Australian National University

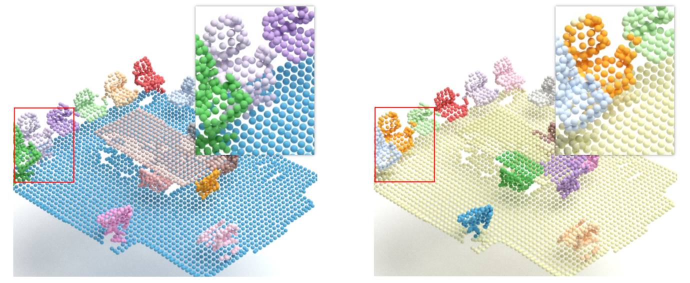
Spatially Invariant Unsupervised 3D Object-Centric Learning and Scene Decomposition

Tianyu Wang , Miaomiao Liu, Kee Siong Ng
Australian National University

ECCV TEL AVIV 2022

Goal and Contributions

Goal. Achieve unsupervised object-centric learning and 3D scene decomposition from 3D point clouds via a generative model.



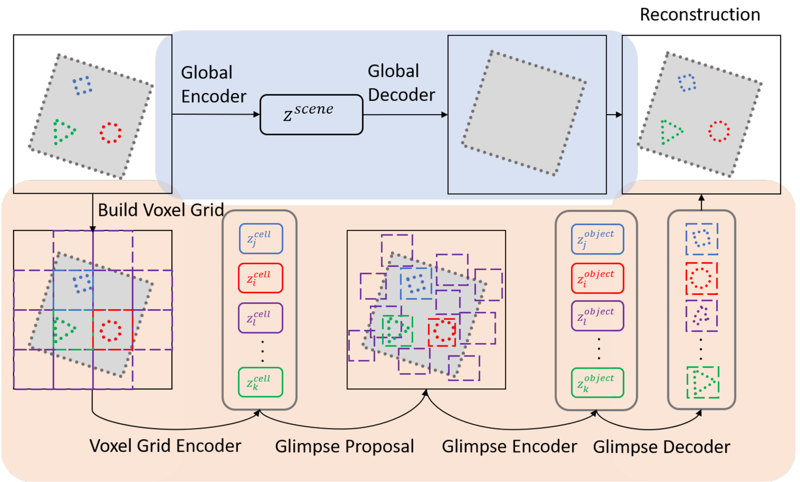
GT Ours

- We introduce a framework, **SPAIR3D**, to factorize a 3D point cloud into a spatial mixture model where each component corresponds to one object.
- We aim to maximize the likelihood for a point cloud $\mathcal{X}$ is $p(\mathcal{X}) = \int_{\mathbf{z}} p(\mathbf{z})p(\mathcal{X}|\mathbf{z})d\mathbf{z}$, where $\mathbf{z} = (\bigcup_i \mathbf{z}_i^{cell}) \cup (\bigcup_i \mathbf{z}_i^{object}) \cup \mathbf{z}^{scene}$, given the latent representations of objects and the scene.

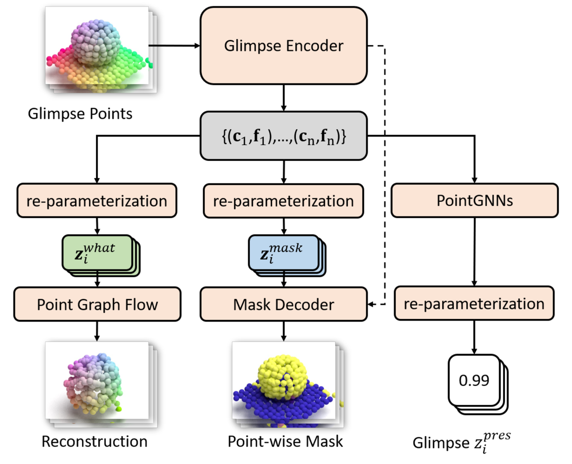
Contributions.

- To the best of our knowledge, the first unsupervised object-centric learning pipeline for point cloud data, named SPAIR3D.
- A new *Chamfer Mixture Loss* function tailored for learning mixture models over point cloud data with a novel graph neural network that can be used to model and generate a variable number of 3D points.
- Learn meaningful object-centric representation and decompose point clouds scene with an arbitrary number of objects in an object-oriented manner.

Overview



Structure of SPAIR3D



Structure of Glimpse VAE

- Latent variables : $\mathbf{z}_i^{cell} = \{\mathbf{z}_i^{where}, \mathbf{z}_i^{apothem}\}$ enocdes the position and dimension of the proposed object bounding box. $\mathbf{z}_i^{object} = \{\mathbf{z}_i^{what}, \mathbf{z}_i^{mask}, z_i^{pres}\}$ encodes the object structure, the mask for points and presence status of the object, respectively. $\mathbf{z}^{scene} = \{\mathbf{z}_0^{what}\}$ encode the scene structure information.
- Generated point cloud: $\hat{\mathcal{X}}$ and the input point cloud: $\mathcal{X}$
- Point cloud is firstly discretized into cells, each of which proposes an object glimpse.

Our Approach

Our solution.
We introduce GlimpseVAE and GlobalVAE for object-centric learning and scene decomposition.

- The Glimpse VAE is composed of a Glimpse Encoder, Point Graph Decoder, Mask Decoder and a multi-layer PointGNN network.
- The Global VAE consisting of the Global Encoder and a PGD outputs the reconstructed scene layout.

We design an encoder network $q_\phi(\mathbf{z}|x)$ to obtain $\{\mathbf{z}_i^{cell}\}_{i=1}^n$ and $\{\mathbf{z}_i^{object}\}_{i=1}^n$ from a point cloud $\mathcal{X}$.

1. Voxel Grid Encoding. Taking $\mathcal{X}$ as input, generating for each voxel cell $\mathcal{C}_i$ two latent variables $\mathbf{z}_i^{where} \in \mathbb{R}^3$ and $\mathbf{z}_i^{apothem} \in \mathbb{R}^3$ to propose a glimpse $\mathcal{G}_i$ potentially occupied by an object.
2. Glimpse Encoding. Encode each glimpse $\mathcal{G}_i$ into one point $\mathbf{a}_i = (\mathbf{c}_i, \mathbf{f}_i)$, defining the glimpse center coordinate and feature vector, then to generate $\mathbf{z}_i^{what}$ and $\mathbf{z}_i^{mask}$ from $\mathbf{a}_i$.
3. Global Encoding. Encode scene glimpse $\mathcal{G}_0$. to $\mathbf{z}_0^{what}$ with $z_0^{pres} = 1$.

We now introduce the decoders used for point-cloud and mask generation.

1. Point Graph Decoder. Decode $\mathbf{z}_i^{what}$ of each glimpse to point-cloud reconstruction.
2. Mask Decoder. The Mask Decoder decodes $(\mathbf{c}_i, \mathbf{z}_i^{mask})$ to the mask value, $\pi_i^x \in [0,1]$, of each point within a glimpse $\mathcal{G}_i$.

Our Loss. Denote the i^{th} glimpse as $\mathcal{G}_i$, $i \in \{0,\ldots,n\}$ and its reconstruction as $\hat{\mathcal{G}}_i$, $i \in \{0,\ldots,n\}$, the scene glimpse as the 0^{th} glimpse $\mathcal{G}_0 = \mathcal{X}$.

- $\mathcal{L} = -\log \mathcal{L}_{CD}(\mathcal{X}, \hat{\mathcal{X}}) + \mathcal{L}_{KL}(\mathbf{z}^{cell}, \mathbf{z}^{object}, \mathbf{z}^{scene})$, where $\mathcal{L}_{KL}$ is the KL divergence between the prior and posterior of the latent variables,
- We define *Chamfer Mixture Loss* as $\mathcal{L}_{CD}(\mathcal{X}, \hat{\mathcal{X}}) = \mathcal{L}^F(\mathcal{X}) \cdot \mathcal{L}^B(\hat{\mathcal{X}})$.
- The total forward likelihood of $\mathcal{X}$ is then defined as $\mathcal{L}^F(\mathcal{X}) = \prod_{x\in\mathcal{X}} \mathcal{L}^F(x)$, where the mixture model for an input point x is $\mathcal{L}^F(x) = \sum_{i=0}^n \alpha_i^x \mathcal{L}_i^F(x)$.
- For each input point x in the i^{th} glimpse, the glimpse-wise forward likelihood of that point is defined as $\mathcal{L}_i^F(x) = \frac{1}{u_i} \max_{\hat{x}\in\hat{\mathcal{G}}_i} \mathcal{N}(x|\hat{x}, \sigma_c)$, where $u_i = \int_{x\in\mathcal{X}} \max_{\hat{x}\in\hat{\mathcal{G}}_i} \mathcal{N}(x|\hat{x}, \sigma_c)dx$ is the normalizer.
- For each glimpse $\mathcal{G}_i$, $i \in \{0,\ldots,n\}$, $\alpha_i^x \in [0,1]$ defines a mixing weight for point x in the glimpse and $\sum_{i=0}^n \alpha_i^x = 1$ which further defines the segmentation mask.
- The backward regularization is then defined as $\mathcal{L}^B(\hat{\mathcal{X}}) = \prod_{i=0}^n \prod_{\hat{x}\in\hat{\mathcal{G}}_i} \mathcal{L}^B(\hat{x})^{\alpha_i^{x(\hat{x})}}$.
- For each predicted point $\hat{x}$, the point-wise backward regularization is $\mathcal{L}^B(\hat{x}) = \max_{x\in\mathcal{G}_{i(\hat{x})}} \mathcal{N}(\hat{x}|x, \sigma_c)$, where $i(\hat{x})$ returns the glimpse index of $\hat{x}$. We denote $x(\hat{x}) = \arg\max_{x\in\mathcal{G}_{i(\hat{x})}} \mathcal{N}(\hat{x}|x, \sigma_c)$ and $\hat{\mathcal{X}} = \bigcup_{i=0}^n \hat{\mathcal{G}}_i$.

Experiments

Metrics.

- We use the Adjust Rand Index (ARI) [2] to measure the segmentation performance against the ground truth instance labels.
- We also employ foreground Segmentation Covering (SC)[3] and foreground unweighted mean Segmentation Covering (mSC) for performance measurements as ARI does not penalize object over-segmentation[3].

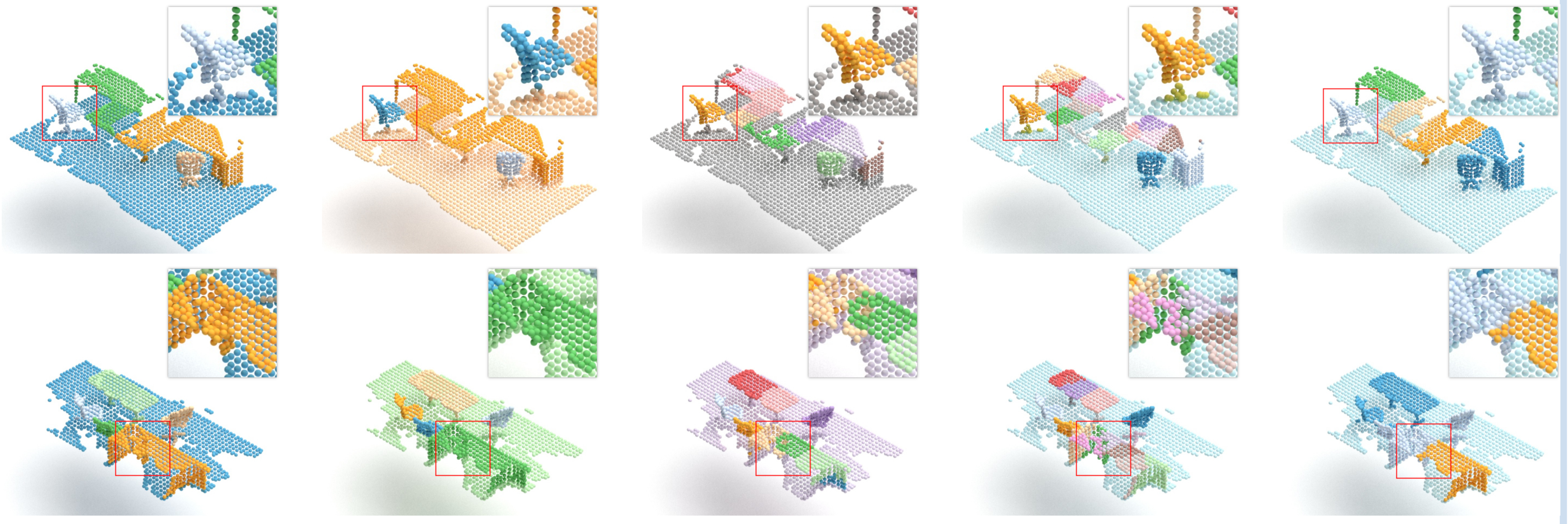
Datasets. We evaluate our method on synthetic datasets, such as the Unity Object Room (UOR) dataset and the Unity Object Table (UOT) dataset and real dataset such as S3DIS.

Results on UOR and UOT.

UOR / UOT	PG [1]	Ours	voxel size $0.75l$ / voxel size $1.25l$	$6-12$ objects	object matrix
ARI↑	0.976	0.915 ± 0.03	0.932	0.912	0.872
	0.923	0.901 ± 0.02	0.922	0.892	0.879
SC↑	0.907	0.832 ± 0.04	0.853	0.846	0.856
	0.917	0.835 ± 0.03	0.857	0.843	0.877
mSC↑	0.900	0.836 ± 0.04	0.850	0.842	0.861
	0.907	0.831 ± 0.03	0.861	0.834	0.886

Table 1. 3D point cloud segmentation results on UOR (blue) and UOT (red).

Results on S3DIS [4].



GT Label PG [1] Ours MS 0.06 MS 0.15
Fig.1 S3DIS Segmentation Results.

References

[1] Jiang, Li and Zhao, Hengshuang and Shi, Shaoshuai and Liu, Shu and Fu, Chi-Wing and Jia, Jiaya PointGroup: Dual-Set Point Grouping for 3D Instance Segmentation. In *CVPR*, 2020.

[2] L. Hubert and P. Arabie Convolutional Sequence to Sequence Model for Human Dynamics. In *Journal of Classification*, 1985.

[3] M. Engelcke and A. R. Kosiorek and O. Jones and I. Posner GENESIS: Generative Scene Inference and Sampling with Object-Centric Latent Representations. In *ICLR*, 2020.

[4] I. Armeni and A. Sax and A. R. Zamir and S. Savarese Joint 2D-3D-Semantic Data for Indoor Scene Understanding. In *arXiv preprint arXiv:1702.01105*, 2017.