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

Tianyu Wang, Miaomiao Liu, Kee Siong Ng

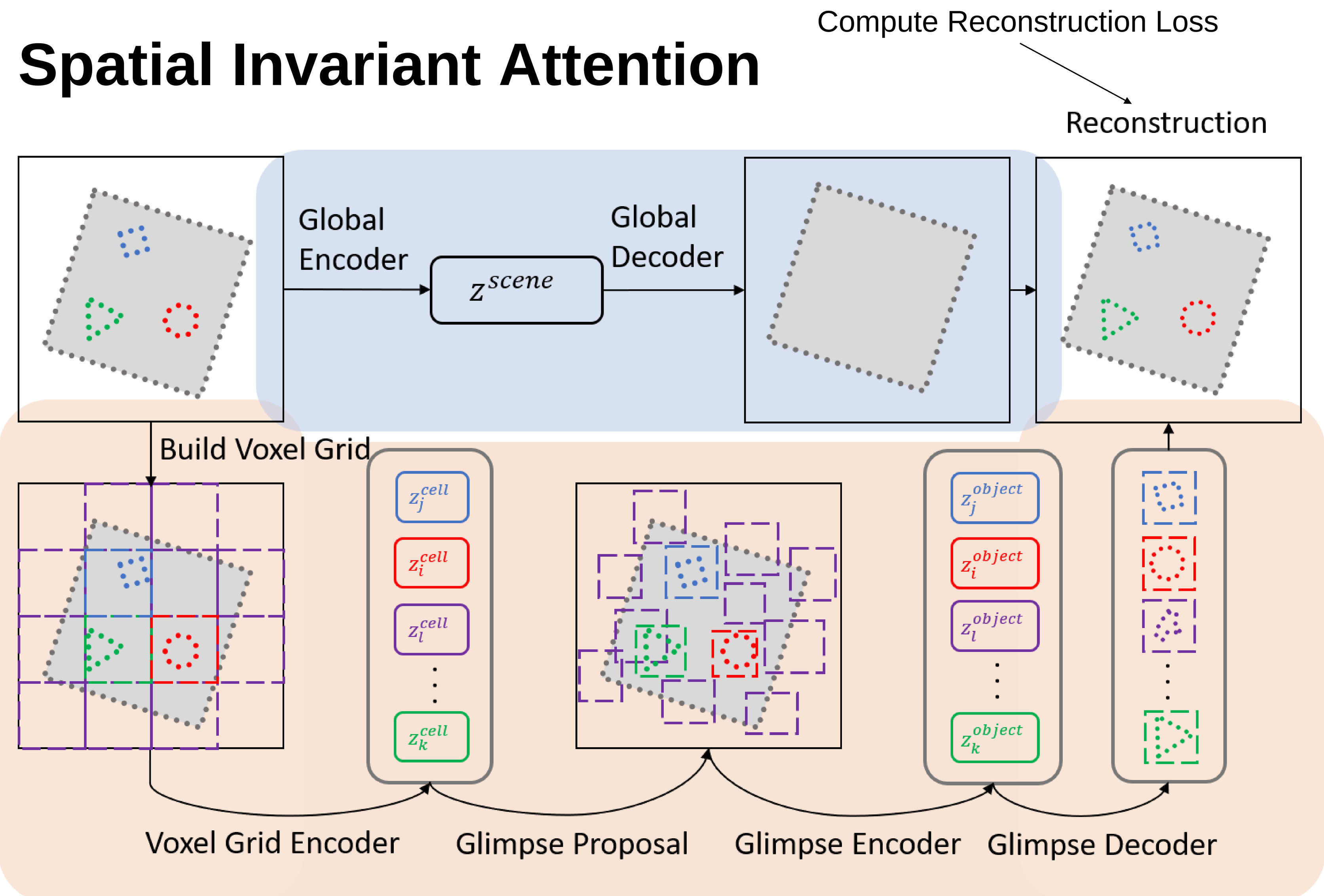
Overview

We tackle the unsupervised 3D point cloud object-centric problem combining variational inference and spatial invariant attention mechanism. We demonstrate that our model learns to detect object on both simulated and real-world dataset. t-SNE visualization shows our model learns meaningful object representations.

Motivation

Unsupervised object-centric learning is an act of defining objectness. We define an object to be a collection of highly correlated matters that can be exploited during a compression-decompression process. For a scene point cloud (colorless), we exploit variational autoencoder to trade-off between compression rate and reconstruction quality.

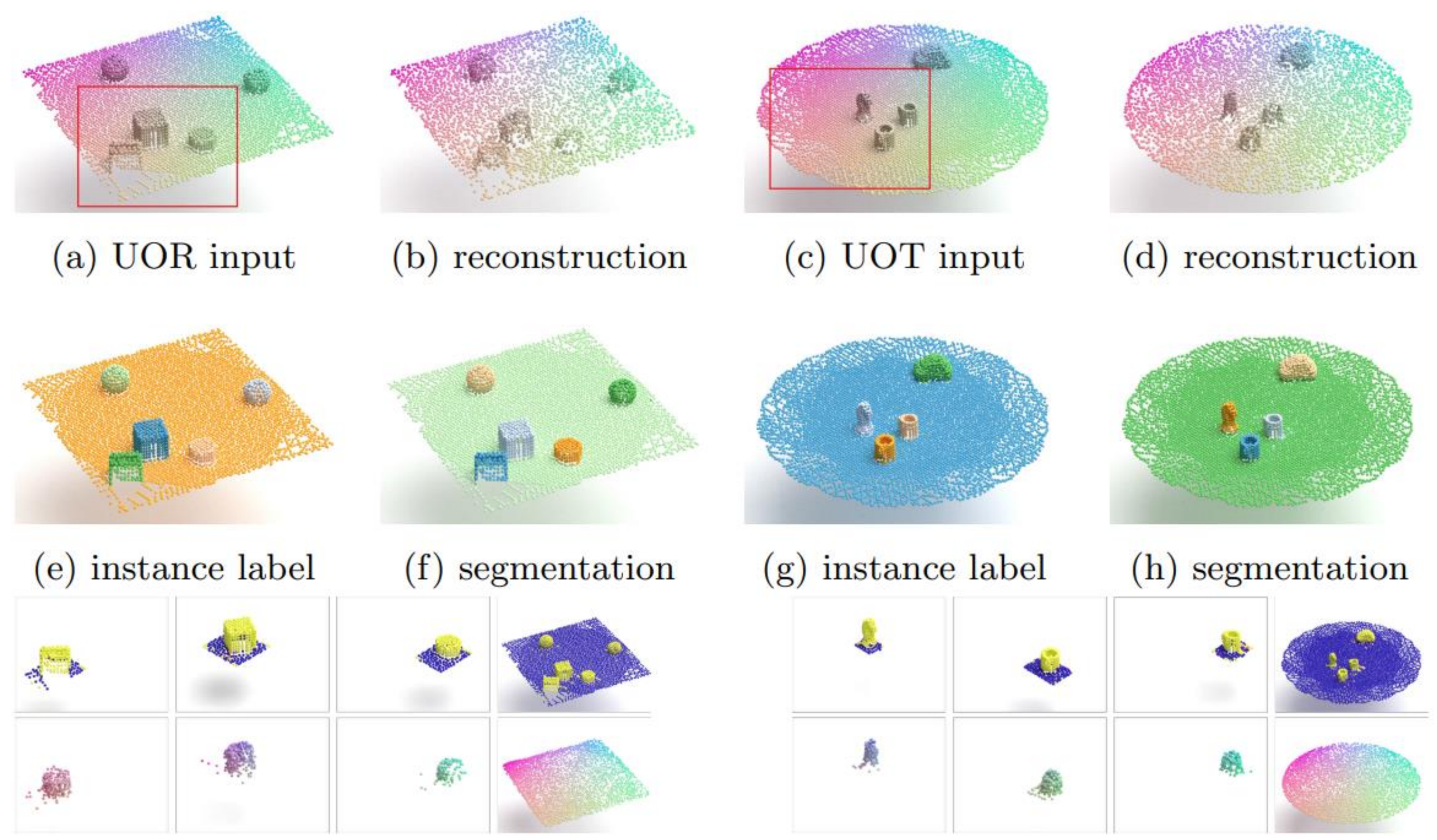
Spatial Invariant Attention



We evenly divide a 3D point cloud scene with a voxel grid and infer an object candidate from each grid cell. Each object candidate is defined by location, span $\mathbf{z}_i^{cell} = \{\mathbf{z}_i^{where}, \mathbf{z}_i^{apothem}\}$ and its structure, reconstruction weights and its existence flag $\mathbf{z}_i^{object} = \{\mathbf{z}_i^{what}, \mathbf{z}_i^{mask}, z_i^{pres}\}$.

Object-centric learning on simulated dataset

Segmentation and reconstruction results on simulated UOR and UOT dataset.PointGroup (PG) baseline is fully supervised.



(a) UOR input (b) reconstruction (c) UOT input (d) reconstruction

(e) instance label (f) segmentation (g) instance label (h) segmentation

(i) UOR close-up glimpses visualization, foreground alpha and scene layout reconstruction.

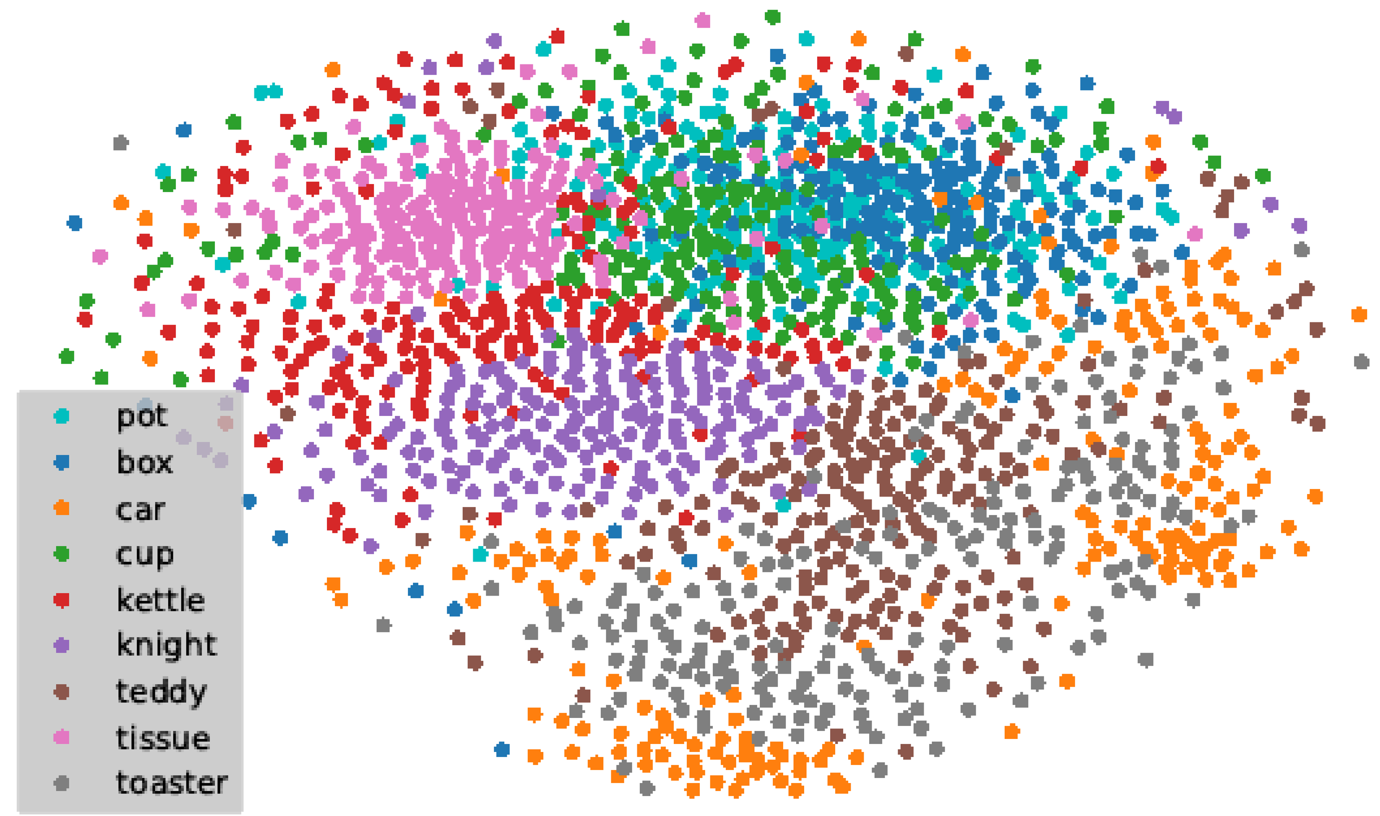
(j) UOT close-up glimpses visualization, foreground alpha and scene layout reconstruction.

	UOR / UOT	PG	SPAIR3D (ours)	voxel size $0.75l$ / voxel size $1.25l$	$6-12$ objects	object matrix
ARI↑	UOR	0.976	0.915 ± 0.03	0.932	0.912	0.872
	UOT	0.923	0.901 ± 0.02	0.922	0.892	0.879
SC↑	UOR	0.907	0.832 ± 0.04	0.853	0.846	0.856
	UOT	0.917	0.835 ± 0.03	0.857	0.843	0.877
mSC↑	UOR	0.900	0.836 ± 0.04	0.850	0.842	0.861
	UOT	0.907	0.831 ± 0.03	0.861	0.834	0.886

High quality segmentation that can generalize to scenes denser than training set.

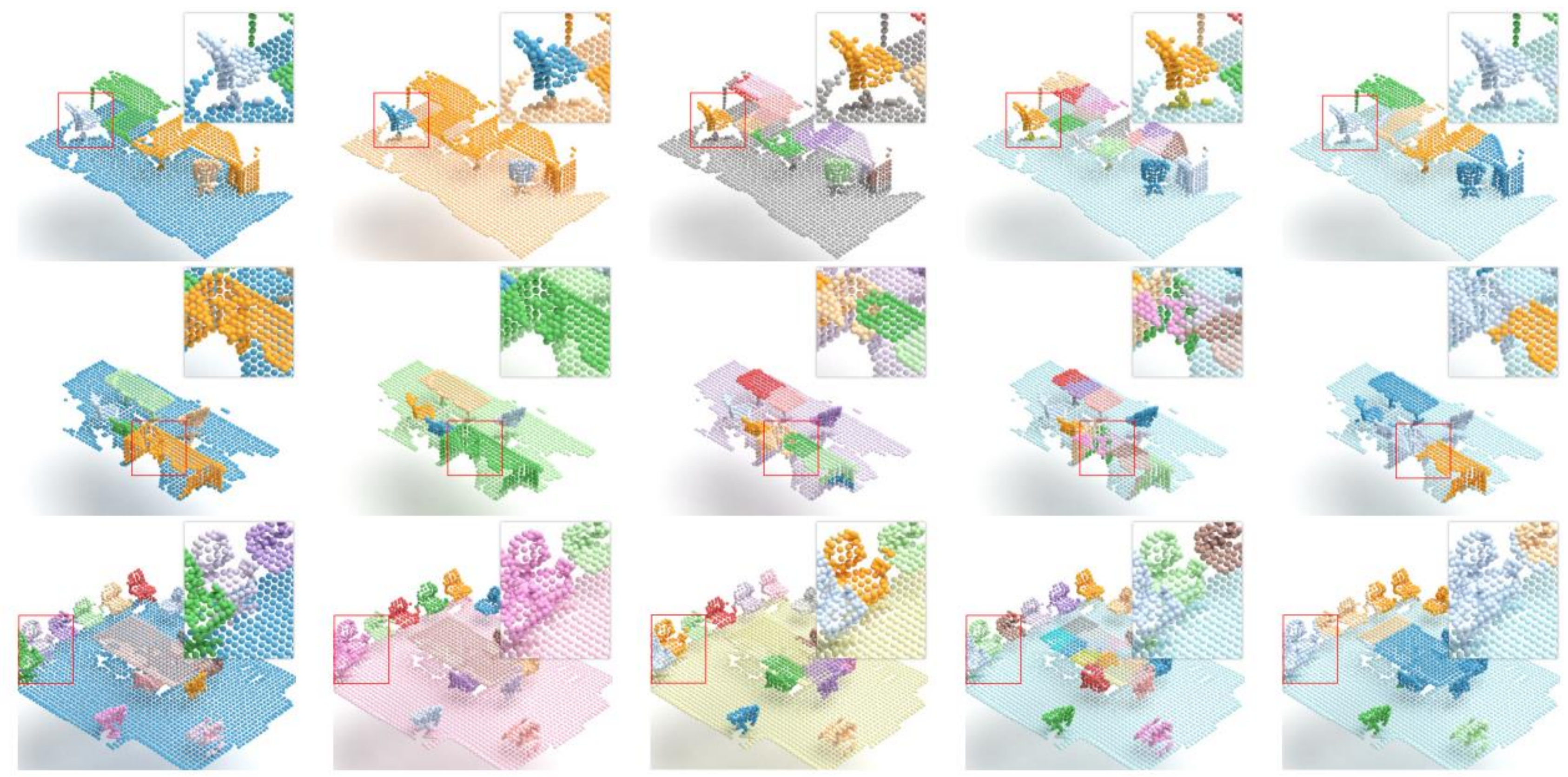
(a) instance label (b) segmentation (c) instance label (d) segmentation

Latent space visualization



Real-world scene segmentation

Segmentation and reconstruction results on simulated UOR and UOT dataset compared with PointGroup (PG) (supervised baseline) and mean-shift (rule-based baseline).



(a) GT Label (b) PG (c) Ours (d) MS 0.06 (e) MS 0.15

		Chair ↑	Table ↑	Sofa ↑	macro-avg ↑
PG	S	0.61	0.69	0.52	0.60
MS 0.06	U	0.75	0.34	0.36	0.48
MS 0.15	U	0.33	0.46	0.38	0.39
SPAIR3D (ours)	U	0.59	0.43	0.49	0.50

Per-class mIoU score

Australian National University