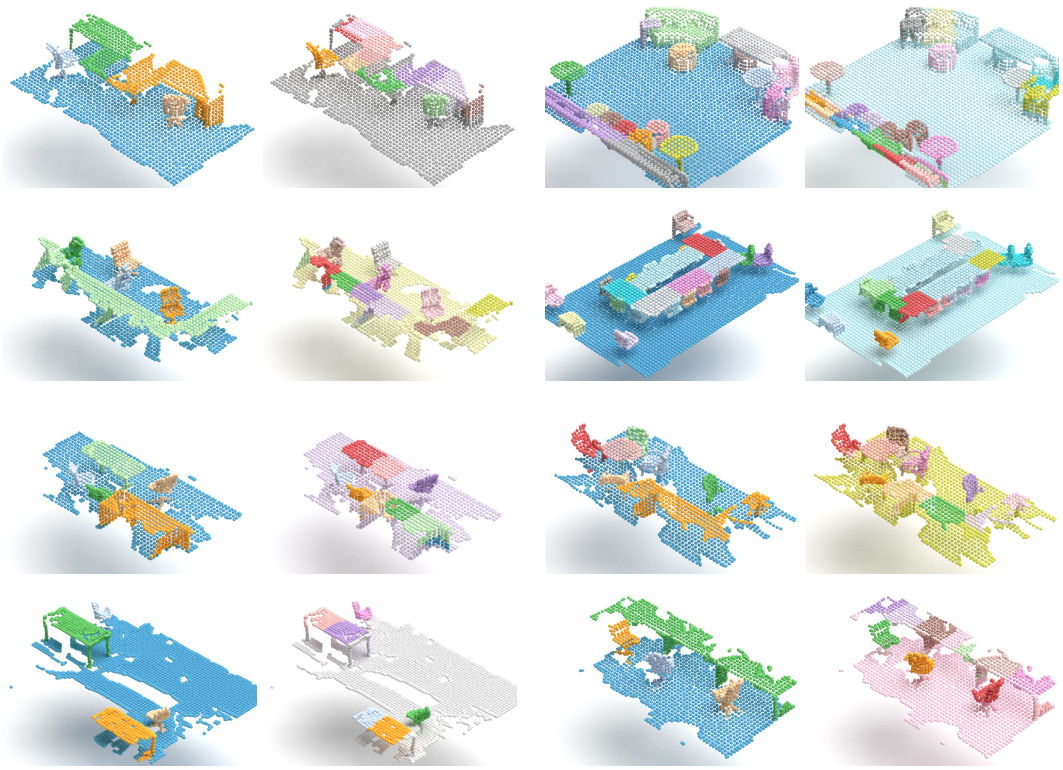
ADDITIONAL RESULTS

Anonymous authors
Paper under double-blind review

1 RESULTS ON S3DIS DATASET

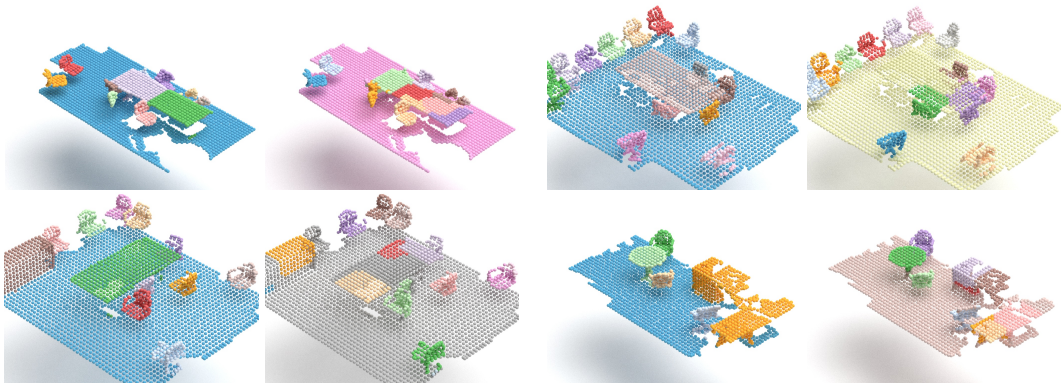
To further evaluate the effectiveness of our framework, we train our model on S3DIS (Armeni et al., 2017) dataset which includes 13 semantic categories. In particular, we focus on categories with regular structures, including chairs, tables, and sofas. Note that generative unsupervised object-centric learning models require a large amount of training data. For example, to achieve object-centric learning for 2D images in an unsupervised manner, IODINE (Greff et al., 2019) is trained on CLEVR (Johnson et al., 2017) dataset that consists of 100K images. MONET (Burgess et al., 2019) is trained on Object Room datasets with 1M scenes. For our reported results, we train our model on UOR and UOT datasets, which include 50K scenes each. In contrast, S3DIS dataset only consists of 271 rooms/scenes which is relatively very small and insufficient for unsupervised learning. As our approach focuses on object-centric learning, we thus manually inspect the dataset and remove rooms that are too empty (such as hallway), containing no regular objects such as clutter (such as storage room), and connected objects (such as lecture theater with connected chairs). After processing the dataset, we kept 174 scenes in total. We perform the data augmentation by applying random horizontal rotation, reflection, and crop to the point clouds. We report that our model achieves **AIR: 0.572**, **SC: 0.501**, and **mSC: 0.541**. When measured in mIoU, our model achieves **Chair: 0.603**, **Table: 0.375**, and **sofa: 0.482**, which demonstrates our performance on three main categories for foreground objects in the scene.

Qualitative results for our segmentation are shown in Fig. 1 and Fig. 2.



(a) (b) (c) (d)

Figure 1: S3DIS segmentation results. Column (a) and column (c) are instance labels. Column (b) and column (d) are the corresponding SPAIR3D segmentation results.



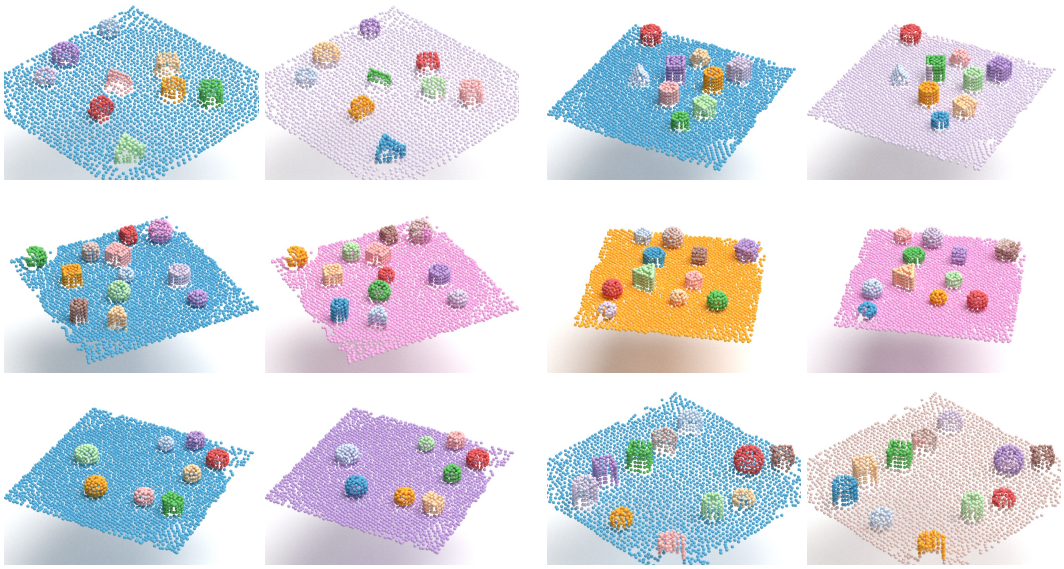
(a) (b) (c) (d)

Figure 2: S3DIS segmentation results. Column (a) and column (c) are instance labels. Column (b) and column (d) are the corresponding SPAIR3D segmentation results.

Those qualitative results show that chairs are largely segmented with high accuracy except for the ones that are placed under the table with incomplete structure. Long tables are over-segmented into multiple parts as their sizes are larger than the maximum glimpse size. Small round tables are well segmented. Note that the sizes of scenes vary from small offices to large meeting rooms and the spatially invariant property enables that the scene factorization process is not influenced by the sizes of the scenes. We expect, that with more training data, our model can achieve better results.

2 More Results on UOR Dataset

To demonstrate the effectiveness of our framework, we train our model on the UOR dataset but with different point densities or with noise. First, we set the point density to $\frac{2}{3}$ of the standard UOR dataset. Our model achieves **AIR: 0.921**, **SC: 0.827**, and **mSC: 0.836**. Qualitative results are shown in Fig. 3.



(a) (b) (c) (d)

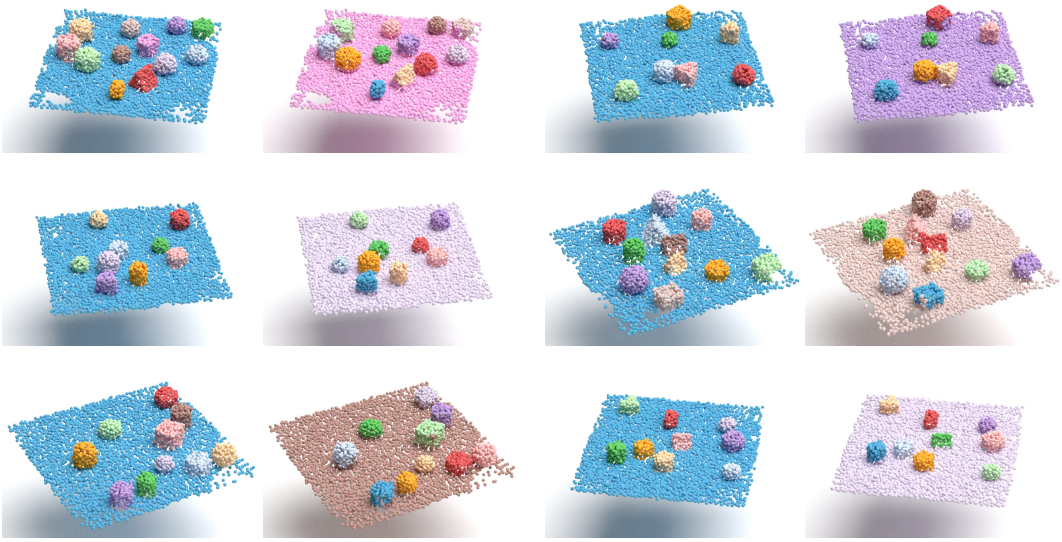
Figure 3: Segmentation results on UOR low density variant. Column (a) and column (c) are instance labels. Column (b) and column (d) are the corresponding SPAIR3D segmentation results.

Then, we set the point density to $\frac{4}{3}$ of that of the standard UOR dataset. Our model achieves **AIR: 0.914**, **SC: 0.831**, and **mSC: 0.833**. Qualitative results are shown in Fig. 4.

(a) (b) (c) (d)

Figure 4: Segmentation results on UOR high density variant. Column (a) and column (c) are instance labels. Column (b) and column (d) are the corresponding SPAIR3D segmentation results.

Finally, we add Gaussian distributed small perturbations to point coordinates. In particular, the standard deviation for the noise is 0.08, which corresponds to objects with the average object radius around 1. Our model achieves **AIR: 0.903**, **SC: 0.817**, and **mSC: 0.825**. Qualitative results are shown in Fig. 5.



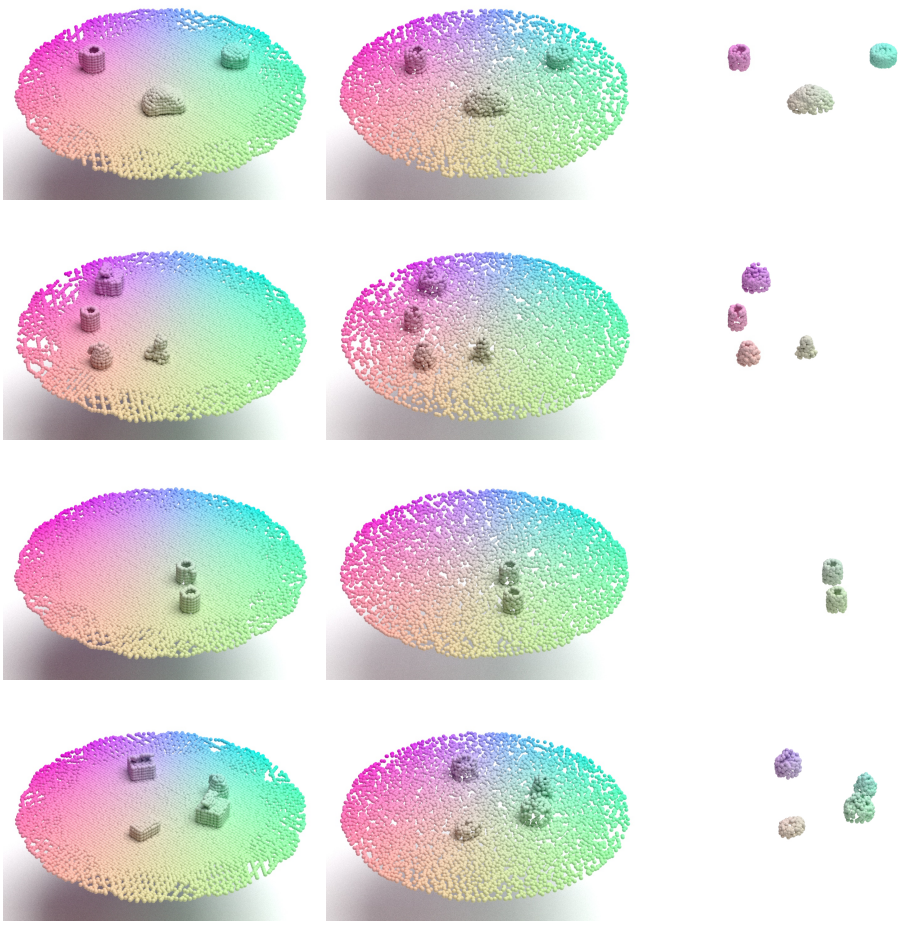
(a) (b) (c) (d)

Figure 5: Segmentation results on UOR noisy variant. Column (a) and column (c) are instance labels. Column (b) and column (d) are the corresponding SPAIR3D segmentation results.

The above three experiments demonstrate the robustness of our model against input data with varying point density and noises.

3 More Glimpse Visualization

In this section, we show more reconstruction visualization in individual glimpses.



(a) (b) (c)

Figure 6: Reconstruction results on UOT dataset. Column (a) is the raw input. Column (b) is the complete reconstruction. Column (c) is the visualization of all foreground glimpses with $z_{pres} > 0.5$.

References

I. Armeni, A. Sax, A. R. Zamir, and S. Savarese. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints*, February 2017.

Christopher Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. Monet: Unsupervised scene decomposition and representation. *ArXiv*, abs/1901.11390, 01 2019. URL `https://arxiv.org/abs/1901.11390`.

Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loïc Matthey, Matthew M Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *ICML*, 2019.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. pp. 1988–1997, 07 2017. doi: 10.1109/CVPR.2017.215.