Self-Supervised 3D Representation Learning for Robotics

Ishika Singh[1] Anthony Liang[1] Mohit Shridhar[2] Jesse Thomason[1]

***Abstract*— Recent work on visual representation learning from images and videos has shown to be efficient for robotic manipulation tasks. However, learning to act in a 6-DoF 3D action space from 2D observations is a hard problem. As a result, 2D representation learning methods require huge amounts of data for pretraining. To this end, we investigate a self-supervised 3D representation learning framework that works with limited data. Our model learns 3D scene representations from self-supervised masked reconstruction of 3D voxel grids, alongside imitation learning, from few-shot task demonstrations. We use Perceiver-Actor as the backbone for 3D representation learning. Our preliminary experiments show improved task success rates on the training task and its visual variations compared to the base Perceiver-Actor.**

I. Introduction

Working towards the goal of generalizable robotics is challenged by learning a general perception and interaction framework that works across robot agents, tasks, and environments. Recent works on visual representation learning on large-scale task-agnostic datasets has shown to be effective on downstream policy learning for a variety of visuomotor control tasks [1], [2], [3], [4], [5]. Most representation learning approaches for robotics employ a self-supervised masked reconstruction objective for pretraining. They then use the learned visual encoder to encode observations for downstream policy learning. All of these methods pretrain on large scale offline datasets of 2D observations such as images and videos with an objective to capture generalizable visual features.

While we have seen much success on learning representations for vision-and-language tasks, there is an open question of whether these representations are operating in the "right" space for robotics. Given the objective to predict actions in a 3D space, we argue that *the learned representation space for robotics should encode 3D features*.

In this paper, we present our preliminary study of a framework we propose for learning 3D representations for robotic manipulation tasks. We work with a problem formulation that aligns the observation and action spaces. This formulation enables both representation learning and action prediction to be in the same output space. To study self-supervised 3D representation learning, we utilize Perceiver-Actor [6], that trains a transformer-based behavior cloning policy conditioned on a 3D observation space, as our base model. Concretely, we optimize for a masked 3D scene reconstruction objective jointly with the behavior cloning objective. Our initial experiments show that these learned 3D representations aid the task performance on the training task, as well as are robust to visual perturbations.

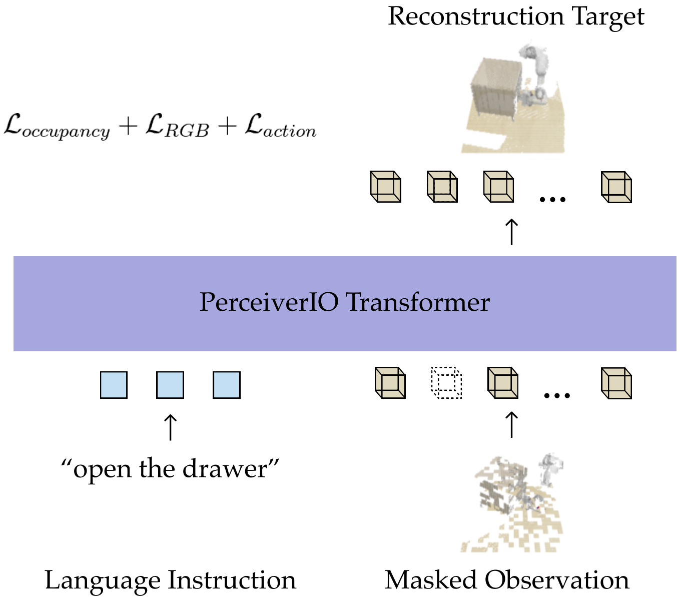


Fig. 1: Self-supervised 3D representation learning: We propose a 3D representation learning framework that jointly optimizes the 3D masked reconstruction and the behavior cloning objectives.

II. Related Work

A. 2D Representation learning for Robotics

An emerging area of work in robot learning has been studying representation learning for robotics, that are learned from egocentric images or videos of humans doing everyday activities. Existing works train visual encoders on out of domain task-agnostic images to obtain rich visual embeddings, which improves the downstream agent's scene understanding and helps with achieving higher task success [2], [3]. R3M [1] learns representation using human videos with captions using video-language contrastive and temporal loss objectives. Others learn representation on single or multi-view images on task-agnostic data using a self-supervised masked reconstruction objective [4]. Finally, Voltron [5] proposes a "language-aware" multimodal representation learning modal. It jointly models a sequence of frames and generates visually grounded language to describe the visual input. These works advocate the general efficacy of learning self-supervised representations, however they work with 2D inputs both during representation learning and policy adaption. We propose a

[1] University of Southern California `ishikasi@usc.edu`
[2] University of Washington

framework to jointly learn 3D scene representations and the policy using only a limited number of task demonstrations.

B. 3D Representation Learning for Robotics

3D4RL[7] proposes a 3D visual representation learning approach consisting of a pretraining phase using external data and a joint training phase using in-domain data collected by an RL agent. They pretrain a 3D voxel-based autoencoder with an offline 3D object dataset and use it as initialization for downstream task learning. Their proposed 3D objective requires at least two views of a scene, while our objective can work with the RGB-D from a single view. NeRF-RL [8] explores using Neural Radiance Fields (NeRFs) as a way to incorporate 3D inductive biases into learned state representations. They demonstrate that NeRF representations improve sample efficiency of learning manipulation tasks that depend on the geometry of the objects. They use their pretrained NeRF model as a frozen feature extractor, while we jointly optimize both the policy and learned representations. Additionally, training a NeRF model requires observations from multiple views and can be extremely computationally expensive. More broadly, researchers in the computer vision community have investigated learning multimodal 3D representations for 3D vision tasks. ULIP [9] introduces a unified representation of image, text, and 3D point clouds by pretraining with triplets from all three modalities. While such large-scale pretrained 3D representations are generally useful for vision tasks, they have not yet been used for robotic applications.

III. Method

We propose a method that jointly learns generalizable 3D representations and a policy from an offline dataset of expert demonstrations.

A. Dataset

The demonstration dataset consists of N trajectories, $\tau_i = \{(o_j, a_j)\}_{j=1}^{T}$ where o is the observation and a is a continuous robot action. An action a_j is 6-DoF pose and gripper open position, and an observation o_j is an RGB-D image from a given number of cameras. Each trajectory is paired with a natural language instruction l.

B. Problem Formulation

Following [10], [11], we utilize voxel grid as the observation space and discretize action prediction via voxel keypoint-based action prediction. The voxel grid is constructed from RGB-D images fused though triangulation using known camera extrinsics and intrinsics. We use a voxel grid of size 100^3, corresponding to an actual volume of 1.0m^3. The actions are then predicted as the "next best voxel" that is closest to the center of the gripper fingers. Moreover, instead of predicting continuous 7-DoF actions, we use keypoint action prediction. The keypoint actions are discovered using intuitive heuristics, such as instances where the arm's joint velocities are close to zero, and whether the gripper open state has changed.

C. Agent: Perceiver-Actor

Perceiver-Actor or PERACT [11] is a transformer-based robot learning framework that takes tokenized voxel grid and language instruction as the input, to predict discretized voxel grid translation point, and discrete rotation in Euler angles. PERACT works with 3D voxel grid token, akin to visual patch tokens or language tokens in vision or language transformers. The patch tokens of size 5^3 are encoded via a 3D convolution layer with kernel-size and stride of 5, resulting in 20^3=8000 voxel observation tokens. PERACT uses PerceiverIO Transformer [6], a latent-space Transformer architecture designed to handle large multimodal input sequences.

D. Self-supervised 3D Representation Learning

We use the PERACT framework to learn 3D representations for transferable and generalizable pretraining for robot learning. Inspired by the techniques in 2D representation learning, we propose 3D masked voxel reconstruction as a self-supervised training objective for learning 3D scene representations. In addition to action prediction, we incorporate two reconstruction terms in the loss to encourage learning 3D information. Specifically, our model learns to reconstruct occupancy and RGB channels for the input voxel grid using binary cross entropy and L2 loss respectively, given as

$$L_{Occupancy} = \sum_{c \in \text{voxel grid}} -\, w\ O(c)\ log(o') \\ -\,(1-w)(1-O(c))log(1-o') \quad (1)$$

$$L_{RGB} = \sum_{c \in \text{voxel grid}} \mathbb{1}_O(c)[-w(rgb_c - rgb'_c)^2] \quad (2)$$

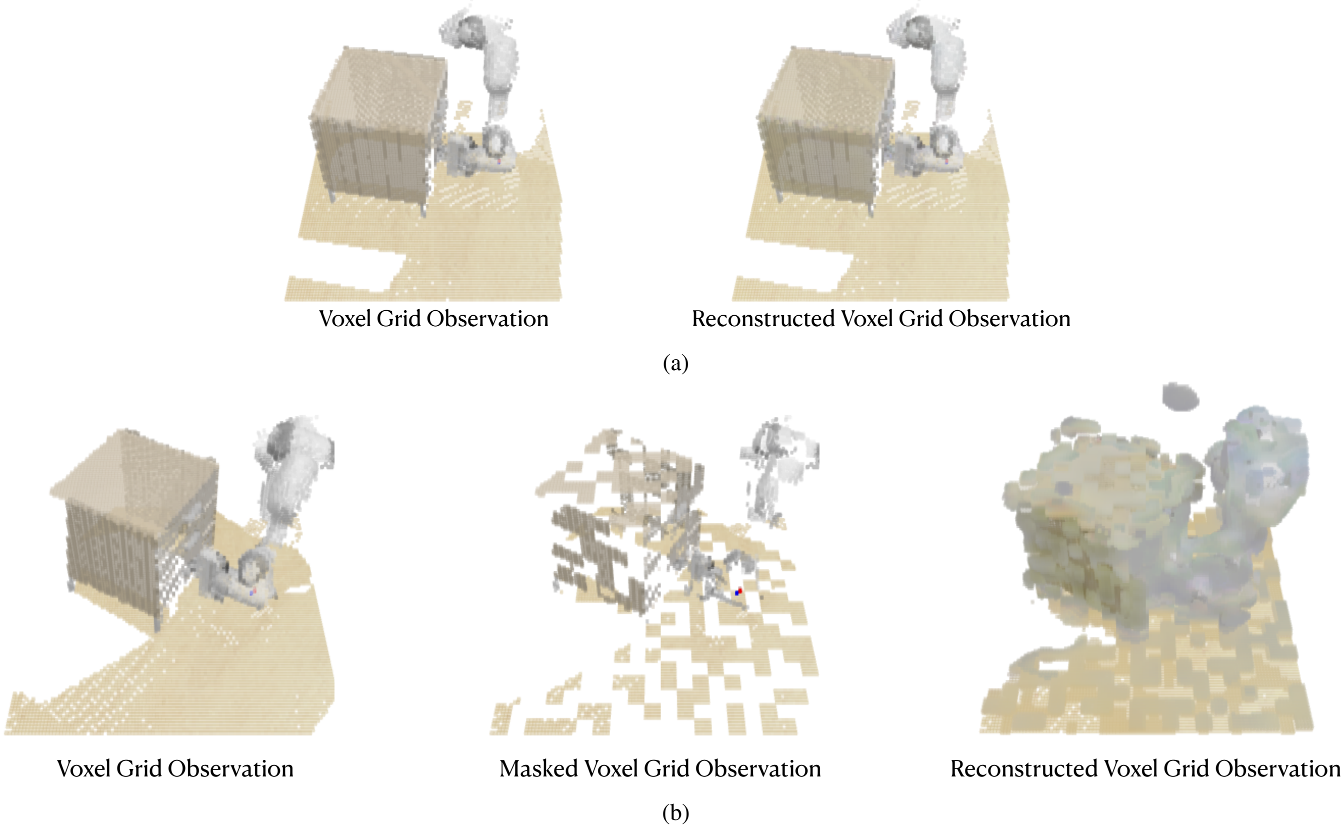
where $O(\cdot) \in \{0,1\}$ indicates whether a voxel c is occupied, $o' \in (0,1)$ is the predicted occupancy, and w is a weight term to correct for the class imbalance between occupied and unoccupied voxels. We apply a Sigmoid layer to occupancy logits for each voxel to map them between 0 and 1. For visualization, we use a threshold of 0.5 for determining occupied voxels. Similarly, $rgb_c \in \mathbf{R}^3 \in (-1,1)$ is the ground truth RGB value of voxel c and rgb' is the predicted RGB value.

We also apply our scalar factor w for RGB prediction, and we only consider RGB prediction for voxels that are occupied. The reconstruction objective is optimized together with the original action prediction objectives in [11].

We apply patch masking to the input voxel grid. We first tokenize the voxel grid into 5^3 patches. We randomly mask out 80% of the voxel patches, inspired by prior works [3] that show heavy masking helps learn more robust and useful representations. We only apply masking to 50% of the observations during training, so that during zero-shot evaluation, the unmasked input is not out of distribution.

IV. Experiments

PERACT [11] uses 2048 latents of dimension 512. However, for our initial experiments, we use 64 latents of


Voxel Grid Observation

Reconstructed Voxel Grid Observation

(a)

Voxel Grid Observation

Masked Voxel Grid Observation

Reconstructed Voxel Grid Observation

(b)

Fig. 2: Voxel grid reconstruction (a) without masking and (b) with masking

dimension 64: $\mathbb{R}^{64\times 64}$. We train the model for 70k iterations on the `Open Drawer` task in the RLBench simulation environment with 10 demonstrations collected using a scripted policy. We use the same procedure to generate a test set of 10 demonstrations, where each demonstration has a randomly sampled initial object pose. In our case, the drawer's pose changes, forcing the model be invariant to the object's starting location. We select the model with best training performance based on success rate and evaluate this on the heldout set of test configurations. Each test episode is given a score of 100 if the model completes the task fully, else 0. We also test our model on task variations (shown in Figure 3) with varied drawer color, size, and handle shape, to measure the robustness of our learned representations.

V. Results

In Figure 2, we visualize the reconstruction of the voxel grid with and without masking. Without masking, we get close to ground truth reconstruction. With patch masking, we observe a more noisy reconstruction of the scene. We hypothesize that the poor quality of masked reconstruction can be attributed to model capacity. We expect that scaling the size of our backbone model will enable it to better capture contextual features to accurately reconstruct the scene from few masked voxels, as in the case of 2D reconstruction.

In Table I, we report results on evaluating of our model trained with full reconstruction and masked reconstruction. We observe that our self-supervised reconstruction objective improves task learning and enables the learned agent to be robust to a variety of visual perturbations. We hypothesize that 3D reconstructions force the model to focus on the 3D structure of the object rather than visual attributes, which are not relevant for completing the task. Though we see robustness against visual shifts, the performance on the "small body" variation remains 0. This could be an inherent issue with imitation learning, in which the learned policy struggles to generalize to drastically new out-of-distribution action trajectories.

Our reconstruction objective forces the agent to construct the scene from its own scene encoding, which leads to better understanding of object features such as the drawer's shape and pose. In turn, these richer representations aid the action prediction, which in our case also happens in the same voxel grid space. Masking on top of reconstruction forces the model to infer the 3D shape of the object from the partially visible input. Moreover, the noisily masked inputs also implicitly augment the training samples.

VI. Conclusion

We propose a simple 3D representation learning framework. Our preliminary results show that it achieves strong downstream task performance and generalizes to different task variations.

In our ongoing work, we plan to evaluate our framework on more downstream task variations to further test the robustness of the learned representations. We will conduct

TABLE I: Task success rates (%) for `Open Drawer` task. Our proposed 3D representation learning approach improves task performance across most of the visual task purturbations.

Models	`Open Drawer`	Random Color Frames	Random Color Full Body	Random Texture Frame	Small Body	Square Handles	**Average Success Rate**
PERACT	60	10	10	40	0	60	30.0
w/ Full Reconstruction	70	**60**	**20**	80	0	**90**	**53.3**
w/ Masked Reconstruction	**90**	10	**20**	**90**	0	80	48.3

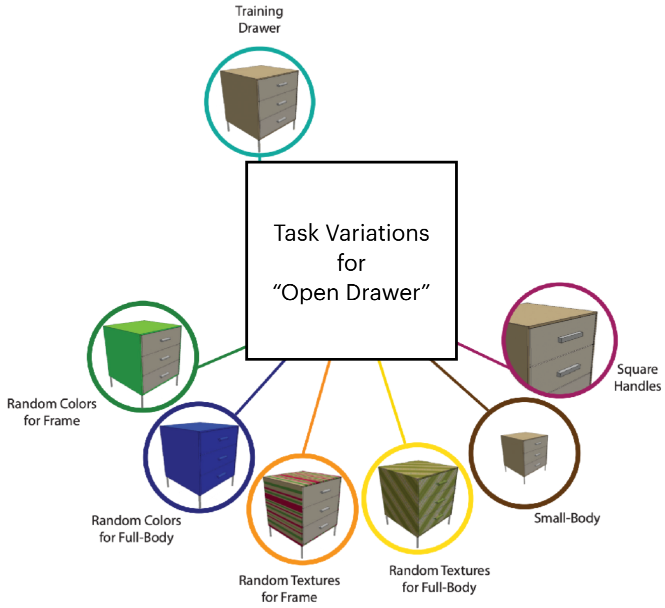

Training Drawer
Task Variations for "Open Drawer"
Square Handles
Random Colors for Frame
Small-Body
Random Colors for Full-Body
Random Textures for Frame
Random Textures for Full-Body

Fig. 3: Task variations for the `Open Drawer` task proposed in recent work [11].

ablation studies to better understand the effect of 3D masked reconstruction and what type of 3D inductive biases are being captured by our model. We also hope to study a multi-task pretrained model, which may learn more robust and useful 3D representation for downstream task adaptation. Moreover, we will conduct sim-to-real experiments in which we use our pretrained 3D representations in simulation as initialization for zero-shot or few-shot transfer to real-world manipulation tasks.

References

[1] S. Nair, A. Rajeswaran, V. Kumar, C. Finn, and A. Gupta, "R3m: A universal visual representation for robot manipulation," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=tGbpgz6yOrI

[2] S. Parisi, A. Rajeswaran, S. Purushwalkam, and A. Gupta, "The unsurprising effectiveness of pre-trained vision models for control," in *International Conference on Machine Learning*. PMLR, 2022, pp. 17 359–17 371.

[3] I. Radosavovic, T. Xiao, S. James, P. Abbeel, J. Malik, and T. Darrell, "Real-world robot learning with masked visual pre-training," in *6th Annual Conference on Robot Learning*, 2022. [Online]. Available: https://openreview.net/forum?id=KWCZfuqshd

[4] Y. Seo, J. Kim, S. James, K. Lee, J. Shin, and P. Abbeel, "Multi-view masked world models for visual robotic manipulation," *arXiv preprint arXiv:2302.02408*, 2023.

[5] S. Karamcheti, S. Nair, A. S. Chen, T. Kollar, C. Finn, D. Sadigh, and P. Liang, "Language-driven representation learning for robotics," *arXiv preprint arXiv:2302.12766*, 2023.

[6] A. Jaegle, S. Borgeaud, J.-B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, O. J. Henaff, M. Botvinick, A. Zisserman, O. Vinyals, and J. Carreira, "Perceiver IO: A general architecture for structured inputs & outputs," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=fILj7WpI-g

[7] Y. Ze, N. Hansen, Y. Chen, M. Jain, and X. Wang, "Visual reinforcement learning with self-supervised 3d representations," *IEEE Robotics and Automation Letters*, 2023.

[8] D. Driess, I. Schubert, P. Florence, Y. Li, and M. Toussaint, "Reinforcement learning with neural radiance fields," *arXiv preprint arXiv:2206.01634*, 2022.

[9] L. Xue, M. Gao, C. Xing, R. Martín-Martín, J. Wu, C. Xiong, R. Xu, J. C. Niebles, and S. Savarese, "Ulip: Learning unified representation of language, image and point cloud for 3d understanding," *arXiv preprint arXiv:2212.05171*, 2022.

[10] S. James and A. J. Davison, "Q-attention: Enabling efficient learning for vision-based robotic manipulation," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 1612–1619, 2022.

[11] M. Shridhar, L. Manuelli, and D. Fox, "Perceiver-actor: A multi-task transformer for robotic manipulation," in *Proceedings of the 6th Conference on Robot Learning (CoRL)*, 2022.