Extended Abstract Track

# Symmetry-based Object-centric Learning for Rigid Objects

**Zhiwei Han** HAN@FORTISS.ORG
**Stefan Matthes** MATTHES@FORTISS.ORG
**Hao Shen** SHEN@FORTISS.ORG
**Yuanting Liu** LIU@FORTISS.ORG
*Guerickestraße 25, 80805 München, Germany*

## Abstract

In this work, we present SymObjectRF, a symmetry-based method that learns object-centric representations for rigid objects from one dynamic scene without hand-crafted annotations. SymObjectRF learns the appearance and surface geometry of all dynamic object in their canonical poses and represents individual object within its canonical pose using a canonical object field (COF). SymObjectRF imposes group equivariance on rendering pipeline by transforming 3D point samples from world coordinate to object canonical poses. Subsequently, a permutation-invariant compositional renderer combines the color and density values queried from the learned COFs and reconstructs the input scene via volume rendering. SymObjectRF is then optimized by minimizing scene reconstruction loss. We show the feasibility of SymObjectRF in learning object-centric representations both theoretically and empirically.

**Keywords:** 3D Computer Vision; Group Symmetry; Object-centric Learning

## 1. Introduction

Learning 3D object-centric representation from sensory inputs is crucial for scene understanding and downstream tasks across various domains. However, visual ambiguities caused by occlusion and entangled information make object-centric learning an under-constrained inverse problem. Previous unsupervised learning methods failed to address this challenging problem unless additional structure in dataset or model architecture is available (Hyvärinen and Pajunen, 1999; Locatello et al., 2019). Although recent researches demonstrate significant progress (Yang et al., 2021; Wu et al., 2022a; Weder et al., 2023; Gao et al., 2023; Yu et al., 2022; Niemeyer and Geiger, 2021; Wu et al., 2022b; Stelzner et al., 2021) in leveraging 3D cues for object-centric learning, existing 3D methods still have certain practical limitations such as the high demand for training data (Yu et al., 2022), the requirement for holistic scene geometry (Gao et al., 2023) or the reliance on hand-annotated supervision (Wu et al., 2022a; Yang et al., 2021). A natural question to ask is: Can we learn 3D object-centric representations from a specific dynamic scene without relying on pixel-level supervision, especially when the holistic scene geometry is not available?

In this work, we introduce group symmetry for learning 3D object-centric representations and impose Euclidean equivariance on the rendering process through scene reconstruction. The experimental results demonstrate that the learned object representations capture invariance in objects such as appearance and surface geometry.

## 2. Approach

### 2.1. Preliminaries

The goal of object-centric learning is to learn a dedicated representation for every object in the presented scenes. In the rest of the text, we use capital and lowercase letters to denote spaces and space elements and $[\cdot]_i$ to index the $i$-th independent subspace or the element of the $i$-th independent subspace. Let $W = W_1 \times W_2$ be the source space of two independent objects, $Z = Z_1 \times Z_2$ be the representation space of two objects. Let $\psi$ and $b$ denote a decoding function and a data generating process with variable input object number and $dist$ be a distance measure. Object-centric learning aims to find a representation space $Z$, such that $dist(\psi(z'), b(w'))$ is minimized for all $z' \in \mathcal{P}_z$ and all $w' \in \mathcal{P}_w$, where $\mathcal{P}_w$ and $\mathcal{P}_z$ denote the power sets of the subspace element sets of $w \in W$ and $z \in Z$. Without loss of generality, above setup is applicable to scenarios with an arbitrary number of objects.

### 2.2. Proposed Construction

A key challenges for learning object-centric representations is the object ordering ambiguity. Similar to Locatello et al. (2020); Kipf et al. (2022); Elsayed et al. (2022), we propose the construction of fixed object ordering in $W$ and $Z$. Let $I_Z$ and $I_W$ be the sets of the subspace indices of $Z$ and $W$. Let $P$ and $P'$ be permutation groups on the subspaces of $Z$ and $W$ with group actions $\beta_Z : P \times I_Z \to I_Z$ and $\beta_W : P' \times I_W \to I_W$ respectively. Let $Z_{/\sim_p} = Z_{\beta_Z(p,1)} \times Z_{\beta_Z(p,2)}, \forall p \in P$ and $W_{/\sim_{p'}} = Z_{\beta_W(p',1)} \times W_{\beta_Z(p',2)}, \forall p' \in P'$.

**Property 1 (Construction of fixed object ordering)** (a) $\forall p \in P, Z_{/\sim_p} \in Z/P$ and $\forall p' \in P', W_{/\sim_{p'}} \in W/P'$ (b) $\bigcup_{p \in P} Z_{/\sim_p} = Z$ and $\bigcup_{p' \in P'} W_{/\sim_p} = Z$.

In the rest of this work, we refer $p$ and $p'$ to a fixed subspace permutation of $Z$ and $W$ for the sake of simplicity. Instead of $Z$, we prospose to learn the representation $Z_{/\sim_p}$ of $W_{/\sim_{p'}}$.

To tackle with the lack of supervision elaborated in Section 1, we introduce group symmetry and equivariant decoding process as additional structure.

**Property 2 (Assumption of source space structure)** (a) There exists a group composition $G = G_1 \times G_2$ of two non-trivial subgroups with a group action $\alpha_W : G \times W \to W$ on $W$, (b) There exists a permutation group on the subspaces of $G$ with group action $\beta_G : R \times I_G \to I_G$, where $I_G$ denotes the set of the subspace indices of $G$, (c) There exists a permutation group isomorphism $m' : P' \to R$, (b) $\bigcup W_{/\sim_{p'}}/G_{/\sim_{m'(p')}} = W_{/\sim_{p'}}$, where $G_{/\sim_r} = G_{\beta_G(r,1)} \times G_{\beta_G(r,2)}, \forall r \in R$.

Similarly, we impose the group-induced structure in $Z_{/\sim_p}$ by adopting the formalism of symmetry-based disentangled representation (SBDR) from Higgins et al. (2018) and we propose the construction for object-centric represention used in this work. Let $\alpha_Z : G \times Z \to Z$ be the group action of $G$ on $Z$ and $g_i^{(e)}$ be the identity element of $G_i$.

**Property 3 (Construction of object-centric representations)** The representation $Z_{/\sim_p}$ is said to be **object-centric** if (a) There exists a permutation group isomorphism $m : P \to R$, (b) $Z_{/\sim_p}$ is a homogeneous space for $G_{/\sim_{m(p)}}$, (c) There exists a $G$-equivariant map $f : W_{/\sim_{p'}} \to Z_{/\sim_p}$ such that $f(\alpha_W(g, w)) = \alpha_Z(g, f(w)), \forall g \in G_{/\sim_{m(p)}}, w \in W_{/\sim_{p'}}$,

(d) $(Z_{/\sim_p})_i$ is invariant to the group action of $(G_{/\sim_{m(p)}})_j, \forall i \in \{1,2\}$ and $i \neq j$, i.e., $z_i = \alpha_Z(g,z)_i, \forall z \in Z_{/\sim_p}, i \in \{1,2\}, g \in \{g | g \in G_{/\sim_{m(p)}}$ and $g_i = g_i^{(e)}\}$.

A major difference between our method and prior methods is that we aim to learn the object set of a particular scene rather than certain object categories. Since the object set to be learned can be considered as a connected subspace of $W$, our objective is equivalent to learn a locally object-centric representations $Z_{local}$, a connected subspace of $Z_{/\sim_p}$.

**Property 4 (Construction of locally object-centric representations)** The representation $Z_{local}$ is said to be **locally object-centric** if (a) $Z_{local} \in Z_{/\sim_p}/G_{/\sim_{m(p)}}$, (b) There exists $W' \in W_{/\sim_{p'}}/G_{/\sim_{m(p)}}$ and a $G$-equivariant map $f_{local} : W' \to Z_{local}$ such that $\exists w' \in W', \forall g \in G_{/\sim_{m(p)}}, f_{local}(\alpha_W(g,w')) = \alpha_Z(g, f_{local}(w'))$, (c) $(Z_{local})_i$ is invariant to the group action of $(G_{/\sim_{m(p)}})_j, \forall i \in \{1,2\}$ and $i \neq j$, i.e., $z_i = \alpha_Z(g,z)_i, \forall z \in Z_{local}, i \in \{1,2\}, g \in \{g | g \in G_{/\sim_{m(p)}}$ and $g_i = g_i^{(e)}\}$.

Let $b$ be a permutation-invariant data generating process and $X' \subseteq X$ be the observation set, we propose our final construction for locally object-centric learning.

**Property 5 (Construction for locally object-centric learning)** (a) $X' \subseteq b(W'), \exists W' \in W_{/\sim_{p'}}/G_{/\sim_{m'(p')}}$, (b) We have the access to a mapping $\eta : X \to G_{/\sim_{m'(p')}}$, (c) We impose $G$-equivariance on the permutation-invariant decoding function $\psi : Z \to X$ and the learned locally object-centric representations $Z' \in Z_{/\sim_p}/G_{/\sim_{m(p)}}$, i.e., we want to learn a $z' \in Z'$, $\forall x' \in X', x' = \psi(\alpha_Z(\eta(x'), z'))$.

### 2.3. Problem Setup

In this work, we focus on learning a locally object-centric representation from one single dynamic scene. More specifically, we learn a **canonical** representation $z' \in Z'$ from a homogeneous space $Z' \in Z_{/\sim_p}/G_{/\sim_{m(p)}}$ that jointly reconstructs scene observations with the application of group symmetry via property 5 (c).

## 3. Algorithm and Experiments

### 3.1. Algorithm Overview

The main idea of SymObjectRF is to reconstruct every dynamic object in a scene video from its canonical representation. For every dynamic object in the scene, SymObjectRF maintains a dedicated canonical object field (COF), a neural radiance field that maps bounded coordinates to density and RGB values. The learned objects are represent in their canonical pose in COFs. During volume rendering, 3D point samples are transformed from world coordinate to the canonical poses of the learned objects by their canonical transformations, which are the inverse transformations of corresponding object placements. The input scenes are then reconstructed by combing the density and RGB values queried by the transformed points from all the COFs. The minimization of reconstruction losses, i.e., RGB- and depth loss, and the application of canonical transformations ensures the Euclidean equivariance of the rendering pipeline.

| Methods | MoviSim | | MoviCmplx | |
|---|---|---|---|---|
| | ARI↑ | FG-ARI↑ | ARI↑ | FG-ARI↑ |
| Slot AttentionLocatello et al. (2020) | 0.255 | 0.617 | 0.199 | 0.603 |
| SAVi Kipf et al. (2022) | 0.442 | 0.632 | 0.412 | **0.663** |
| SymObjectRF(Fixed viewpoint) | **0.732** | **0.727** | **0.773** | 0.656 |
| DYNAVOLGao et al. (2023) | 0.027 | 0.112 | 0.033 | 0.257 |
| uORFYu et al. (2022) | 0.831 | 0.866 | 0.707 | 0.739 |
| SAMHu et al. (2018) | 0.422 | 0.512 | 0.529 | 0.562 |
| SymObjectRF | **0.937** | **0.873** | **0.822** | **0.795** |

Table 1: ARI and FG-ARI scores for the quantitative evaluation of scene decomposition problem.

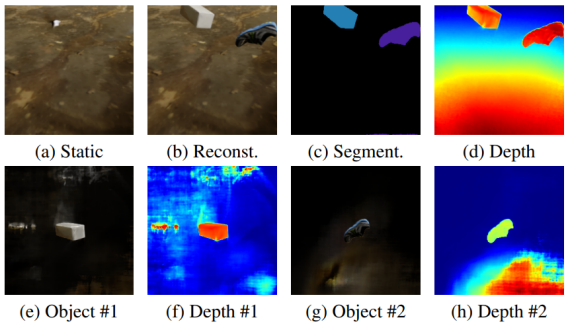


Figure 1: Rendering results from one held-out novel viewpoint

### 3.2. Experimental Setup

The training dynamic scenes were captured using Kubric (Greff et al., 2022) from a monocular camera with smooth linear movement, where captured scene geometry does not necessarily include the geometries of all objects. We assume the access to object poses for computing canonical transformations.

We evaluate SymObjectRF on the scenes generated using two distinct configurations. The first configuration *MoviSimple* consists of a CLEVER-styled background and objects from the kubasic object set, while a more photo-realistic configuration *MoviCmplx* shares the same background and object arrangement as the Movi-E Greff et al. (2022) setup.

To assess the performance of SymObjectRF on the scene decomposition task, we use the adjusted rand index (ARI) for the segmentation quality of the whole scene and foreground adjusted rand index (FG-ARI) for object segmentation. ARI measures pixel-wise clustering similarity between a predicted segmentation and the ground truth segmentation. The ARI score varies from 0 for a random clustering to 1 for a perfect segmentation.

### 3.3. Experimental Results

To evaluate the object decomposition capability of SymObjectRF for 3D dynamic scenes, we compare SymOjbectRF to 2D segmentation method Slot Attention (Locatello et al., 2020), its video version SAVi(Kipf et al., 2022), SAM(Kirillov et al., 2023), 3D object-centric method uORF (Yu et al., 2022) and unsupervised 3D object-centric method DYNAVOL(Gao et al., 2023). We compute the contribution of each canonical object field (including the background field) during rendering and assign a sampled ray the object class associated with the highest contribution for object segementation. We report the ARI and FG-ARI scores averaged across all 4 scenes from the 12 held-out camera positions for quantitative analysis in Table 1.

In the performance comparison presented in Table 1, our method outperforms nearly all 2D and 3D baseline methods. DYNAVOL potentially suffers from the absence of holistic scene geometry and perform less competitively, while SymObjectRF showcases its ability to learn object-centric representations even when the presented object geometry is incomplete.

We present both scene-level and object-level rendering results of SymObjectRF from one held-out novel viewpoint in Figure 1. Figure 1 (a) to (d) illustrate the scene-level reconstructed RGB, segementation and depth map. Figure 1 (e) to (f) plot the corresponding object-level rendering results. The plausible object-level rendering results and the perfect alignments of objects to the coordinate origin imply the effectiveness of SymObjectRF in learning object within a canonical pose.

## References

Gamaleldin F. Elsayed, Aravindh Mahendran, Sjoerd van Steenkiste, Klaus Greff, Michael C. Mozer, and Thomas Kipf. SAVi++: Towards end-to-end object-centric learning from real-world videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Siyu Gao, Yanpeng Zhao, Yunbo Wang, and Xiaokang Yang. Unsupervised object-centric voxelization for dynamic scene understanding. *CoRR*, 2023.

Klaus Greff, Francois Belletti, Lucas Beyer, Carl Doersch, Yilun Du, Daniel Duckworth, David J Fleet, Dan Gnanapragasam, Florian Golemo, Charles Herrmann, Thomas Kipf, Abhijit Kundu, Dmitry Lagun, Issam Laradji, Hsueh-Ti (Derek) Liu, Henning Meyer, Yishu Miao, Derek Nowrouzezahrai, Cengiz Oztireli, Etienne Pot, Noha Radwan, Daniel Rebain, Sara Sabour, Mehdi S. M. Sajjadi, Matan Sela, Vincent Sitzmann, Austin Stone, Deqing Sun, Suhani Vora, Ziyu Wang, Tianhao Wu, Kwang Moo Yi, Fangcheng Zhong, and Andrea Tagliasacchi. Kubric: a scalable dataset generator. 2022.

Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a definition of disentangled representations. arxiv. *arXiv preprint arXiv:1812.02230*, 2018.

Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4233–4241, 2018.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999.

Thomas Kipf, Gamaleldin F. Elsayed, Aravindh Mahendran, Austin Stone, Sara Sabour, Georg Heigold, Rico Jonschkowski, Alexey Dosovitskiy, and Klaus Greff. Conditional Object-Centric Learning from Video. In *International Conference on Learning Representations (ICLR)*, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. *Advances in Neural Information Processing Systems*, 33:11525–11538, 2020.

Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021.

Karl Stelzner, Kristian Kersting, and Adam R Kosiorek. Decomposing 3d scenes into objects via unsupervised volume segmentation. *arXiv preprint arXiv:2104.01148*, 2021.

Silvan Weder, Guillermo Garcia-Hernando, Aron Monszpart, Marc Pollefeys, Gabriel J Brostow, Michael Firman, and Sara Vicente. Removing objects from neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16528–16538, 2023.

Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In *European Conference on Computer Vision*, pages 197–213. Springer, 2022a.

Yizhe Wu, Oiwi Parker Jones, and Ingmar Posner. Obpose: Leveraging pose for object-centric scene inference and generation in 3d. 2022b.

Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *International Conference on Computer Vision (ICCV)*, October 2021.

Hong-Xing Yu, Leonidas J. Guibas, and Jiajun Wu. Unsupervised discovery of object radiance fields. In *International Conference on Learning Representations*, 2022.