

SculptBot: Pre-Trained Models for 3D Deformable Object Manipulation

Alison Bartsch¹, Charlotte Avra¹, and Amir Barati Farimani¹

Abstract—Deformable object manipulation presents a unique set of challenges in robotic manipulation by exhibiting high degrees of freedom and severe self-occlusion. Choosing state representations for materials that exhibit plastic behavior, like modeling clay or bread dough, is also difficult because they permanently deform under stress and are constantly changing shape. In this work, we investigate each of these challenges using the task of robotic sculpting with a parallel gripper. We propose a system that uses point clouds as the state representation and leverages a pre-trained point cloud reconstruction transformer to learn a latent dynamics model to predict material deformations given a grasp action. We design a novel action sampling algorithm that reasons about geometrical differences between point clouds to further improve the efficiency of model-based planners. All data and experiments are conducted entirely in the real world. Our experiments show the proposed system is able to successfully capture the dynamics of clay, and is able to create a variety of simple shapes. Videos and additional figures are available on our project page at: <https://sites.google.com/andrew.cmu.edu/sculptbot>

I. INTRODUCTION

Deformable objects are prevalent in tasks related to cooking [1]–[4], manufacturing [5]–[7], medical procedures [8]–[12], and more. It is therefore necessary to develop robotic systems that can effectively handle and manipulate these objects before being deployed in these environments. However, due to their mechanical properties, deformable objects permanently deform with robot interaction, increasing dynamic complexity and the difficulty of state estimation. Therefore, unlike rigid object manipulation, common assumptions to estimate the pose and shape of an object from a single viewpoint, even with self-occlusion, cannot be made when perceiving deformable objects, such as clay, that have no inherent shape.

There have been numerous works focused on modeling and manipulating deformable one-dimensional objects (DOOs) or deformable linear objects (DLOs) such as rope, cables, sutures, and wires, [13]–[18] and 2D deformable objects such as cloth [19]–[25], however systems handling 3D deformable objects, such as modeling clay, remain relatively underexplored, with [26], and the follow-up work [27] being the only known soft body manipulation work conducted primarily on physical robot hardware. In this work, we focus on the challenge of predicting how 3D deformable objects, in this case modeling clay, will deform given a robotic action through the task of sculpting clay with a parallel gripper. We propose a novel method leveraging an existing large pre-trained model for point cloud reconstruction, Point-BERT

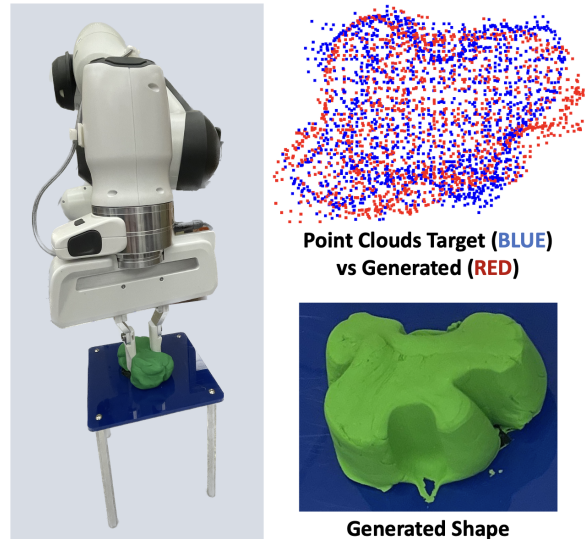


Fig. 1. SculptBot is a pipeline that leverages the pre-trained point cloud model Point-BERT to learn a latent dynamics model that predicts the next state point cloud of the clay conditioned on the current point cloud and grasp action. When combined with planning methods, SculptBot is able to recreate a variety of 3D shapes. Left shows the robot actively deforming the clay. Right shows the shape the system created compared to the target point cloud.

[28], to provide a quality latent embedding of the clay states without requiring any training on our particular dataset. We use this embedding to train a latent forward dynamics model to predict the next state of the clay given the current state and a grasp action. The key contributions of this work are as follows:

- We present the first method to our knowledge leveraging large pre-trained models for deformable object manipulation.
- We propose a novel action sampling algorithm that uses the geometrical relationships between point clouds to select candidate actions to achieve the desired deformations.
- We present a robust system for sculpting clay with a parallel gripper that is able to successfully replicate target shapes, including those that require changes in the z-axis.

II. RELATED WORKS

Representations of 3D Data: Point clouds [29] are often used to represent 3D data as they can represent relatively complex objects with a small number of elements. 3D data can also be represented as voxels [30], where a 3D shape is discretized into a set of cubes of a specific size. Voxel grids

¹With the Department of Mechanical Engineering, Carnegie Mellon University {abartsch, cavra, afariman}@andrew.cmu.edu

can then be easily processed using 3D volumetric convolutional neural networks. However, 3D voxel representations tend to be memory intensive and the voxelization process can lead to information loss. A more recent alternative representation is learned signed distance functions (SDF) [31] which represents the shape with a continuous volumetric field. Researchers have also developed Neural Radiance Fields (NeRFs), neural network-based representations that are able to synthesize novel views of 3D scenes [32], and follow-up works have extended NeRFs to handle dynamics scenes [33].

Point Cloud Networks: Point clouds are permutation invariant, and thus special neural networks are needed to handle these inputs. PointNet [34] and PointNet++ [35] use a shared MLP to learn local features of the point cloud. Point transformer [36] leverages self attention to learn the local features of the point cloud. More recently, numerous works have created novel self supervised learning methodologies to learn robust encodings of point clouds. PointMAE [37] extends the concept of image masking for pretraining into the point cloud domain, where the point cloud is clustered and randomly masked. The follow-up work of PointM2AE [38] includes hierarchical masking, motivated by encouraging the model to learn a variety of higher-level and fine-grained features of the point cloud. Point-BERT [28] devises a BERT-style pretraining method to predict the tokens of the masked patches of the point cloud.

Learned Dynamics Models: There have been many successful works focused on learning dynamics models from data. In [39], researchers learn particle-based dynamics models for a variety of materials as an alternative to particle-based simulators. Similarly, in [40], researchers learn a graph-based particle dynamics model for fluids. Learned dynamics models are also prevalent within the domain of model-based reinforcement learning. In DREAMER [41], and the follow-up DREAMER-V2 [42], a learned dynamics model is used to predict the agent’s future success by leveraging dynamics model rollouts.

Manipulation of Deformable Objects: The task of manipulating deformable objects has been investigated previously. In RoboCraft [26], researchers have similarly learned a dynamics model to deform clay into target shapes. However, in this work the authors focus their evaluation on a 2D alphabet test set. We hope to explore a wider variety of more complex, 3D target shapes in this work. In an extension to RoboCraft, researchers present RoboCook [27], which uses a similar dynamics model with a set of diverse tools to create the 2D alphabet shapes. In [20], researchers train a latent dynamics model with contrastive learning to handle 2D deformable objects, such as cloth. A similar task of folding cloth was explored in [43]. In [44], researchers focus on handling semi-rigid deformable objects, such as silicone spatulas. Beyond the tasks of handling deformable objects, research has also focused on building better simulators, such as SoftGym [45], or PlasticineLab [46], learning material properties such as elasticity [47], and improving the possibility of sim-to-real transfer with these complex objects [48].

III. METHODOLOGY

Our method primarily consists of three modules - the vision system, the pre-trained tokenizer from Point-BERT [28], and the learned latent dynamics model. One of the key considerations with this work was the choice of representation for the 3D shape of the clay. Point clouds were a natural choice due to the richness of information, while minimizing memory usage. The choice of state representation for the clay directly informed both the vision system as well as the dynamics model. Point clouds are an unstructured representation, which requires a special neural network architecture that can handle this.

A. Vision System

Our vision system consists of 4 Intel RealSense D415 RGB-D cameras (shown in Figure 5). We found that the standard technique for calculating extrinsics between cameras using a calibration checkerboard pattern resulted in relatively inaccurate combined point clouds. Instead, we used an asymmetrical 3D object to calculate the extrinsics between cameras using global point cloud registration, specifically RANSAC [49], and fine-tuned them with the point-to-point variant of the Iterative Closest Point algorithm (ICP) [50]. This significantly improved our final combined point cloud accuracy, to approximately 0.5 cm error, comparable to that of RoboCraft’s vision module [26]. Once we have a quality 3D point cloud of the environment, we perform a simple position-based crop to isolate the elevated stage (shown in Figure 5), and use color-based cropping in the LAB colorspace to isolate the clay from the table. We then extract the points closest to the base of the clay (by indexing points below a z threshold). We use this base shape outline to crop a plane of points to form the bottom of the clay. We combine this base plane with the original cloud to form a fully enclosed point cloud shell. This processing is based on the assumption that the clay is always resting on the elevated stage, thus the base of the clay will be at that z -position. Once we have the clay shell, we uniformly downsample the point cloud to 2048 points. A full visualization of the point cloud preprocessing pipeline is shown in Figure 3.

B. Pre-Trained Point-BERT Tokenizer

One of the challenges with manipulating 3D deformable objects is that models handling 3D representations typically require a significant amount of data to properly train. However, when working within the robotic space, collecting sufficient data can be incredibly time consuming. To address this challenge, we leverage the tokenizer from Point-BERT, which is pre-trained on the ShapeNet dataset [51], a collection of 3D point clouds of 3,135 common categories. We find in practice that the discrete variational autoencoder (dVAE) from Point-BERT learned a sufficient latent representation, that needs no fine-tuning on the clay-specific dataset. A visualization of the reconstruction quality of our real-world clay data is shown in Figure 4.

Point-BERT is originally a pre-training strategy for point cloud transformers. First, the point clouds are grouped into

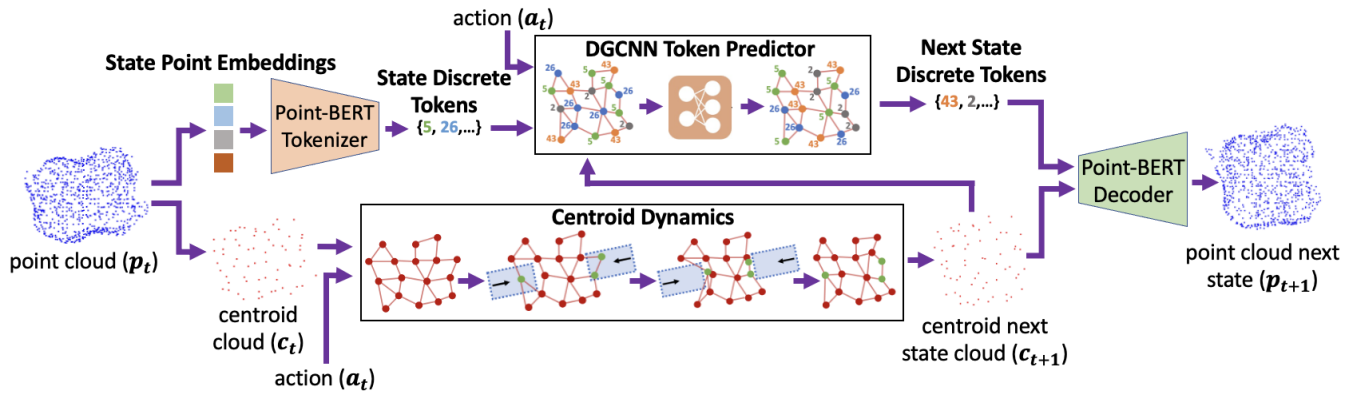


Fig. 2. The dynamics prediction pipeline. We first use farthest point sampling and k-nearest neighbors to cluster the original point cloud into 64 clusters. These 64 clusters become a much smaller and sparser point cloud. We then use the pre-trained dVAE from Point-BERT to tokenize each cluster. The centroid point cloud is passed through a simple physics-based dynamics approximator to predict the next state centroid point cloud given the grasp action. This predicted next state centroid point cloud is passed to the point token predictor dynamics model along with the state centroid tokenization and the grasp action. The point token predictor predicts the tokens for each next state centroid, which represent the geometrical structure of the points within that region of the cloud. These predicted tokens along with the predicted centroid point cloud are then passed through the dVAE decoder to reconstruct the full dense predicted next state point cloud.

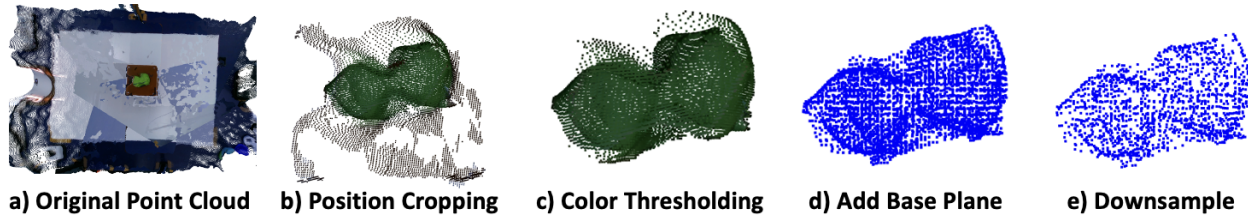


Fig. 3. The full preprocessing pipeline for the point clouds. a) The original point cloud of the scene. b) the scene after position-based cropping to eliminate the elevated stage. c) The point cloud after color-based thresholding. d) The point cloud after removing statistical outliers and adding in a base plane. e) The point cloud uniformly downsampled to 2048 points.

local patches, or clusters using farthest-point sampling and k-nearest neighbors. Each cluster has a positional embedding to maintain the regional locations for reconstruction. The points in each local patch are subtracted with the centroid position to eliminate bias and only represent the regional point cloud structure. Each cluster is then passed through a small PointNet [34] to output a discrete point embedding that describes the general point cloud shape within that region. These discrete tokens and the positional embeddings are then passed through the decoder to reconstruct the original point cloud. In this work, we use the dVAE tokenizer from Point-BERT and train a dynamics model in the latent space of the positional and point embeddings.

C. Latent Dynamics Model

Given the latent representation from the dVAE of Point-BERT consists of two components (positional embedding of the clusters, and the point tokens for each cluster), we developed two separate dynamics modules to propagate the dynamics of a grasp action. First, we developed a simple physics approximator that propagates the point cloud based on the trajectory of the gripper fingertips during the grasp. Next, based on the next state centroids and the grasp action, we train a DGCNN [52] token predictor model to predict how

each cluster’s point token changes given the grasp action. When combined with the pre-trained dVAE to encode the state point cloud and reconstruct the predicted cloud, the system predicts the next state of the clay point cloud given the current state point cloud and the grasp action. The grasp action is defined as the x,y,z position of the end-effector, the rotation about the z-axis, and the distance between the fingertips.

1) *Centroid Dynamics*: Each cluster is passed through a physics-based dynamics approximator that propagates the point clouds based on the grasp action. This dynamics approximator is light weight, moving the points that are inliers in the approximated gripper trajectory mesh based on the grasp pose. These points are moved normal to the surface of the gripper the distance between their initial location and the gripper’s final position. Additionally, we impose some distance constraints on the point cloud, ensuring that the points cannot exceed a pre-specified distance. After we approximate the point motion due to the grasp, we update the point cloud to reinforce these distance constraints. While this dynamics approximator is relatively naive, when paired with the second stage learned dynamics model we expect it to account for the simplicity.



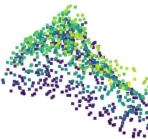
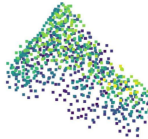
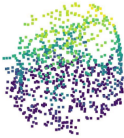


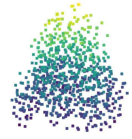
Shape	Input	Reconstruction	CD
X			0.004077
T			0.004275
Cylinder			0.005135
Pyramid			0.005540

Fig. 4. The dVAE from Point-BERT provides quality reconstructions of the real-world clay point clouds without requiring any finetuning on our dataset. This allows us to train a latent dynamics model to predict the material deformation in the Point-BERT embedding space.

2) *DGCNN Point Token Dynamics*: The token dynamics model takes the predicted next state positional embeddings from the centroid dynamics model, the state point token embeddings and the grasp action as input to predict the next state point tokens that minimizes the Chamfer distance between the reconstructed predicted next state and ground truth point clouds. The Chamfer distance is the sum of squared distances between the nearest neighbor correspondences between point clouds, and is a common point cloud reconstruction metric shown below in Equation 1. The predicted next state point tokens are passed back through the dVAE decoder to reconstruct the full predicted next state point cloud.

$$d_{CD}(X, Y) = \sum_{x \in X} \min_{y \in Y} \|x - y\|_2^2 + \sum_{y \in Y} \min_{x \in X} \|x - y\|_2^2 \quad (1)$$

D. Geometry-Informed Action Sampling

We developed an action sampler that leverages geometric knowledge of the point clouds to efficiently sample quality potential actions. The concept of the geometric sampler is to identify the regions of the current point cloud that are most different from the target cloud, and search for actions that would apply a change to the state in those areas. Especially as the action space increases in dimensionality, it is important to have a more efficient action sampling method than random parameter selection. The full algorithmic description of the action sampler is shown in Algorithm 1. We first utilize k-means [53] to cluster each point cloud into N regions. We

Algorithm 1 Geometric Sampler

Input: P_{state} ▷ current state point cloud
Input: P_{target} ▷ target shape point cloud
Input: $N_{clusters}$ ▷ algorithm parameter (64)
Input: EE_{width} ▷ width of gripper

$\mu_{state} = KMeans(P_{state}, N_{clusters})$
 $\mu_{target} = KMeans(P_{target}, N_{clusters})$
 $\vec{\delta} = (\mu_{target} - \mu_{state})$
 $dist_s = |\mu_{state, i} - \mu_{target, i}|$
Sample pair of centroids with probability $p = \frac{dist_s}{\sum dist_s}$
 $(x, y, z) = \mu_{state} + \frac{1}{2}\vec{\delta}EE_{width}$
 $r_z = \arctan2(\vec{\delta}_y, \vec{\delta}_x)$
 $d = \frac{1}{2}|\mu_{target} - (x, y, z)|$
 $actions = [x, y, z, r_z, d]$

Return: $actions$

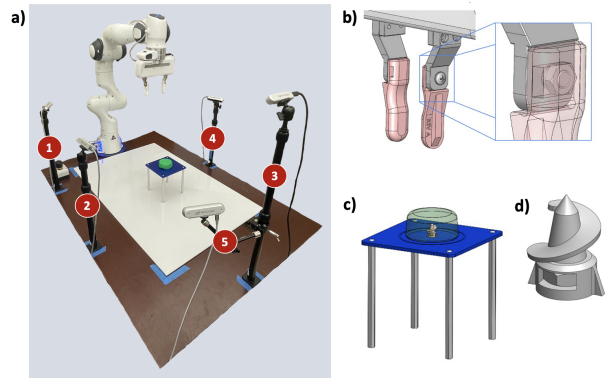


Fig. 5. The hardware setup: a) general workspace with 4 Intel RealSense D415 cameras for 3D scene reconstruction (1-4), and one Intel RealSense D455 camera for capturing video (5), b) the custom 3D printed fingertips, c) the elevated stage, d) screw-in anchor for the clay on the elevated stage.

then pair the clusters between the two point clouds based on those with the smallest Euclidean distance. The cluster pairs are sampled with a probability proportional to the distance, and an action is generated to push the cluster in the state point cloud towards that in the target point cloud.

IV. EXPERIMENTAL SETUP DETAILS

All data collection and experiments were conducted on a physical 7-axis Franka Emika Panda manipulator equipped with a two-finger parallel gripper. There are 4 Intel RealSense D415 cameras mounted to the workspace. The clay sits atop an elevated stage in the center of the workspace, and is loosely attached to the stage with a small screw to ensure the clay generally remains on the stage. Details of the hardware setup are shown in Figure 5. It is important to note that in our current setup there is no force information. We assume the clay volume remains consistent throughout data collection, always initialized to a short cylinder 6 cm in diameter and 2.5 cm in height.

A. Data Collection

We collect two separate datasets to train alternate versions of the latent dynamics model, 1) a random action dataset, 2)

a human demonstration dataset. The random action dataset is collected by randomly sampling action parameters and executing the generated grasps. The point cloud of the clay is collected before and after each grasp. We consider a grasp trajectory to be 10 consecutive grasps applied to the clay before the clay is re-initialized to the starting shape. We collect a total of 15 trajectories, totalling approximately 30 minutes of data collection time. To increase the size of our dataset, we augment the dataset by applying varying rotation transforms about the z-axis to the point cloud (in intervals of 6°) and the grasp action. This augmentation strategy is sound, as we have the assumption that the clay always remains fixed on the elevated stage surface. The human demonstration dataset is collected with kinesthetic teaching where the human controls the end-effector position, rotation, and the distance between the fingertips. We chose this over alternative human demonstration collection techniques, such as [54], because we found kinesthetic teaching produced more accurate sculptures. The human operator collected 5 trajectories of varying length for creating a cone, tall cylinder, line, pyramid, T, and X. The same rotation augmentation technique is applied to these trajectories. While this requires more human time than the random actions, we collected this dataset to explore differences in model performance trained on each dataset. These human demonstrations are simply unlabeled state and action pairs to train our point cloud dynamics model, meaning at test time we are not restricted to evaluate only on these target shapes.

V. RESULTS

Once we have the trained latent dynamics model, we can plan action trajectories to reach a variety of target point clouds. In section V-A we first evaluate the quality of the learned latent dynamics model. In section V-B we have a human subject teleoperate the robot to generate a variety of shape targets to investigate the difficulty of the task with the robot embodiment. Finally, in section V-C we deploy the learned dynamics model with model predictive control for the sculpting task. We evaluate the performance of the dynamics model trained on the random and human demonstration datasets, as well as the proposed geometric sampling strategy compared to random shooting.

A. Evaluation of the Dynamics Model

To evaluate the quality of the next state point cloud predictions of our dynamics model, we report the Chamfer Distance (CD) between the predicted next state and ground truth next state of the combined test sets of the human demonstration and random action datasets. A visualization of the predictions for a few states are shown in Figure 6. The performance of the dynamics models on the entirety of the dataset are shown below in Table I. The dynamics model trained on the human demonstration dataset had a lower mean Chamfer distance and lower variance for the next state predictions. However, this difference is relatively small, and may not be great enough to motivate the additional human effort of collecting the demonstrations.

TABLE I
CHAMFER DISTANCE (CD) FOR THE DYNAMICS MODELS TRAINED ON THE DIFFERENT DATASETS COLLECTED.

	Centroid Next State CD	Full Next State CD
Human Demos	0.0450 ± 0.0069	0.0109 ± 0.0024
Random Actions	0.0451 ± 0.0069	0.0113 ± 0.0028




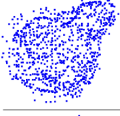

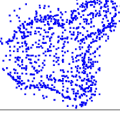
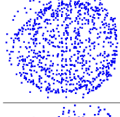

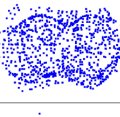



State	Predicted Next State	Ground Truth Next State	Prediction CD
			0.0115
			0.0109
			0.0131
			0.0111

Fig. 6. A visualization of some next state predictions on the test set by the latent dynamics model trained on the human demonstration dataset. This is a top-down view of 3D point clouds. The results show that the model is able to capture and predict the large geometric changes caused by various grasp actions. However, some of the details may not be captured, likely due to the shape reconstruction, as the quality appears similar to some of the detail lost during reconstruction, shown in Fig 4. This loss is a side effect of leveraging the pre-trained model from Point-BERT, and is not sufficient to justify training our own point cloud encoder, as it would require substantially more data.

B. Human Teleoperation Baseline

The task of deforming a small block of clay with a parallel gripper is incredibly difficult, as a relatively simple action results in complicated changes to the clay shape. To better understand what quality of results we could expect for this task, we conducted human teleoperation experiments. We used a simple kinesthetic teaching system where the human controls the end-effector position, rotation, and the distance between the fingertips. The results of this experiment are shown in Figure 7 and are compared to the results of our dynamics system. Although sculpting with a parallel gripper can be unintuitive, the human teleoperators were able to successfully create a range of simple shapes.

C. Hardware Deployment

We deploy the fully trained dynamics model on hardware for the sculpting task by unrolling trajectories with model predictive control (MPC). For these experiments, we plan one step into the future, and selecting the action that minimizes the Chamfer distance between the expected next state and the target point cloud. All target point clouds are created from human teleoperation to ensure the goals are feasible to create with our hardware. We first compare the

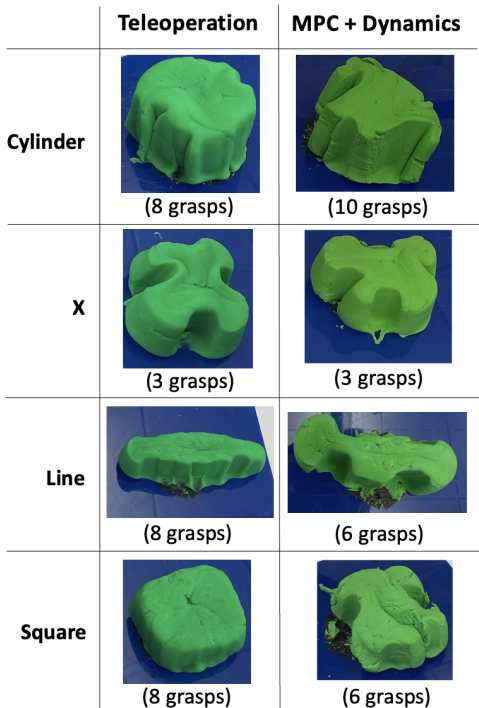


Fig. 7. The shapes a human was able to create when guiding the robot and its parallel gripper (Left) compared to the shapes the human demonstration trained dynamics model combined with MPC and geometric sampling was able to create (Right). Visually, it is clear that our system is not able to perform better than human oracle, but based on the reconstruction metrics, it is able to successfully recreate the key structural aspects of the shapes.

performance of the proposed geometric action sampler as compared to random shooting (shown in Table II). For these experiments, we are using the dynamics model trained on the human demonstration dataset. For the random shooting action sampling, we randomly sample 2500 actions for MPC, whereas with the geometric-based sampler we only need to sample 35 actions. For each comparison, we run the system 3 times for each target, and report the mean and standard deviation. The geometric-based sampler produces sculptures of similar quality compared to random shooting, while being significantly more sample efficient. Additionally, the geometric-based sampler achieves these results with on average fewer grasp actions as compared to random shooting. Next, we evaluate the performance of the models trained on the human demonstration and random datasets with MPC with geometric-based sampling (shown in Table III) on a variety of target shapes. The model trained with the human demonstration dataset performs substantially better on a few of our test shapes, particularly 'X' and 'triangle'. Our human demonstration dataset did include examples of humans creating an 'X', but not 'triangle'. As the human demonstration dataset was used to train a single-step dynamics model, this means that the model trained on this data is able to more accurately predict the dynamics of the actions that are useful for creating these shapes. In the future, it would be interesting to explore a middle-ground dataset collection technique that

TABLE II
COMPARING SCULPTING PERFORMANCE OF THE PROPOSED GEOMETRIC ACTION SAMPLING AS COMPARED TO RANDOM SHOOTING.

Target	Sampler	# Grasps	CD
X	Random	4.2 ± 0.84	0.0024 ± 0.0003
	Geometric	3.6 ± 0.89	0.0029 ± 0.0007
T	Random	10.7 ± 3.1	0.0069 ± 0.0012
	Geometric	6.7 ± 3.5	0.0058 ± 0.002
Square	Random	3.7 ± 0.6	0.0044 ± 0.0005
	Geometric	4.7 ± 1.5	0.0042 ± 0.0004
Line	Random	10.7 ± 3.1	0.00739 ± 0.001
	Geometric	5.2 ± 2.5	0.0068 ± 0.0015
Cylinder	Random	11.5 ± 2.1	0.0027 ± 0.0003
	Geometric	12.7 ± 2.5	0.0041 ± 0.0002
Triangle	Random	4.0 ± 2.0	0.0048 ± 0.0009
	Geometric	2.3 ± 0.6	0.0031 ± 0.0005

TABLE III
COMPARING SCULPTING PERFORMANCE OF THE DYNAMICS MODELS TRAINED ON THE RANDOM ACTION DATASET AND THE HUMAN DEMONSTRATION DATASET WITH THE GEOMETRIC SAMPLER.

Target	Dataset	# Grasps	CD
X	Human Demo	3.0 ± 0.0	0.0025 ± 0.0007
	Random	3.7 ± 0.6	0.0041 ± 0.0005
T	Human Demo	6.7 ± 3.5	0.0058 ± 0.0018
	Random	4.0 ± 2.6	0.0060 ± 0.0013
Square	Human Demo	4.7 ± 1.5	0.0042 ± 0.0004
	Random	3.3 ± 0.6	0.0039 ± 0.0011
Line	Human Demo	5.3 ± 2.5	0.0068 ± 0.0015
	Random	3.3 ± 0.6	0.0054 ± 0.0007
Cylinder	Human Demo	12.7 ± 2.5	0.0041 ± 0.0002
	Random	10.5 ± 0.7	0.0051 ± 0.0015
Triangle	Human Demo	2.3 ± 0.6	0.0031 ± 0.0005
	Random	3.3 ± 1.2	0.0071 ± 0.0015

is more intelligent than random actions, but not as time consuming as human demonstrations. A visualization of the quality of some of the shapes the system is able to create are shown in Figure 7. Our system is able to reliably generate replications of target sculptures with reasonable Chamfer distances that capture the key structural aspects of the target shape.

VI. CONCLUSION

In this work we present SculptBot, a system that leverages pre-trained point cloud reconstruction models to learn a latent dynamics model for 3D deformable object manipulation. Through our experiments, we demonstrate that our model is able to successfully capture the dynamics of the clay, and is able to create a variety of simple shape sculptures when combined with model predictive control and a geometric-based action sampling algorithm. One of the key limitations of this work is that a parallel gripper may not be the best tool for the sculpting task. From our human teleoperation experiments, the quality of the shapes our human oracle was able to create are relatively coarse and simplistic. In future work, we hope to investigate alternative sculpting actions, first by expanding the types of actions the robot can take, such as pinch and twist, smooth, etc. Additionally, we hope to provide the robot with a set of sculpting tools and incorporate actions for both additive and subtractive clay sculpting.

REFERENCES

- [1] P. Long, A. A. Moughlby, W. Khalil, and P. Martinet, "Robotic meat cutting," in *Ict-pamm workshop*, 2013.
- [2] A. Dikshit, A. Bartsch, A. George, and A. B. Farimani, "Robochop: Autonomous framework for fruit and vegetable chopping leveraging foundational models," *arXiv preprint arXiv:2307.13159*, 2023.
- [3] A. Petit, V. Lippiello, G. A. Fontanelli, and B. Siciliano, "Tracking elastic deformable objects with an rgb-d sensor for a pizza chef robot," *Robotics and Autonomous Systems*, vol. 88, pp. 187–201, 2017.
- [4] X. Mu, Y. Xue, and Y.-B. Jia, "Robotic cutting: Mechanics and control of knife motion," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 3066–3072.
- [5] K. Kimble, J. Albrecht, M. Zimmerman, and J. Falco, "Performance measures to benchmark the grasping, manipulation, and assembly of deformable objects typical to manufacturing applications," *Frontiers in Robotics and AI*, vol. 9, p. 999348, 2022.
- [6] J. Cortsen, J. A. Jørgensen, D. Sølvason, and H. G. Petersen, "Simulating robot handling of large scale deformable objects: Manufacturing of unique concrete reinforcement structures," in *2012 IEEE International Conference on Robotics and Automation*. IEEE, 2012, pp. 3771–3776.
- [7] N. Lv, J. Liu, and Y. Jia, "Dynamic modeling and control of deformable linear objects for single-arm and dual-arm robot manipulations," *IEEE Transactions on Robotics*, vol. 38, no. 4, pp. 2341–2353, 2022.
- [8] L. Han, H. Wang, Z. Liu, W. Chen, and X. Zhang, "Vision-based cutting control of deformable objects with surface tracking," *IEEE/ASME Transactions on Mechatronics*, vol. 26, no. 4, pp. 2016–2026, 2020.
- [9] Y. Wang, Y. Long, S. H. Fan, and Q. Dou, "Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2022, pp. 431–441.
- [10] P. M. Scheikl, E. Tagliabue, B. Gyenes, M. Wagner, D. Dall'Alba, P. Fiorini, and F. Mathis-Ullrich, "Sim-to-real transfer for visual reinforcement learning of deformable object manipulation for robot-assisted surgery," *IEEE Robotics and Automation Letters*, vol. 8, no. 2, pp. 560–567, 2022.
- [11] F. Liu, Z. Li, Y. Han, J. Lu, F. Richter, and M. C. Yip, "Real-to-sim registration of deformable soft tissue with position-based dynamics for surgical robot autonomy," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 12 328–12 334.
- [12] Y. Li, F. Richter, J. Lu, E. K. Funk, R. K. Orosco, J. Zhu, and M. C. Yip, "Super: A surgical perception framework for endoscopic tissue manipulation with surgical robotics," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 2294–2301, 2020.
- [13] M. Saha and P. Isto, "Manipulation planning for deformable linear objects," *IEEE Transactions on Robotics*, vol. 23, no. 6, pp. 1141–1150, 2007.
- [14] M. Yu, K. Lv, C. Wang, M. Tomizuka, and X. Li, "A coarse-to-fine framework for dual-arm manipulation of deformable linear objects with whole-body obstacle avoidance," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 10 153–10 159.
- [15] J. Zhu, B. Navarro, R. Passama, P. Fraise, A. Crosnier, and A. Cherubini, "Robotic manipulation planning for shaping deformable linear objects with environmental contacts," *IEEE Robotics and Automation Letters*, vol. 5, no. 1, pp. 16–23, 2020.
- [16] 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.
- [17] W. Zhang, K. Schmeckpeper, P. Chaudhari, and K. Daniilidis, "Deformable linear object prediction using locally linear latent dynamics," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 503–13 509.
- [18] H. Lu, Y. Teng, and Y. Li, "Learning latent dynamics for autonomous shape control of deformable object," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–8, 2022.
- [19] Z. Huang, X. Lin, and D. Held, "Self-supervised cloth reconstruction via action-conditioned cloth tracking," *arXiv preprint arXiv:2302.09502*, 2023.
- [20] W. Yan, A. Vangipuram, P. Abbeel, and L. Pinto, "Learning predictive representations for deformable objects using contrastive estimation," in *Conference on Robot Learning*. PMLR, 2021, pp. 564–574.
- [21] A. Longhini, M. Moletta, A. Reichlin, M. C. Welle, D. Held, Z. Erickson, and D. Kragic, "Edo-net: Learning elastic properties of deformable objects from graph dynamics," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 3875–3881.
- [22] R. Jangir, G. Alenyà, and C. Torras, "Dynamic cloth manipulation with deep reinforcement learning," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 4630–4636.
- [23] J. Borràs, G. Alenyà, and C. Torras, "A grasping-centered analysis for cloth manipulation," *IEEE Transactions on Robotics*, vol. 36, no. 3, pp. 924–936, 2020.
- [24] N. Sunil, S. Wang, Y. She, E. Adelson, and A. R. Garcia, "Visuotactile affordances for cloth manipulation with local control," in *Proceedings of The 6th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, K. Liu, D. Kulic, and J. Ichnowski, Eds., vol. 205. PMLR, 14–18 Dec 2023, pp. 1596–1606. [Online]. Available: <https://proceedings.mlr.press/v205/sunil23a.html>
- [25] H. Ha and S. Song, "Flingbot: The unreasonable effectiveness of dynamic manipulation for cloth unfolding," in *Proceedings of the 5th Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, A. Faust, D. Hsu, and G. Neumann, Eds., vol. 164. PMLR, 08–11 Nov 2022, pp. 24–33. [Online]. Available: <https://proceedings.mlr.press/v164/ha22a.html>
- [26] H. Shi, H. Xu, Z. Huang, Y. Li, and J. Wu, "Robocraft: Learning to see, simulate, and shape elasto-plastic objects with graph networks," *arXiv preprint arXiv:2205.02909*, 2022.
- [27] H. Shi, H. Xu, S. Clarke, Y. Li, and J. Wu, "Robocook: Long-horizon elasto-plastic object manipulation with diverse tools," *arXiv preprint arXiv:2306.14447*, 2023.
- [28] 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*, 2022, pp. 19 313–19 322.
- [29] R. B. Rusu and S. Cousins, "3d is here: Point cloud library (pcl)," in *2011 IEEE international conference on robotics and automation*. IEEE, 2011, pp. 1–4.
- [30] Z. Liu, H. Tang, Y. Lin, and S. Han, "Point-voxel cnn for efficient 3d deep learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [31] 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.
- [32] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.
- [33] 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.
- [34] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "Pointnet: Deep learning on point sets for 3d classification and segmentation," 2017.
- [35] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," 2017.
- [36] H. Zhao, L. Jiang, J. Jia, P. Torr, and V. Koltun, "Point transformer," 2021.
- [37] Y. Pang, W. Wang, F. E. Tay, W. Liu, Y. Tian, and L. Yuan, "Masked autoencoders for point cloud self-supervised learning," in *European conference on computer vision*. Springer, 2022, pp. 604–621.
- [38] R. Zhang, Z. Guo, P. Gao, R. Fang, B. Zhao, D. Wang, Y. Qiao, and H. Li, "Point-m2ae: multi-scale masked autoencoders for hierarchical point cloud pre-training," *Advances in neural information processing systems*, vol. 35, pp. 27 061–27 074, 2022.
- [39] Y. Li, J. Wu, R. Tedrake, J. B. Tenenbaum, and A. Torralba, "Learning particle dynamics for manipulating rigid bodies, deformable objects, and fluids," 2019.
- [40] Z. Li and A. B. Farimani, "Graph neural network-accelerated lagrangian fluid simulation," *Computers & Graphics*, vol. 103, pp. 201–211, 2022.
- [41] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi, "Dream to control: Learning behaviors by latent imagination," *arXiv preprint arXiv:1912.01603*, 2019.
- [42] D. Hafner, T. Lillicrap, M. Norouzi, and J. Ba, "Mastering atari with discrete world models," *arXiv preprint arXiv:2010.02193*, 2020.
- [43] Y. Li, Y. Yue, D. Xu, E. Grinspun, and P. K. Allen, "Folding deformable objects using predictive simulation and trajectory opti-

- mization,” in *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2015, pp. 6000–6006.
- [44] Y. Wi, P. Florence, A. Zeng, and N. Fazeli, “Virdo: Visio-tactile implicit representations of deformable objects,” in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 3583–3590.
- [45] X. Lin, Y. Wang, J. Olkin, and D. Held, “Softgym: Benchmarking deep reinforcement learning for deformable object manipulation,” in *Proceedings of the 2020 Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, J. Kober, F. Ramos, and C. Tomlin, Eds., vol. 155. PMLR, 16–18 Nov 2021, pp. 432–448. [Online]. Available: <https://proceedings.mlr.press/v155/lin21a.html>
- [46] Z. Huang, Y. Hu, T. Du, S. Zhou, H. Su, J. B. Tenenbaum, and C. Gan, “Plasticinelab: A soft-body manipulation benchmark with differentiable physics,” *arXiv preprint arXiv:2104.03311*, 2021.
- [47] B. Frank, R. Schmedding, C. Stachniss, M. Teschner, and W. Burgard, “Learning the elasticity parameters of deformable objects with a manipulation robot,” in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 1877–1883.
- [48] J. Matas, S. James, and A. J. Davison, “Sim-to-real reinforcement learning for deformable object manipulation,” in *Proceedings of The 2nd Conference on Robot Learning*, ser. Proceedings of Machine Learning Research, A. Billard, A. Dragan, J. Peters, and J. Morimoto, Eds., vol. 87. PMLR, 29–31 Oct 2018, pp. 734–743. [Online]. Available: <https://proceedings.mlr.press/v87/matas18a.html>
- [49] M. A. Fischler and R. C. Bolles, “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, 1981.
- [50] P. J. Besl and N. D. McKay, “Method for registration of 3-d shapes,” in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.
- [51] A. X. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su *et al.*, “Shapenet: An information-rich 3d model repository,” *arXiv preprint arXiv:1512.03012*, 2015.
- [52] A. V. Phan, M. Le Nguyen, Y. L. H. Nguyen, and L. T. Bui, “Dgcnn: A convolutional neural network over large-scale labeled graphs,” *Neural Networks*, vol. 108, pp. 533–543, 2018.
- [53] J. MacQueen *et al.*, “Some methods for classification and analysis of multivariate observations,” in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA, 1967, pp. 281–297.
- [54] A. George, A. Bartsch, and A. B. Farimani, “Openvr: Teleoperation for manipulation,” *arXiv preprint arXiv:2305.09765*, 2023.