a block of clay into two bowl shapes. We incorporate structural priors specific to the pottery wheel task to simplify the skill-learning process. Namely, we show that because of the pottery task's radial symmetry, we can reduce the dimension of the explicit clay state representation to the 2D plane by taking a cross-section. Additionally, we demonstrate that we can exploit the structure of a mesh initialized over the starting shape of the clay to reason about the occluded points throughout the task and learn skills effectively. To summarize, we made the following contributions in this work:

1) We design the novel, long-horizon task of robotic pottery-making wherein an agent deforms a block of clay into two bowls of different dimensions.
2) We develop a system, RoPotter, to collect and register point clouds of the rotating clay deformable object and a pipeline for collecting demonstrations from users.
3) We leverage a structural prior-based approach to shape recovery and estimation of clay undergoing continuous deformation, which we evaluate with ablations.
4) We provide a behavior cloning pipeline for policy learning, which uses the proposed representations of the clay shape, and we evaluate performances with both geometric and semantic metrics.

II. Related Work

A. Deformable Object Perception and Representation

Conventional methods of representing deformable objects use analytical models such as mass-spring mechanics [1, 12, 13]. However, these methods rely on careful system identification, require knowledge about the object's material properties, and are sensitive to empirically-derived parameters [1, 14]. Recent advancements in learning representations with 3D geometry such as PointNet [15], PointNet++ [16], and PointBERT [17] have significantly improved robot capabilities in various applications—from autonomous driving to robot manipulation [4]. Additionally, methods in the broader family of Graph Neural Networks (GNN) have enabled researchers to introduce structure to the representation and dynamics-learning problem [2].

Subsequent works demonstrated that the use of priors can introduce structure and thereby regularize the representation learning process, yielding improvements in sample-efficiency of model training [18]. In the domain of deformable robot shape representation, researchers have noted that mechanics-based priors can help ground the soft body states on physically admissible configurations [18]. However, learning such explicit dynamics models requires either extensive exploration steps to perturb the object and observe state changes [2, 3], or exploitation of the simulation environments' privileged information that is not available in the real world [19]. Instead of explicit dynamics models, we leverage structural prior that does not require laborious steps. We propose a pipeline for learning deformable object manipulation skills for robot pottery-making, where, inspired by previous works, we incorporate structural priors based on task-dependent mechanics.

B. Deformable Object Manipulation

Various works have studied robot manipulation of deformable objects [1, 2]. Prior works can be broadly categorized by the dimensionality of the deformable objects. Researchers have previously proposed various methods to address robotic manipulation of one-dimensional deformable objects such as ropes and cables [20, 21], two-dimensional deformable objects such as cloth and flattened dough [22], and three-dimensional deformable objects such as clay and plasticine [2]. The ability to reason about manipulating three-dimensional geometries has broad application including in enabling new avenues of human expression through art for instance via human-robot collaborative sculpting [23]. In this work, we are specifically interested in robot pottery-making with clay, which falls under the category of three-dimensional deformable object manipulation. To the best of our knowledge, this work presents the first methods for robot pottery-making.

III. Problem Statement

Let S denote a set of states representing the space of 3D shapes that can be made of clay on a pottery wheel; and A, a set of end-effector actions available to a robot. The task of robot pottery-making can be defined as modifying an

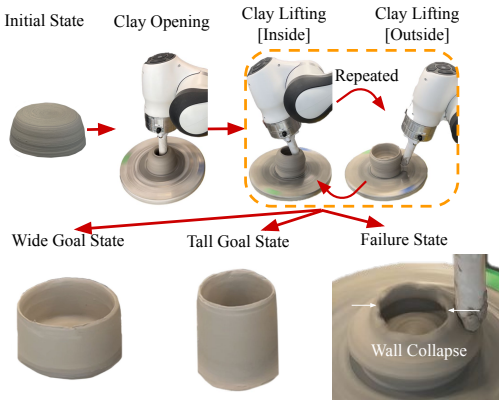

Fig. 3. The pottery-making task. We demonstrate the RoPotter's ability to produce two bowls of different dimensions.

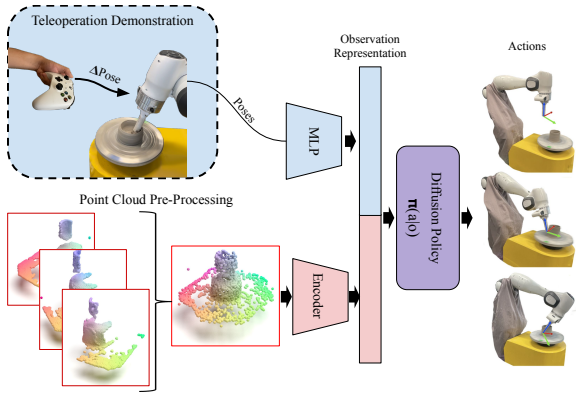

Fig. 4. Proposed pipeline for the robotic pottery. We collected expert demonstrations for the two bowl shapes using teleoperation with a gaming console controller. We paired the expert-demonstrated actions with the point cloud observation of the clay shape that was stiched together from multiple virtual perspectives.

initial state of clay $s_i \in S$ into a desired shape state $s_g \in S$, implicitly captured by the demonstrated set of final shape states. The objective here is to learn a policy, $\pi_{s_g}(a|s)$, which can prescribe optimal action $a \in A$ defined with respect to the end-effector poses given diverse state $s \in S$ towards achieving goal shape $s_g \in S$.

IV. Methods

We propose a learning from demonstration (LfD) approach to develop a pottery-making robotic system. Our approach is designed to address the sample complexity issue of LfD as well as the occlusions often present during continuous deformable object manipulation. Specifically, structural and mechanics-based priors can help simplify the robot learning problem by grounding predictions to the set of physically permissible spaces [2, 18]. In the RoPotter approach, we propose the use of two types of structural priors of the pottery task to reduce the complexity of the problem and account for occlusion.

A. RoPotter-2D: Compact Representation with Structural Priors

Because of the rotating base of the clay on a pottery wheel, we assume radial symmetry of the clay shapes during the task. We also assume quasi-static conditions for the clay and ignore dynamic effects during manipulation, modeling the clay shapes as discrete states at each time step.

In our first proposed pipeline, RoPotter-2D, we reduce the dimension of the observation space by taking a cross section composed of the points within a 5mm threshold of a plane containing the clay's center of mass. The reduction of dimension has the benefit of reducing the complexity of the dataset that the policy network must reason over and filtering out uninformative features. The 2D cross-section of the clay concisely captures a corresponding 3D shape state in terms of the diameter, height, and thickness of the walls.

Algorithm 1 RoPotter-Mesh Reconstruction

Input: $P \in \mathbb{R}^{N\times 3}$ ▷ Initial point cloud
Input: $x_t \in \mathbb{R}^7$ ▷ Robot joint angles at time step t
$P_{\text{xz}} \leftarrow \{(x,z) \mid (x,y,z) \in P, |y| < T_{\text{thresh}}\}$
$H \leftarrow \text{ConvexHull}(P_{\text{xz}})$
$H_{\text{aug}} \leftarrow \text{AugmentInteriorPoints}(H)$
$M_0 \leftarrow \text{DelaunayTriangulation}(H_{\text{aug}})$ ▷ Initial mesh
$M_t \leftarrow M_0$
for each time step t **do**
  $F_t \leftarrow \text{ComputeFingerPoints}(x_t)$
  $P_t \leftarrow P \setminus \text{ConvexHull}(F_t)$ ▷ Remove points in robot finger convex hull
  $P_{\text{xz},t} \leftarrow \{(x,z) \mid (x,y,z) \in P_t, |y| < T_{\text{thresh}}\}$
  $M_t \leftarrow \text{MoveVerticesToFit}(M_t, P_{\text{xz},t})$ ▷ Move vertices to fit observable points
  for each vertex $v \in M_t$ **do**
    if $v \in \text{ConvexHull}(F_t)$ **then**
      $n \leftarrow \text{SurfaceNormal}(F_t)$
      $v \leftarrow v + \delta \cdot n$ ▷ Push vertex in the direction of surface normal
    end if
  end for
  $M_t \leftarrow \text{ARAPUpdate}(M_t, M_0)$ ▷ Update the mesh using ARAP algorithm
end for
return M_t

TABLE I
RoPotter-Mesh Reconstruction Accuracy Ablation

Method	Input	CD [mm] ↓* CS	FP
Reference CS [Observed]	P	0.0	2.46
Mesh w/o Contact	P	15.41	17.01
Mesh w/o ARAP	x_t, P	4.45	7.74
Mesh [Proposed]	x_t, P	**1.50**	**2.58**

*CS refers to cross-section and FP refers to full point cloud.

B. RoPotter-Mesh: Shape Estimation with Structural Priors

In our RoPotter-Mesh reconstruction pipeline as outlined in Algorithm 1, we use structural priors on how the clay deforms and an assumption of local smoothness of the clay geometry to recover occluded points. Previous works have

shown that deformable soft bodies represented by discrete meshes often follow the constraints of local rigidity, also known as As-Rigid-As-Possible (ARAP) deformation [18, 24].

ARAP includes a penalty on the rotations of the neighboring edges, producing physically admissible mesh manipulation [18, 24, 25]. The ARAP energy that we minimize during ARAP deformation is given by:

$$E_{\text{smoothed}}(M, M') = \min_{R_1,\ldots,R_m} \sum_{k=1}^{m} \Big(\sum_{i,j \in e_k} c_{ijk} \|e_{ij} - R_k e_{ij}\|^2 + \lambda \hat{A} \sum_{e_l \in N(e_k)} w_{kl} \|R_k - R_l\|^2 \Big) \quad (1)$$

Here, M, M' define the mesh initialized with the initial point cloud's convex hull (line 4 Algorithm 1) and the deformed mesh respectively, c_{ijk} are the cotan weights [26], λ is the regularization weight, $R_1, ..., R_m \in SO(3)$ are the local rotations for each of the edges $e_k \in E$ where $m = |E|$, $\hat{A}$ is the triangle area and w_{kl} are the scalar weight terms defined by the cotan weights of the dual mesh of e_{kl} [26]. As with previous works on using ARAP for shape recovery and deformation, we use a local-global solver to iteratively minimize the energy defined in Eq. 1. During the local optimization step, we compute the locally best-fitting rigid transformation R_k for each of the edge simplices to best map to M_t from M_0. In the global step, we update the metric positions of the vertices $i, j \in e_k$ with least squares fitting to make them as consistent with rigid transformation R_k as possible. We repeat the local-global steps until convergence within the energy threshold or maximum iteration.

As outlined in Algorithm 1, we use ARAPUpdate to recover occluded points in the partial point cloud of the clay cross-section. With MoveVerticesToFit, we move the vertices of the mesh to their closest neighbors in the partial point cloud observation. We also displace the points in contact with the RoPotter finger, which are identified by checking for points in the convex hull of the finger with respect to the surface normal. Using the displaced points and the bottom row of points that are attached to the pottery wheel as constraint points, we minimize the energy in Eq. 1.

C. Learning From Demonstration with 3D representation

Learning from Demonstrations via Imitation Learning provides an intuitive approach to robot skill-acquisition, where expert demonstrations are directly used by the robot agent policy to quickly learn the involved underlying skills [27]. In Behavior Cloning, specifically, the policy learns to map observations to the demonstration actions. Such approaches benefit from intuitive training processes, few implementation-level challenges, and strong supervision, at the cost of requiring observations labeled with corresponding expert actions [6]. Our RoPotter approach learns robot pottery-making skills from human demonstration data, using action and observation pairs provided through teleoperation.

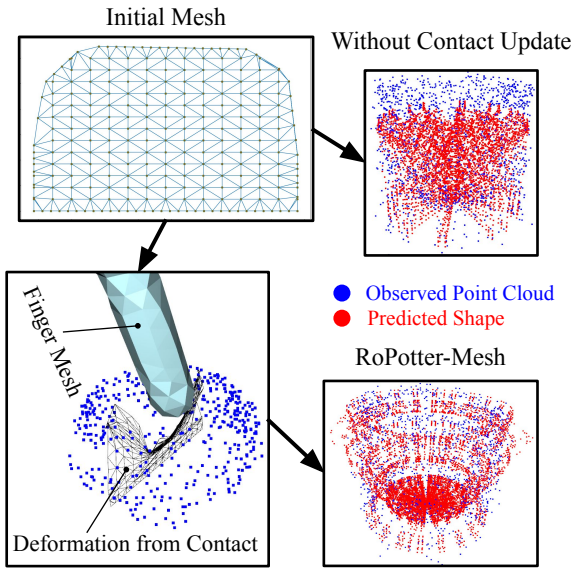

Fig. 5. RoPotter-Mesh method outline and ablation results.

Demonstration Collection: To collect demonstrations of the robot deforming the clay into a desired bowl shape, we controlled the change in pose of the finger with a gaming console controller (Xbox Wireless Controller, Microsoft) as shown in Fig. 4. The point clouds are pre-processed and merged into discrete observations, as described in Section V-C. In each demonstration, we deformed the clay procedurally until the overall measurements of the bowls matched the ones outlined in Section III. We trained a behavior cloning policy after collecting 40 demonstrations.

Policy Learning: We build on the previous work on diffusion policy with 3D point cloud observations [11] for the presented results. The end-effector orientation is represented with the continuous 6-dimensional representation as previously proposed [28]. The full pose of the end-effector with position and orientation is encoded with an MLP to a 64-dimensional feature. The merged and down-sampled point clouds are encoded with the DP3 encoder [11] to a 64-dimensional observation feature. For the baseline, we directly use the DP3 encoder architecture as described in Ze, et al. [11]. For our two proposed methods, we use the two-dimensional variant of the encoder. Next, the robot pose and point cloud features are concatenated. The diffusion policy then denoises random noise into an action sequence, conditioning on these state features. We use a single clay shape observation as input to the policy and predict a sequence of 8 denoised actions, which helps with the temporal consistency of the trajectories as noted in the literature [10, 11].

V. Implementation Details

A. Robotic Pottery Setup

As seen in Fig. 2, our robotic pottery setup has a pottery wheel with two colored markers (green and blue) on the edge of the pottery wheel surface that allow us to track the orientation of the clay at a given instance as the clay rotates at approximately three rotations a second.

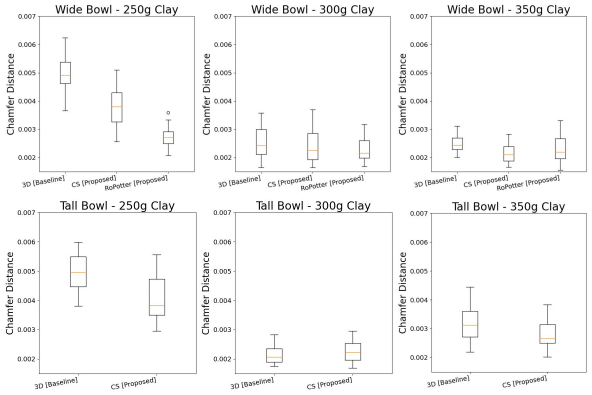

Fig. 6. Boxplots showing the full distributions of chamfer distances. For a bowl produced by a given policy and clay mass, the chamfer distance is calculated between this bowl's final state and every final bowl state in the corresponding set of demonstrated bowls, creating a single distribution.

B. Sensor Setup

Two RGB-D cameras (Realsense D415, Intel) are positioned, pointed at the clay, to observe the deforming clay shape at approximately 25 Hz and to capture partial point cloud views of top- and side-view perspectives. The two cameras were calibrated to the pottery wheel frame, where the origin is defined to lie on the wheel's rotation axis. During demonstrations, we capture the robot joint states as well as the point cloud of the bowl from virtual perspectives provided by the rotation of the clay.

C. Point Cloud Pre-Processing

In the initial frame, we detect both of the markers' centroid positions. We then define the orientation reference vectors from the centroid of each marker to the origin of the pottery wheel frame, and normalize to unit length as $\vec{r_0} = \frac{p_m - p_o}{\|p_m - p_o\|_2}$ where $p_m, p_o \in \mathbb{R}^2$ are respectively the marker and origin points projected onto the pottery wheel surface. In the subsequent frames, we track one of the two markers' positions to get the new orientation vector $\vec{r_t}$ (the position of the markers is susceptible to occlusion, but both markers are never occluded simultaneously). We then compute the rotation angle $\theta = arccos(\vec{r_t} \cdot \vec{r_0})$. The observed point clouds of the clay are merged as they are rotated from the start orientation until they make a full rotation. The combined point cloud is downsampled to 1,024 points and is used as a single full observation of the clay at a time instance.

VI. Evaluation

A. Experimental Setup

For the purpose of evaluation, we design two types of pottery-making as follows. We define two goal bowl geometries, where a 'wide bowl' goal is defined by outer diameter of 100.0 mm and height of 60.0 mm, and a 'tall bowl' goal has an outer diameter of 70.0 mm and height of 70.0 mm.

All models received 40 demonstrations of teleoperated continuous robot deformable manipulation of clay into each of the two bowl types; demonstrated bowl shapes will have variance caused by imperfect human demonstrations.

TABLE II
Policy Evaluation Against Demonstrated Pottery

Method	Task	Clay Mass [g]	CD [mm] ↓	CLIP Similarity ↑
DP3 [Baseline]	Wide	250	4.95	0.945
		300	2.55	0.962
		350	2.50	0.971
	Tall	250	4.93	**0.967**
		300	**2.14**	**0.966**
		350	3.19	0.952
RoPotter-2D [Proposed]	Wide	250	3.80	**0.954**
		300	2.44	**0.972**
		350	2.15	**0.973**
	Tall	250	**4.05**	0.957
		300	2.26	0.961
		350	**2.81**	**0.954**
RoPotter-Mesh [Proposed]	Wide	250	**2.75**	0.951
		300	**2.28**	0.971
		350	2.28	0.972

To test the generalizability of the policy, we also controlled for and precisely varied the amount of clay that was on the pottery wheel. We present the metrics that we propose for the evaluation of these policies and discuss the results.

B. Metrics

Previous works on learning from demonstration for manipulation relied on the tasks having clear success criteria, where the researchers can clearly classify the success of trials. However, our novel task of robotic pottery does not afford such clear evaluation criteria. As shown in Fig. 7, failure cases can sometimes be clear when the robotic agent fails to continue through the task or causes the catastrophic collapse of the clay structure. In other cases, however, success evaluation may require arbitrary distinction on which mistakes are disqualifying. To quantitatively evaluate the proposed methods, we propose two metrics: a geometric similarity and a semantic similarity to the demonstrated pottery.

Chamfer Distance (CD) metric (Geometric Similarity): We use the Chamfer Distance (CD) metric for assessing geometric similarity. The unidirectional CD metric from the source point cloud P_S to target point cloud P_T is defined by [19, 29]:

$$d_{UCD}(P_S, P_T) = \frac{1}{|P_S|} \sum_{x \in P_s} \min_{y \in P_T} \|x - y\|_2 \qquad (2)$$

When elements from both P_S and P_T can be accurately matched to each other, we can use the bidirectional variant of CD defined as the average of $d_{BCD}(P_S, P_T) = \frac{1}{2}[d_{UCD}(P_S, P_T) + d_{UCD}(P_T, P_S)]$. When we evaluate the RoPotter structural prior-based mesh state estimation, we use d_{UCD} metric from the observed real-world point cloud to the mesh vertex point cloud because the mesh includes internal points of the clay cross-section points that are not visible in the real world with external RGB-D cameras. When we evaluate the real-world point clouds of the bowls produced by the trained policy roll-out against the demonstrated bowl real-world point clouds, we use d_{BCD}.

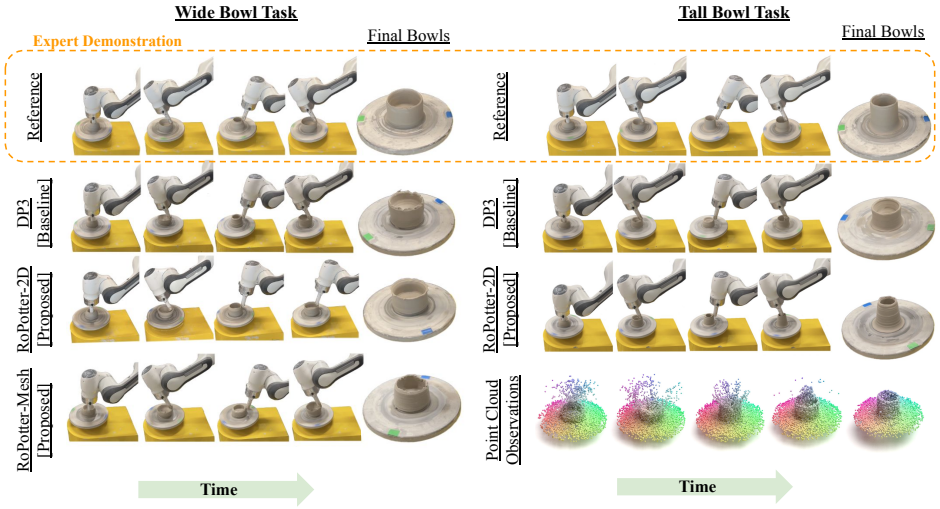

Fig. 7. Results of RoPotter robot pottery pipelines. Reference row shows the demonstrated bowls.

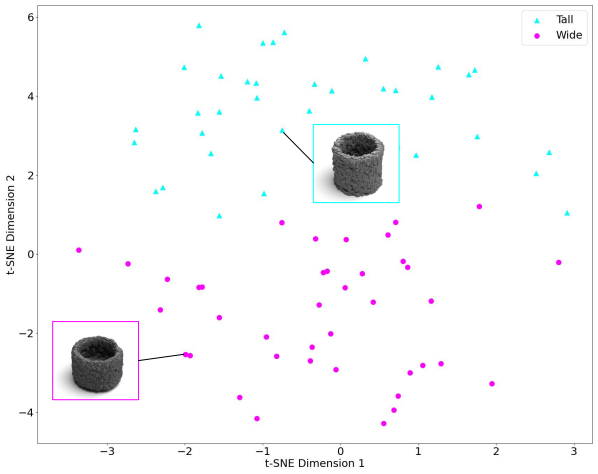

Fig. 8. Visualization of the bowl demonstrations' semantic features.

CD can broadly offer a measurement of geometric accuracy, capturing details such as metric height, diameter, and wall thickness. However, it is also prone to biases from outlier points by averaging the distances in Eq. 2. CD also tends to filter or average out high-frequency geometric features [30].

CLIP Score (Semantic Similarity): Additionally, in our problem statement in Section VI-A, goals such as "wide," "tall," and even "bowls" are largely only semantically meaningful descriptors that were implicitly captured with geometric goals of desired diameter and height. Toward evaluating our methods in the semantic space, we propose using CLIP [31] to compute feature similarity between the images of the bowls produced by the policy and the demonstrated bowls. To account for the differing lighting conditions, we re-render the voxelized point clouds using a physics-based renderer [32] and we use cosine similarity to compare the features. Fig. 8 shows a visualization of the feature space spanned by the demonstrated bowls. For both the geometric and semantic metrics, we report the results of the methods in Table II.

C. Experiments

We first evaluated the RoPotter-Mesh algorithm as outlined in Algorithm 1 to recover occluded points. For evaluation, we rolled out the demonstrated actions and corresponding partial point clouds for each wide and tall pottery task. We first computed the CD of RoPotter-Mesh's cross-section prediction to the final unoccluded cross-section. We then rotated the predicted cross-section by the pottery wheel's rotational axis and computed the CD to the unoccluded final point cloud of the clay without the finger in the scene. We included ablation results with the contact displacement step and the ARAP step as discussed in Section IV-B. To validate that the clay is indeed radially symmetric, we also included a comparison between the unoccluded cross-section that was rotated on the axis of the pottery wheel rotation and compared to the fully observed point cloud of the clay. The observed results in this row provide an upper limit to the achievable performance of the presented methods.

We compared training a diffusion policy with RoPotter-2D and RoPotter-Mesh against using 3D point clouds as outlined in state-of-the-art 3D diffusion policy work [11]. To show the ability of the policy to generalize well to changing initial conditions, we tested the two bowl-making tasks with 250, 300, and 350 g of clay on the pottery wheel. The initial shape of the clay was consistent across the trials and methods to fairly evaluate the methods.

D. Discussion

The ablation with the Fig.5 and Table I showed that each of the steps in RoPotter-Mesh are necessary to produce reliable mesh recovery of the clay cross-section. As shown in the diffusion policy roll-out in Fig. 7, the policies trained in this work on 40 demonstrations can effectively deform the lump of clay into a bowl. Notably, the policy learned to replicate the strategy from most of the demonstrations, where

we started with pushing the clay out from the center and then iterated on lifting the wall of the bowl from either side until the bowl had the approximately desired diameter and height.

When the robot finger first opens up the lump of clay into a concave shape, the resulting outer diameter depends on the amount of clay present, because the diameter is governed by the amount of clay that is available to be displaced laterally. Because the initial outer diameter varies with clay mass, different trajectories are required for the following phase, during which the robot finger lifts the clay walls up from either side. If these trajectories are not adjusted accordingly, the resulting shape may still have the same diameter, as the robot finger can directly constrain this dimension. However, the height may be significantly different. This is because the height of the bowl depends on how much the robot finger pushes on the walls of the bowl.

In our experiments, we observe distinct differences in the policies' abilities to adjust their trajectories given initial clay masses that are slightly out of the distribution of their training data. Across the board, the 3D encoder policy seemed to adjust its trajectory less to the observed initial state, consistently resulting in final bowl shapes that were too tall when a large amount of clay was provided, and too short when a small amount of initial clay was provided. The bowls produced by the 2D and mesh encoder policies also had the same challenges, but to a lesser degree. Most obviously, for the tall bowl task, the bowl produced by the 3D encoder policy was much shorter than the others, and for much of the trial we noticed that the 3D encoder policy seemed unable to deduce a trajectory that qualitatively resembled the demonstrations once it completed the clay-opening step. Quantitatively, we show in Fig. 6 that for all tasks performed with 250g or 350g of clay, the CD distribution between the policy-produced bowls and demonstration bowls is lower for the bowls produced by the two proposed policies. For the tasks performed with 300g of clay, all three policies perform extremely well, and the differences between the CD distributions for these trials are small.

We hypothesized that the policy trained with RoPotter-Mesh would have an advantage when it came to reasoning about the internal shape of the deformed clay as it directly reasons about the contact between the finger and the clay to update the clay states even when it may not be visible from the external perspective. With the experiments with wide bowl-making task, the hypothesis was validated. Additionally, since the mesh is deformed over time but retains information between timesteps, RoPotter-Mesh most likely acted as a filter for observation noise. This may help the continuity of the predicted trajectories over time, and qualitatively we noticed the mesh policy seemed to progress in the tasks more quickly than the other policies with fewer erratic actions compared to the baseline and RoPotter-2D policies. To summarize, the RoPotter-Mesh policy performed 44.4% better on the wide bowl-making task with 250 g of clay compared to the baseline 3D policy and overall performed best in the wide bowl-making task. The baseline policy only performed best on the tall bowl-making task with 300 g of clay; however, the proposed methods also performed exceptionally well on this task, indicating that all three policies may have reached upper-limit of performance. The semantic metric scores were consistent with the geometric metric, which can be explained by the fact that the semantic goals of pottery-making were captured well by the geometric task definition.

VII. Conclusion

In this work, we present RoPotter, a behavior cloning pipeline for learning a policy for the novel task of robot pottery-making, aided by dimensional reduction of the state space and structural priors. We demonstrated that the RoPotter-Mesh pipeline can recover the occluded points of the clay during continuous clay manipulation tasks with a pottery wheel. The results also showed that using the cross-sectional representation of the clay shape with RoPotter-2D and RoPotter-Mesh allowed the behavior cloning policies trained on them to achieve better performance on the pottery-making tasks even with varying initial conditions. An extension of the work will be toward learning goal-conditioned policies for RoPotter. Similar to the challenges of the metrics defined in Section VI-B, defining what is an intuitively useful goal representation for the bowl will be a challenge that we will first approach with optimizing over the pre-trained vision-language representations such as CLIP [31]. We also hope to build on these works toward enabling human-robot collaborative sculpting similar to the recent emergence of collaborative human-robot painting [33].

Acknowledgements

This work is supported by NSF Graduate Research Fellowship under Grant No. DGE2140739 and by the Technology Innovation Program (20018112, Development of autonomous manipulation and gripping technology using imitation learning based on visual tactile sensing) funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea). We thank John Zhang, Sofia Kwok, Eliot Xing, Jeong Hun Lee and Peter Schaldenbrand for the discussions and feedback.

References

[1] H. Yin, A. Varava, and D. Kragic, "Modeling, learning, perception, and control methods for deformable object manipulation," *Science Robotics*, vol. 6, no. 54, p. eabd8803, 2021. 1, 2

[2] H. Shi, H. Xu, Z. Huang, Y. Li, and J. Wu, "Robocraft: Learning to see, simulate, and shape elasto-plastic objects in 3d with graph networks," *The International Journal of Robotics Research*, vol. 43, no. 4, pp. 533–549, 2024. 1, 2, 3

[3] H. Shi, H. Xu, S. Clarke, Y. Li, and J. Wu, "Robocook: Long-horizon elasto-plastic object manipulation with diverse tools," in *Conference on Robot Learning*, pp. 642–660, PMLR, 2023. 2

[4] A. Bartsch, C. Avra, and A. B. Farimani, "Sculptbot: Pre-trained models for 3d deformable object manipulation," *arXiv preprint arXiv:2309.08728*, 2023. 1, 2

[5] X. Lin, Y. Wang, Z. Huang, and D. Held, "Learning visible connectivity dynamics for cloth smoothing," in *Conference on Robot Learning*, pp. 256–266, PMLR, 2022. 1

[6] F. Torabi, G. Warnell, and P. Stone, "Behavioral cloning from observation," in *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pp. 4950–4957, 2018. 1, 4

[7] J. Francis, N. Kitamura, F. Labelle, X. Lu, I. Navarro, and J. Oh, "Core challenges in embodied vision-language planning," *Journal of Artificial Intelligence Research*, vol. 74, pp. 459–515, 2022.

[8] Y. Hu, Q. Xie, V. Jain, J. Francis, J. Patrikar, N. Keetha, S. Kim, Y. Xie, T. Zhang, Z. Zhao, *et al.*, "Toward general-purpose robots via foundation models: A survey and meta-analysis," *arXiv preprint arXiv:2312.08782*, 2023. 1

[9] P. Florence, C. Lynch, A. Zeng, O. A. Ramirez, A. Wahid, L. Downs, A. Wong, J. Lee, I. Mordatch, and J. Tompson, "Implicit behavioral cloning," in *Conference on Robot Learning*, pp. 158–168, PMLR, 2022. 1

[10] C. Chi, S. Feng, Y. Du, Z. Xu, E. Cousineau, B. Burchfiel, and S. Song, "Diffusion policy: Visuomotor policy learning via action diffusion," *arXiv preprint arXiv:2303.04137*, 2023. 1, 2, 4

[11] Y. Ze, G. Zhang, K. Zhang, C. Hu, M. Wang, and H. Xu, "3d diffusion policy," *arXiv preprint arXiv:2403.03954*, 2024. 1, 2, 4, 6

[12] P. Boonvisut and M. C. Çavuşoğlu, "Estimation of soft tissue mechanical parameters from robotic manipulation data," *IEEE/ASME Transactions on Mechatronics*, vol. 18, no. 5, pp. 1602–1611, 2012. 2

[13] P. Guler, K. Pauwels, A. Pieropan, H. Kjellström, and D. Kragic, "Estimating the deformability of elastic materials using optical flow and position-based dynamics," in *2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids)*, pp. 965–971, IEEE, 2015. 2

[14] U. Yoo, Y. Liu, A. D. Deshpande, and F. Alamabeigi, "Analytical design of a pneumatic elastomer robot with deterministically adjusted stiffness," *IEEE robotics and automation letters*, vol. 6, no. 4, pp. 7773–7780, 2021. 2

[15] 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*, pp. 652–660, 2017. 2

[16] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017. 2

[17] X. Yu, L. Tang, Y. Rao, T. Huang, J. Zhou, and J. Lu, "Point-bert: Pre-training 3d point cloud transformers with masked point modeling," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19313–19322, 2022. 2

[18] U. Yoo, Z. Lopez, J. Ichnowski, and J. Oh, "Poe: Acoustic soft robotic proprioception for omnidirectional end-effectors," 2024. 2, 3, 4

[19] U. Yoo, H. Zhao, A. Altamirano, W. Yuan, and C. Feng, "Toward zero-shot sim-to-real transfer learning for pneumatic soft robot 3d proprioceptive sensing," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 544–551, IEEE, 2023. 2, 5

[20] H. Zhang, J. Ichnowski, D. Seita, J. Wang, H. Huang, and K. Goldberg, "Robots of the lost arc: Self-supervised learning to dynamically manipulate fixed-endpoint cables," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4560–4567, IEEE, 2021. 2

[21] M. Yan, Y. Zhu, N. Jin, and J. Bohg, "Self-supervised learning of state estimation for manipulating deformable linear objects," *IEEE robotics and automation letters*, vol. 5, no. 2, pp. 2372–2379, 2020. 2

[22] X. Lin, Y. Wang, J. Olkin, and D. Held, "Softgym: Benchmarking deep reinforcement learning for deformable object manipulation," in *Conference on Robot Learning*, pp. 432–448, PMLR, 2021. 2

[23] S. Duenser, R. Poranne, B. Thomaszewski, and S. Coros, "Robocut: Hot-wire cutting with robot-controlled flexible rods," *ACM Transactions on Graphics (TOG)*, vol. 39, no. 4, pp. 98–1, 2020. 2

[24] O. Sorkine and M. Alexa, "As-rigid-as-possible surface modeling," in *Symposium on Geometry processing*, vol. 4, pp. 109–116, Citeseer, 2007. 4

[25] Z. Levi and C. Gotsman, "Smooth rotation enhanced as-rigid-as-possible mesh animation," *IEEE transactions on visualization and computer graphics*, vol. 21, no. 2, pp. 264–277, 2014. 4

[26] K. Crane, F. de Goes, M. Desbrun, and P. Schröder, "Digital geometry processing with discrete exterior calculus," in *ACM SIGGRAPH 2013 courses*, SIGGRAPH '13, 2013. 4

[27] S. Schaal, "Learning from demonstration," *Advances in neural information processing systems*, vol. 9, 1996. 4

[28] Y. Zhou, C. Barnes, J. Lu, J. Yang, and H. Li, "On the continuity of rotation representations in neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5745–5753, 2019. 4

[29] J. Zhang, X. Chen, Z. Cai, L. Pan, H. Zhao, S. Yi, C. K. Yeo, B. Dai, and C. C. Loy, "Unsupervised 3d shape completion through gan inversion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1768–1777, 2021. 5

[30] T. Wu, L. Pan, J. Zhang, T. Wang, Z. Liu, and D. Lin, "Density-aware chamfer distance as a comprehensive metric for point cloud completion," in *Proceedings of the 35th International Conference on Neural Information Processing Systems*, pp. 29088–29100, 2021. 6

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PMLR, 2021. 6, 7

[32] M. Nimier-David, D. Vicini, T. Zeltner, and W. Jakob, "Mitsuba 2: A retargetable forward and inverse renderer," *ACM Transactions on Graphics (ToG)*, vol. 38, no. 6, pp. 1–17, 2019. 6

[33] P. Schaldenbrand, G. Parmar, J.-Y. Zhu, J. McCann, and J. Oh, "Cofrida: Self-supervised fine-tuning for human-robot co-painting," *arXiv preprint arXiv:2402.13442*, 2024. 7