Self-supervised implicit shape reconstruction and pose estimation for predicting the future

Diego Patiño*, Karl Schmeckpeper*, Hita Gupta, Georgios Georgakis, Kostas Daniilidis
{diegopc,karls,hitagu,ggeorgak,kostas}@seas.upenn.edu

***Abstract*— We present our method for efficiently learning an implicit neural representation for shape reconstruction and pose estimation from raw sensor data. In contrast to recent methods, we utilize signed distance functions (SDFs) and learn this 3D representation in a self-supervised manner from depth observations. Furthermore, we argue that such a representation is suitable for predicting 3D motion that is informed by the shape representation.**

I. INTRODUCTION

A geometric understanding of its environment is critically important for a robotic agent to interact with the world. This understanding requires knowledge about shapes, objects, and how their properties evolve over time, implying that a representation capable of predicting the motion of objects in 3D can encode most necessary information for robotics. However, in realistic scenarios, the agent does not have access to a pre-existing 3D representation of the scene. Therefore, such representation must be learned from partial observations of the environment provided by available sensor measurements.

Recent work on shape reconstruction has attempted to learn object shape representations for the problem of shape completion [1], [2], [3], but has rarely considered how object shapes may influence 3D motion and, subsequently, the dynamics of the scene. On the other hand, prediction methods reason how the environment will evolve over time. Most prior work [4], [5], [6], [7] on prediction has focused on the problem separate from any understanding of shape and structure and focuses on predicting 2D motion by estimating the 2D flow of pixels or other transformations directly in the image space. However, building prediction models that are aware of the 3D structure of the world [8], [9], [10], [11] has several advantages. First, many behaviors that result in large discontinuities in two dimensions, such as occlusions and object permanence, become much simpler when viewed in three dimensions. Second, it is significantly easier to impose physically-grounded constraints on a three-dimensional space than on a two-dimensional space.

In this short paper, we present our method for self-supervised 3D shape and pose estimation from a depth sensor along with preliminary results. In contrast to other methods that use partial point clouds [8], [9] or voxels [10] to facilitate 3D prediction, we utilize signed distance functions (SDFs) [12], [13], [14], [15] as our shape representation. SDFs enable shape completion, allowing our method to reason about unseen portions of the scene, and have shown great performance in representing high-resolution shapes more efficiently than explicit representations, such as voxels. In addition, this representation is physically grounded, allowing us to easily incorporate structure to the problem, including enforcing rigid dynamics and penalizing interpenetration.

*Equal contribution

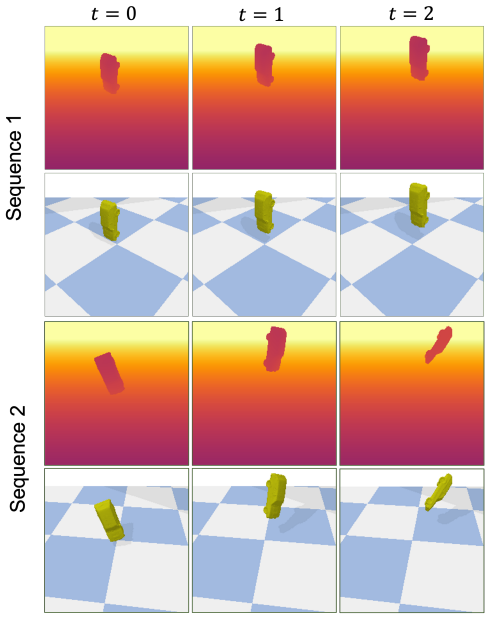

Fig. 1. Sample trajectories from the dataset.

II. SELF-SUPERVISED SHAPE AND POSE FOR PREDICTING THE FUTURE

We seek to perform object-centric prediction with rigid dynamics. In order to do this, we build a pipeline consiting of three components: first, an segmentation and encoding module that extracts object-centric information, second, a signed-distance function module that learns to represent the three-dimensional structure of each object, and third, a prediction module that predicts the poses of each object into the future.

a) Encoding: Our method assumes a static RGB-D camera pointing towards an object overcoming rigid dynamics. The given RGB-D images are masked to separate the moving object from the background. Later, we pass the masked RGB-D image through a convolutional neural network (CNN) to generate pixel-wise features. Because we focus our approach on the prediction task, we use the ground-truth segmentation masks on the input images.

The resulting masked features are passed through another CNN to generate two disentangled latent vectors (z, x) for shape and pose, respectively. The disentanglement of the two latent vectors is enforced alongside our rigid dynamics assumption by combining the shape and pose latent vectors from different timesteps, as described in the next section.

b) Learned Implicit Shape Representation: In contrast to related work on object centric prediction [16], [4], [5], [17], our method reconstructs objects and predict their dynamics as a full 3D geometry rather than individual future frames. To represent a 3D shape, we train a DeepSDF [12] decoder f_θ conditioned jointly on a shape latent z and on query points $p \in \mathbb{R}^3$. We compute a positional encoding [18], [19] $\phi(p)$ of the a query point before feeding it into f_θ.

Simultaneously, we estimate the object's pose through a ReLU-based MLP decoder, such that $(\mathbf{R}, t) = g_\theta(x)$, using the rotation parametrization from [20].

In order to account for the object's motion, we apply a transformation, $(\mathbf{R}, t)$ to each query point. We use the same transformation on all query points for a given object at a given t, enforcing the assumption that all objects are rigid.

$$p_t = \mathbf{R}p + t \tag{1}$$

We use the shape code from the first time step at all future time steps, such that the signed distance at a given point at a given time step is equal to

$$dist = f_\theta\left(z_0, \phi\left(\mathbf{R}_t p + t_t\right)\right), \tag{2}$$

where z_0 is the shape code from the object at time 0 and $(\mathbf{R}, t)$ is the pose of the object at time t.

Training a deep SDF decoder involves a significant amount of SDF ground-truth samples that are rarely available in real-world settings. Therefore, we train our shape decoder in a self-supervised way by approximating the SDF samples with a truncated signed distance function (TSDF) we generate using only the depth input.

Let $\mathcal{S}$ be a set of points on the surface of the ground truth object, and $\mathcal{R}$ be a set of points randomly sampled on the camera rays. We train our model using TSDF values at points $p \in \mathcal{R}$. We compute the TSDF values at training time as the Chamfer Distance d_{cd} between $\mathcal{S}$ and $\mathcal{R}$. We define three loss functions:

$$\mathcal{L}_{eikonal} = \sum_{p \in \mathcal{R}} |\, ||\nabla f_\theta(z_0, \phi(p))|| - 1| \tag{3}$$

$$\mathcal{L}_{surface} = \sum_{p \in \mathcal{S}} |f_\theta(z_0, \phi(\mathbf{R}p + t))| \tag{4}$$

$$\mathcal{L}_{tsdf} = \sum_{p \in \mathcal{R}} |f_\theta(z_0, \phi(\mathbf{R}p + t)) - d_{cd}(\mathbf{R}p + t, S)|. \tag{5}$$

By minimizing the Eikonal loss, $\mathcal{L}_{eikonal}$ we encourage f_θ to be close to a solution of the Eikonal PDF. The surface loss $\mathcal{L}_{surface}$ encourages the SDF decoder's values to vanish at surface points because their distance to the object is zero. Finally, the TSDF loss enforces appropriate distance values on points around the object. Note that our strategy is entirely self-supervised because it only depends on $\mathcal{S}$ and $\mathcal{R}$ that we compute from the input depth image.

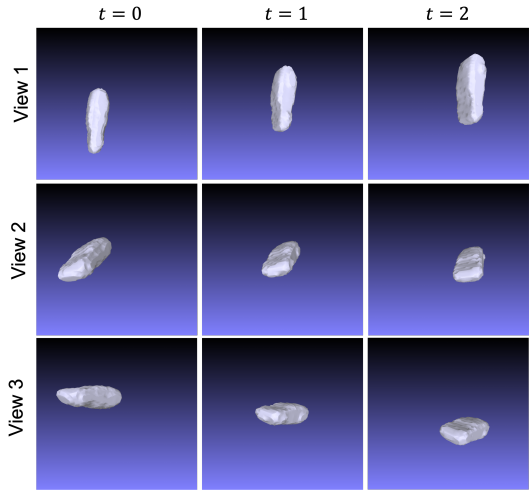

Fig. 2. Pose estimation and reconstruction of the same trajectory from unseen camera views.

c) Prediction: In order to train the prediction component of our model, we first freeze the parameter weights of the encoder and the SDF decoder. We then train an LSTM to predict the pose latent at the next time step, x_{t+1}, from the shape and pose latents at the current time step, (z_t, x_t).

III. Preliminary Results

a) Dataset: We evaluate our method on a simulated dataset of ShapeNet [21] objects launched in a parabolic trajectory at different linear and angular velocities. We restrict the position and initial velocity of the trajectories such that all objects land on a flat space $\mathcal{W} \in [-1, 1]^2$. Our dataset contains $24k$ sequences of RGB images, depth frames, and ground truth poses with the camera pointing towards the center of $\mathcal{W}$.

b) Pose Estimation and Reconstruction Results: We evaluate our method's capacity to jointly predict the shape and pose of objects moving due to parabolic shot dynamics. We summarize our results in Fig. 2. Our method can reconstruct shapes at different poses during the parabolic trajectory. Note that we can render the trajectories from unseen camera viewpoints because our model represents entire 3D geometries and not just the input viewpoint.

IV. Conclusions

We present a method for learning shape and pose representations from RGB-D images. We encode each object in a given RGB-D frame into a set of disentangled latent vectors, which allow us to construct a signed-distance function of the object and apply rigid transformations to it.

Future work includes refining the prediction pipeline and incorporating the model into robotic systems. Additionally, performing prediction in three dimensions allows for the inclusion of physically meaningful losses, such as penalizing the model for causing objects to interpenetrate, which we expect will allow for better performance in scenes with multiple objects.

REFERENCES

[1] Z. Landgraf, R. Scona, T. Laidlow, S. James, S. Leutenegger, and A. J. Davison, "Simstack: A generative shape and instance model for unordered object stacks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 012–13 022.

[2] E. Zobeidi and N. Atanasov, "A deep signed directional distance function for object shape representation," *arXiv preprint arXiv:2107.11024*, 2021.

[3] M. Firman, O. Mac Aodha, S. Julier, and G. J. Brostow, "Structured prediction of unobserved voxels from a single depth image," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5431–5440.

[4] Y. Ye, M. Singh, A. Gupta, and S. Tulsiani, "Compositional video prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 10 353–10 362.

[5] K. Schmeckpeper, G. Georgakis, and K. Daniilidis, "Object-centric video prediction without annotation," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 13 604–13 610.

[6] C. Finn and S. Levine, "Deep visual foresight for planning robot motion," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 2786–2793.

[7] A. X. Lee, R. Zhang, F. Ebert, P. Abbeel, C. Finn, and S. Levine, "Stochastic adversarial video prediction," *arXiv preprint arXiv:1804.01523*, 2018.

[8] A. Byravan and D. Fox, "Se3-nets: Learning rigid body motion using deep neural networks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017, pp. 173–180.

[9] A. Byravan, F. Leeb, F. Meier, and D. Fox, "Se3-pose-nets: Structured deep dynamics models for visuomotor control," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3339–3346.

[10] Z. Xu, Z. He, J. Wu, and S. Song, "Learning 3d dynamic scene representations for robot manipulation," in *Proceedings of the 2020 Conference on Robot Learning*, 2020.

[11] D. Driess, J.-S. Ha, M. Toussaint, and R. Tedrake, "Learning models as functionals of signed-distance fields for manipulation planning," in *Conference on Robot Learning*. PMLR, 2022, pp. 245–255.

[12] J. J. Park, P. R. Florence, J. Straub, R. A. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 165–174, 2019.

[13] S. Liu, Y. Zhang, S. Peng, B. Shi, M. Pollefeys, and Z. Cui, "Dist: Rendering deep implicit signed distance function with differentiable sphere tracing," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2016–2025, 2020.

[14] S. Liu, S. Saito, W. Chen, and H. Li, "Learning to infer implicit surfaces without 3d supervision," in *NeurIPS*, 2019.

[15] M. Michalkiewicz, J. K. Pontes, D. Jack, M. Baktash, and A. P. Eriksson, "Implicit surface representations as layers in neural networks," *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4742–4751, 2019.

[16] M. Janner, S. Levine, W. T. Freeman, J. B. Tenenbaum, C. Finn, and J. Wu, "Reasoning about physical interactions with object-oriented prediction and planning," *arXiv preprint arXiv:1812.10972*, 2018.

[17] T. Kipf, E. van der Pol, and M. Welling, "Contrastive learning of structured world models," in *International Conference on Learning Representations*, 2019.

[18] A. Kazemnejad, "Transformer architecture: The positional encoding," *kazemnejad.com*, 2019. [Online]. Available: https://kazemnejad.com/blog/transformer_architecture_positional_encoding/

[19] M. Tancik, P. P. Srinivasan, B. Mildenhall, S. Fridovich-Keil, N. Raghavan, U. Singhal, R. Ramamoorthi, J. T. Barron, and R. Ng, "Fourier features let networks learn high frequency functions in low dimensional domains," *ArXiv*, vol. abs/2006.10739, 2020.

[20] 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*, 2019, pp. 5745–5753.

[21] A. X. Chang, T. A. Funkhouser, L. J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "Shapenet: An information-rich 3d model repository," *ArXiv*, vol. abs/1512.03012, 2015.