DreamUp3D: Object-Centric Generative Models for Single-View 3D Scene Understanding and Real-to-Sim Transfer

Yizhe Wu[1], Haitz Sáez de Ocáriz Borde[1], Jack Collins[1], Oiwi Parker Jones[1], Ingmar Posner[1]

[1]Applied AI Lab, Oxford Robotics Institute.

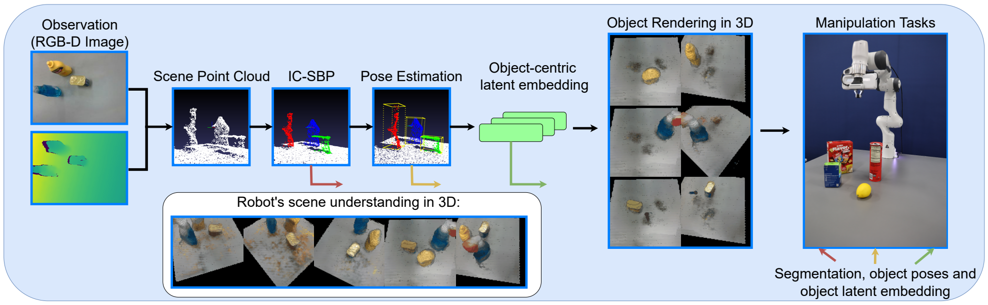
Observation (RGB-D Image)
Scene Point Cloud
IC-SBP
Pose Estimation
Object-centric latent embedding
Object Rendering in 3D
Manipulation Tasks
Robot's scene understanding in 3D:
Segmentation, object poses and object latent embedding

Figure 1. DreamUp3D is an Object Centric Generative Model that performs inference on RGB-D images for 3D scene understanding. The model predicts object segmentations, latent embeddings, 6D pose estimates and 3D reconstructions, all useful outputs for embodied tasks.

Extended Abstract

Reasoning about scenes at an object-centric level is an important abstraction for robots and embodied agents. This abstraction is both interpretable whilst also providing a structured representation useful for reasoning about the physical world. While robots operate in 3D environments and 3D sensors are ubiquitous, it is typical for object-centric representations to use only 2D data, e.g., images and video, to decompose the world. On top of the requirement for 3D scene understanding, object-centric models for embodied agents would also ideally operate in real-time on real-world data, provide informative latent representations, approximate the 6D pose of each object and output 3D reconstructions of individual objects and the full scene. DreamUp3D[1] [6] satisfies all of these requirements.

DreamUp3D is a novel Object-Centric Generative Model (OCGM) which is trained end-to-end in a self-supervised manner. The model is designed to operate over real RGB-D images, utilising both vision and depth to reason about 3D scenes. At inference, DreamUp3D runs in real time, providing object segmentations, 3D reconstructions, 6D pose estimates and latent embeddings. This functionality makes DreamUp3D ideal for real-world deployments requiring object-centric representations.

The architecture of DreamUp3D can be viewed as having 4 distinct modules trained end-to-end. First, a pre-processing module filters and embeds the colourised point cloud derived from the input RGB-D image using a U-Net-like architecture. Second, the embedding is segmented into individual object point clouds using an instance colouring stick-breaking process (IC-SBP) [1]. The third module predicts unseen missing points for each object point cloud and estimates a 6D pose. Shape completion is necessary as otherwise a poor 6D pose estimate would be calculated due to the partial observability of the scene. Finally, each 3D object is reconstructed using a generative radiance field (GRAF) [3] with the entire scene reconstructed using a composition of all the GRAFs. The loss function for training DreamUp3D is composed of five terms that simultaneously maximise the likelihood of the observations, enforce sparsity, improve the attention masks used for segmentation and distill information between the object GRAFs and the shape completion module. DreamUp3D is trained using a dataset of multi-view RGB-D images and associated camera extrinsics on tabletop scenes collected in the real-world.

We compare the capabilities of DreamUp3D to a range of competitive baselines. For 3D scene reconstruction, we compare DreamUp3D to depth-nerfacto [5], a commonly used scene reconstruction baseline based on Neural Radiance Fields, which is frequently employed in robotics applications. We also compare DreamUp3D to another 3D object-centric model, ObSuRF [4], and show DreamUp3D outperforms both baselines in reconstruction accuracy and also beats depth-nerfacto in reconstruction speed at test time. Additionally, when comparing matching accuracy between the learnt latent embedding of DreamUp3D to CLIP features [2] we find that the learnt latent embedding produces better features for matching between objects. Finally, comparing DreamUp3D to the only other 3D OCGM capable of estimating 6D poses, ObPose, we find that DreamUp3D provides better pose estimates.

In summary, DreamUp3D is a novel 3D OCGM capable of producing object segmentations, reconstructions, 6D pose estimates and latent embeddings with fast inference speed from a single RGB-D image. The structured outputs and learned representations from DreamUp3D can be leveraged for a wide variety of applications beyond those demonstrated including reinforcement learning and task and motion planning.

[1] Published in IEEE Robotics and Automation Letters

References

[1] Martin Engelcke, Oiwi Parker Jones, and Ingmar Posner. Genesis-v2: Inferring unordered object representations without iterative refinement. In *Neural Information Processing Systems*, 2021.

[2] Walter Goodwin, Sagar Vaze, Ioannis Havoutis, and Ingmar Posner. Semantically grounded object matching for robust robotic scene rearrangement. *2022 International Conference on Robotics and Automation (ICRA)*, pages 11138–11144, 2021.

[3] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. GRAF: Generative radiance fields for 3D-aware image synthesis. *Advances in Neural Information Processing Systems (NeurIPS)*, 33:20154–20166, 2020.

[4] Karl Stelzner, Kristian Kersting, and Adam R. Kosiorek. Decomposing 3d scenes into objects via unsupervised volume segmentation. *ArXiv*, abs/2104.01148, 2021.

[5] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023.

[6] Yizhe Wu, Haitz Sáez de Ocáriz Borde, Jack Collins, Oiwi Parker Jones, and Ingmar Posner. Dreamup3d: Object-centric generative models for single-view 3d scene understanding and real-to-sim transfer. *IEEE Robotics and Automation Letters*, 9(4):3291–3298, 2024.