Autonomous Clay Sculpting from Human Demonstrations with Point Cloud Goal Conditioned Diffusion Policy

Alison Bartsch[1], Arvind Car[1], Charlotte Avra[1], and Amir Barati Farimani[1]

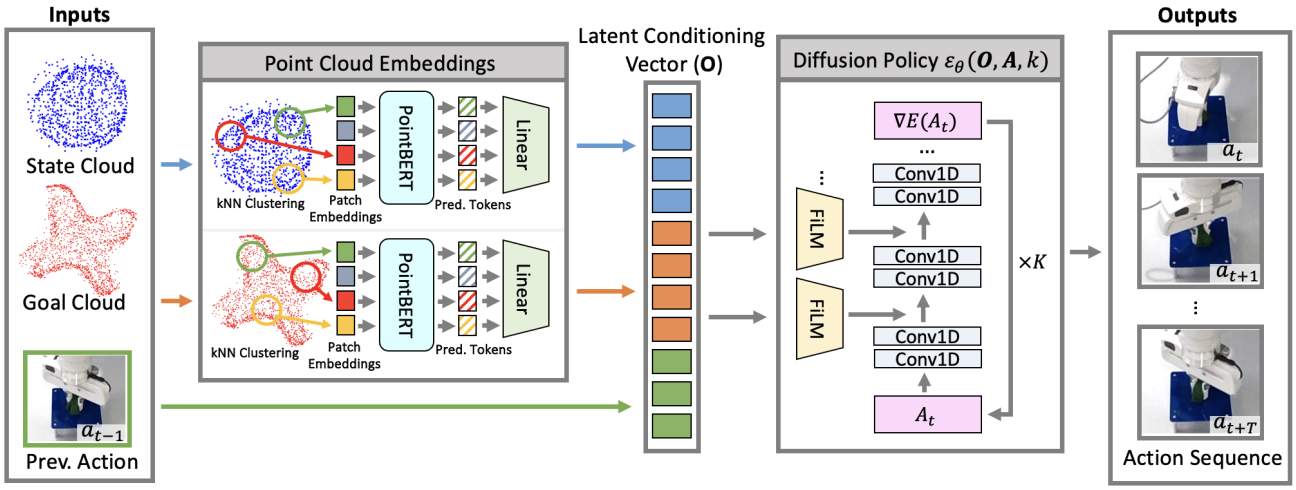

Fig. 1: The learning pipeline of SculptDiff, a modification of diffusion policy [1] to incorporate goal conditioning and leverage 3D point clouds as state representations with PointBERT embeddings [2] for the sculpting task.

***Abstract*— 3D deformable object manipulation remains a challenge due to the difficulties of state estimation, long-horizon planning, and predicting how the object will deform given an interaction. In this work, we propose SculptDiff, a goal-conditioned diffusion-based imitation learning framework that works with point cloud states to directly learn clay sculpting policies for a variety of target shapes. To the best of our knowledge this is the first real-world end-to-end policy for 3D deformable object manipulation. For sculpting videos, see the project website: https://sites.google.com/andrew.cmu.edu/imitation-sculpting/home**

I. INTRODUCTION

Advancements in robotic deformable object manipulation have large-scale implications ranging from manufacturing [3], [4] to surgery [5]. However, deformable object manipulation remains an open challenge within the robotics field due to the complexities of the interaction between the object and robot, as the object permanently changes shape with each grasp. In this work, we aim to explore the challenges of deformable object manipulation with the task of autonomously sculpting clay. The clay sculpting task is a useful benchmark to investigate methods for deformable objects due to the difficulty of the task itself. Firstly, the deformation behavior is difficult to predict as clay has no underlying structure and can be deformed in all three dimensions. When sculpting clay, the system needs to have a representation of the 3D shape, which poses observation and state representation challenges. Additionally, the system needs to have a sense of the goal shape and execute a sequence of actions that result in this final goal shape. This is particularly challenging as multiple sequences of actions can result in the same final 3D shape, but the ordering of the actions themselves are important, presenting a difficult planning problem. Within the realm of deformable object manipulation there has been recent success learning and planning with a dynamics model to predict the complicated interactions between a rigid end-effector and a deformable object. However, for more complicated sculpting tasks with 3D objects, planning with a dynamics model can be very time consuming at test time due to the large state and action space [6], [7], [8]. To address this long planning time, we can instead train a policy to go directly from observations to actions. However, due to the high complexity of deformable objects, it is very challenging to train a robust policy for 3D deformable object sculpting in simulation or in the real-world [9], [10]. This motivates our proposed work, where instead we train a policy directly from human demonstrations to avoid the exploration challenges. In this work, we present SculptDiff, a point cloud-based diffusion policy for the clay sculpting task that can successfully sculpt a 3D target shape from only 10 real-world demonstrations.

II. METHOD

An overview of the SculptDiff pipeline is visualized in Figure 1. In this work, we define the clay sculpting task as applying a sequence of parallel grasps to a piece of clay fixed in the workspace with the goal of replicating a target point

[1]Department of Mechanical Engineering at Carnegie Mellon University (corresponding e-mail: afariman@andrew.cmu.edu)

An extended version of this work is currently under review.

cloud. We collect point cloud states before and after each grasp. The action space is defined as the x, y, z position of the end-effector, the rotation about the z-axis and the distance between the fingertips at the end of the squeeze action. We represent a sculpting trajectory as the sequence of point cloud states and actions given a target shape point cloud.

A. Point Cloud State Representation

The task of sculpting clay into a target shape requires reasoning about 3D geometry. Thus, we hypothesize that our system requires an observation space that explicitly represents this 3D information. Beyond the fact that we are training a policy for an inherently 3D task, past studies have shown that for general cases using point clouds as observations compared with RGB and RGB-D observations show improvement in robot performance on a variety of manipulation tasks [11]. Additionally, point clouds as the state representation allows us to augment our demonstration dataset, minimizing the number of demos that need to be collected to train a quality sculpting policy. Our augmentation strategy is based on the assumption that the clay always remains fixed to the elevated stage (shown in Figure 2). For each demonstration trajectory, we apply a rotational transform about the z-axis in 1° increments to both the state and goal clouds as well as to the parameterized action. This allows us to transform a single human demonstration into 360 varying demonstrations. To acquire the full 3D point cloud of the clay state, we use 4 Intel RealSense D415 cameras that are mounted to a camera cage for simple multi-camera calibration. Our physical camera setup is shown in Figure 2. We use the same point cloud processing pipeline as in our previous work [6].

B. SculptDiff: Point Cloud Diffusion Policy

We combine diffusion policy with point cloud state and goal inputs for the robotics sculpting task. In diffusion policy [1], the robot policy is represented as a denoising diffusion probabilistic model (DDPM). DDPMs [12] are generative models that iteratively denoise an input sampled from Gaussian noise. One of the key innovations of diffusion policy was incorporating visual observation conditioning in which the DDPM approximates the conditional distribution of $p(\mathbf{A_t} \mid \mathbf{O_t})$, where $\mathbf{A_t}$ is the predicted action sequence and $\mathbf{O_t}$ is the observation the action sequence is conditioned on. In the original diffusion policy framework, the observation conditioning $\mathbf{O_t}$ was a vector stacking the flattened latent embeddings of the past N image observations of the scene as well as the robot joint state. In this work, our observation conditioning instead is a learned latent embedding of the clay state and goal point clouds as well as the previous deformation action applied to the clay. To provide a quality latent embedding representing the 3D geometrical information of point clouds, we use PointBERT [2], a point transformer encoder, pre-trained on the ShapeNet dataset [13]. PointBERT is then finetuned end-to-end with the diffusion policy training. PointBERT takes a point cloud, and clusters it into 64 sub-clouds to learn both the overall global geometry as well as the more regional features. The output of PointBERT is $H = \{h_s, h_1, ..., h_g\}$ where h_s represents the global feature and $h_1, ..., h_g$ represents the regional features. While all of this information is very relevant, for the downstream task of learning a policy, we need a much more compact latent representation of the point cloud geometry. Thus, to combine PointBERT with the downstream policy, we add on a two-layer MLP projection head to reduce the latent representation to a compact size of 512. In particular, we combine together the entire global feature h_s and the maxpool of the regional features $h_1, ..., h_g$ before passing this combination through the MLP projection head. We use PointBERT with the MLP projection head to encode the current point cloud observation of the clay as well as the goal point cloud separately into feature vectors of shape 512. Finally, the observation vector $\mathbf{O_t}$ that conditions the diffusion process is the stacked latent representation of the clay state and goal as well as the previous sculpting action applied to the clay.

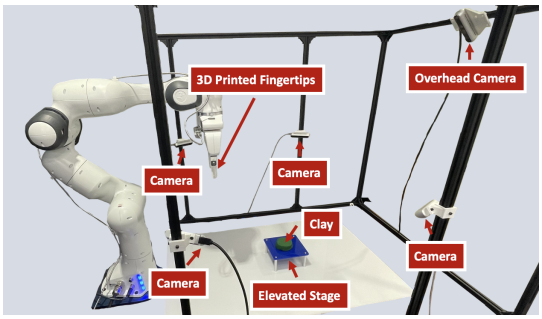

Fig. 2: The experimental setup includes 4 Intel RealSense D415 RGB-D cameras mounted to a camera cage to reconstruct the clay point cloud. An additional camera is used to record videos. We fit the robot with 3D printed fingertips and an elevated stage. We assume the clay remains centered and fixed to the elevated stage throughout the experiments.

III. EXPERIMENTS AND RESULTS

The human demonstration dataset is collected using kinesthetic teaching in which the expert physically moves the robot to sculpt the clay. We collect 10 demonstration trajectories for each target sculpting shape ('X', 'Line' and 'Cone'). We chose these sculpting targets because we believe they allow us to explore a variety of different sculpting behaviors while limiting the amount of time consuming hardware experiments. We apply the rotation transformation as described in section II-A, to our train dataset with an 80/20 split of the raw demonstrations to expand our training dataset to 2880 demonstrations per target shape. To further explore the use of point cloud inputs for imitation learning tasks, we combine the same proposed point cloud embedding strategy with two other state-of-the-art imitation learning frameworks, ACT [14] and VINN [15]. In addition to these imitation learning baselines, we compare the sculpting performance to a human using their own hand, a human operating the robot, and a simple heuristic method where the gripper squeezes the region of the clay with the greatest difference between

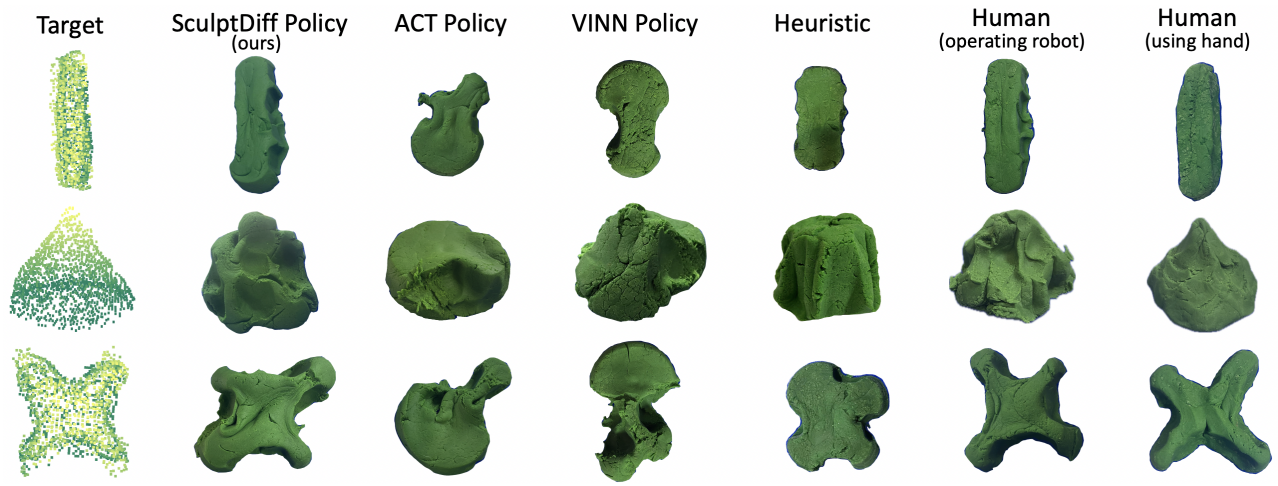

Fig. 3: The final shapes created by the policies trained with point cloud inputs. For the target point cloud on the left-most column, the lightness of each point is correlated with the point's z-value to visualize depth. While both human oracles create the best shapes, point cloud diffusion policy is able to successfully create the closest matches to the human demonstrations.

TABLE I: Performance of SculptDiff compared to baselines.

	Model	CD ↓	EMD ↓	# Grasps
	Diff.	**0.0073 ± 0.001**	**0.0071 ± 0.001**	6.1 ± 0.9
	ACT	0.0077 ± 0.001	0.0077 ± 0.001	8.0 ± 0.0
	VINN	0.0156 ± 0.002	0.0137 ± 0.001	9.0 ± 0.0
X	Heuristic	0.0087 ± 0.000	0.0079 ± 0.000	10.0 ± 0.0
	R. Demo	0.0075 ± 0.000	0.0095 ± 0.001	**8.0 ± 0.0**
	H. Demo	**0.0059 ± 0.001**	**0.0053 ± 0.001**	13.0 ± 0.9
	Diff.	**0.0045 ± 0.000**	**0.0041 ± 0.000**	8.0 ± 0.0
	ACT	0.0159 ± 0.003	0.0169 ± 0.003	8.0 ± 0.0
	VINN	0.0109 ± 0.001	0.0104 ± 0.001	8.0 ± 0.0
Line	Heuristic	0.0071 ± 0.001	0.0079 ± 0.001	**2.8 ± 0.4**
	R. Demo	**0.0064 ± 0.004**	0.0074 ± 0.004	**5.0 ± 0.0**
	H. Demo	0.0065 ± 0.001	**0.0065 ± 0.001**	9.8 ± 0.7
	Diff.	**0.0060 ± 0.001**	**0.0054 ± 0.001**	12.0 ± 0.0
	ACT	0.0070 ± 0.000	0.0079 ± 0.000	12.0 ± 0.0
	VINN	0.0096 ± 0.001	0.0092 ± 0.001	13.0 ± 0.0
Cone	Heuristic	0.0074 ± 0.000	0.0067 ± 0.000	**8.8 ± 0.8**
	R. Demo	0.0057 ± 0.001	0.0074 ± 0.002	**10.2 ± 0.7**
	H. Demo	**0.0038 ± 0.001**	**0.0032 ± 0.001**	16.7 ± 8.5

the current state and goal point clouds.

The numerical results of the sculpting tasks are shown in Table I in which we report the Chamfer Distance (CD) and Earth Mover's Distance (EMD) between the final clay point cloud and the target shape point cloud. We ran each policy 5 times for each shape target and report the mean and standard deviation across experiments. In addition to quantitative similarity, we show the final shapes created by our system compared to a variety of baselines in Figure 3. The SculptDiff policy outperforms the heuristic and learning baselines in terms of CD and EMD as well as visual shape quality for all shape goals. Both ACT and VINN struggled with the sculpting task for all target shapes. This is likely because both frameworks can struggle with multimodality in demonstrations, and the sculpting task is highly multimodal. There are multiple action sequences that can create a 3D shape, and our demonstrations reflect this. ACT and VINN, both deterministic policies, often get stuck in modes repeating similar grasp actions in perpetuity. This due to the rigidity of these policies, where if the first action is not ideal, and the subsequent states deviate more from the training demonstrations and the models fail to handle the compounding errors, a common issue with imitation learning. For both ACT and VINN, the algorithmic techniques to handle these compounding errors involves averaging, with temporal aggregation for ACT and kernel-averaging the actions of clustered states for VINN. However, this averaging scheme is not compatible with our action parameterization, as averaging two different grasps may result in a third with a substantially different deformation behavior than the original two, thus deviating further from the training states. To ensure we were evaluating these algorithms in the best light for this application, for our experiments we did not use temporal aggregation for ACT and limited the number of nearest neighbors for VINN. While this improved overall performance by reducing the errors caused by the averaging mechanisms, the challenges of compounding errors, particularly with a deterministic policy remained for both. In contrast to ACT and VINN, diffusion policy is stochastic and able to successfully capture the distribution of grasps along the clay over time. Based on the results of our experiments we attribute the success of SculptDiff to both the stochastic representation of actions as well as the point cloud state and goal representations.

IV. CONCLUSION

In this work, we present SculptDiff, the first imitation learning policy to successfully create a set of 3D clay sculptures entirely in the real world. Through our experiments, we demonstrate the value of leveraging 3D state representations, in this case point clouds, as well as the importance of a stochastic policy for the complex multi-modal task of sculpting. Our evaluation of sculpting quality compared to human baselines has demonstrated a clear need for further exploration of improving hardware to allow for finer changes to be applied to the clay as well as the development of semantic-based 3D shape similarity metrics.

REFERENCES

[1] 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.

[2] 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 *Conference on Computer Vision and Pattern Recognition*, 2022, pp. 19 313–19 322.

[3] 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.

[4] N. Lv, J. Liu, and Y. Jia, "Dynamic modeling and control of deformable linear objects for single-arm and dual-arm robot manipulations," *Transactions on Robotics*, vol. 38, no. 4, pp. 2341–2353, 2022.

[5] 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 *International Conference on Robotics and Automation*. IEEE, 2021, pp. 12 328–12 334.

[6] A. Bartsch, C. Avra, and A. B. Farimani, "SculptBot: Pre-Trained Models for 3D Deformable Object Manipulation," *arXiv preprint arXiv:2309.08728*, 2023.

[7] 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*, p. 02783649231219020, 2023.

[8] 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.

[9] J. Liu, Z. Li, W. Lin, S. Calinon, K. C. Tan, and F. Chen, "Softgpt: Learn goal-oriented soft object manipulation skills by generative pre-trained heterogeneous graph transformer," in *International Conference on Intelligent Robots and Systems*. IEEE, 2023, pp. 4920–4925.

[10] C. Qi, X. Lin, and D. Held, "Learning closed-loop dough manipulation using a differentiable reset module," *Robotics and Automation Letters*, vol. 7, no. 4, pp. 9857–9864, 2022.

[11] H. Zhu, Y. Wang, D. Huang, W. Ye, W. Ouyang, and T. He, "Point Cloud Matters: Rethinking the Impact of Different Observation Spaces on Robot Learning," *arXiv preprint arXiv:2402.02500*, 2024.

[12] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[13] 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.

[14] T. Z. Zhao, V. Kumar, S. Levine, and C. Finn, "Learning fine-grained bimanual manipulation with low-cost hardware," *arXiv preprint arXiv:2304.13705*, 2023.

[15] J. Pari, N. M. Shafiullah, S. P. Arunachalam, and L. Pinto, "The surprising effectiveness of representation learning for visual imitation," *arXiv preprint arXiv:2112.01511*, 2021.