Open-Vocabulary 3D Semantic Segmentation with Text-to-Image Diffusion Models

Xiaoyu Zhu[1†], Hao Zhou[2], Pengfei Xing[2], Long Zhao[2], Hao Xu[3], Junwei Liang[4], Alexander Hauptmann[1], Ting Liu[2], and Andrew Gallagher[2]

[1] Carnegie Mellon University
[2] Google DeepMind
[3] Google
[4] HKUST

1 Implementation Details

We implement our model using Pytorch [7]. To extract the semantic queries and masks from RGB images, we use Stable Diffusion [9] pre-trained on Laion-5B [11] as the feature backbone. The feature dimensions of diffusion and CLIP features are 256 and 768, respectively. The number of queries of Mask2Former [1] is 100. We use MinkowskiNet18A [2] as the backbone to encode 3D point clouds. The voxel size we use is 2cm following [8] for all datasets. We train the model for 200 epochs with a batch size of 8. Adam optimizer [4] is used with a learning rate of 0.0001. Polynomial learning rate policy is used as the learning rate scheduler with power 0.9. To train the mask distillation loss, the 2D images along with their corresponding 3D point clouds are used as model inputs. The predicted geometric masks from each 3D point cloud are only supervised by the salient masks extracted from the corresponding image. During inference, we ensemble the category logits from multiple image-point cloud pairs for each point by averaging them. In this way, the model can generate consistent and smooth category predictions. For the implementation of OpenScene [8], Openmask3D [13] and ConceptFusion [3], we use their official code bases to obtain the masks and the semantic embeddings of the masks for each scene. Then we compute the per-point class probability by averaging the probabilities of all the masks that contain this point. The class label that has the highest probability is selected for each point.

2 More Experimental Results

2.1 Quantitative Analysis

Comparison with LSeg [5]. We did not compare with LSeg in the main paper as it is pre-trained with ADE20K [14] while our model and all the baselines are pre-trained on MSCOCO [6]. Here, we show the comparison in Table 1. We observe that our model outperforms the OpenScene model with LSeg on the challenging ScanNet200 dataset.

†This work was partially done while the author was a student researcher at Google.

Table 1: Comparison with LSeg on ScanNet200.

Method	Head	Common	Tail	All
OpenScene(LSeg-2D)	22.7	3.8	1.1	9.1
OpenScene(LSeg-3D)	19.3	0.5	0.0	6.5
OpenScene(LSeg-2D/3D)	20.5	2.4	0.7	7.8
Ours	**25.6**	**11.5**	**6.9**	**14.2**

mAcc Results. We primarily use mIoU as the evaluation metric as it considers false positives. We compute the mAcc results on Scannet200, and find that our model achieves the best performance (26.5%) when compared to OpenScene(OS-2D) (25.4%), OpenScene(OS-3D) (12.0%), and OpenScene(OS-2D/3D) (22.6%).
Quantitative results for visual grounding. We perform evaluation on Nr3D. We match the predicted mask and K ground truth masks, and use the matching accuracy as the evaluation metric. We observe that our method achieves the best performance. Besides, our model and fully-supervised 3D visual grounding models (*e.g.* PLA) have different settings and cannot be directly compared.

Table 2: Experimental results on the visual grounding task.

Method	K=# of distractors	K=10	K=15
OpenScene(OS-2D)	79.1	20.0	12.4
OpenScene(OS-3D)	74.3	19.1	13.3
OpenScene(OS-2D/3D)	72.4	21.0	14.3
Ours	**81.9**	**22.9**	**17.1**

2.2 Qualitative Analysis

Visualization Results on Scannet200 [10]. In Fig. 1, we provide qualitative analysis of our approach and OpenScene for the zero-shot 3D semantic segmentation task on the validation set of Scannet200 [10]. Compared with OpenScene, our model generates coherent and consistent masks thanks to the mask-instance representations. For example, the table mask predictions of our model in column 2 are coherent, while the table predictions of Openscene are incomplete. Moreover, our method predicts accurate semantic labels for both head and tail categories by leveraging both CLIP and diffusion features. Specifically, our model can make accurate category predictions for the chair category in column 1 and the toilet category in column 4, while OpenScene fails in those categories.
Visualization Results on Replica [12]. In Fig. 2, we provide qualitative analysis of our approach and OpenScene for the zero-shot 3D semantic segmentation task on the Replica dataset [12]. We observe that the model can make accurate class predictions in both head class (e.g. chair) and tail class (e.g. vase). OpenScene tends to misclassify the semantic labels for an entire object (e.g. the

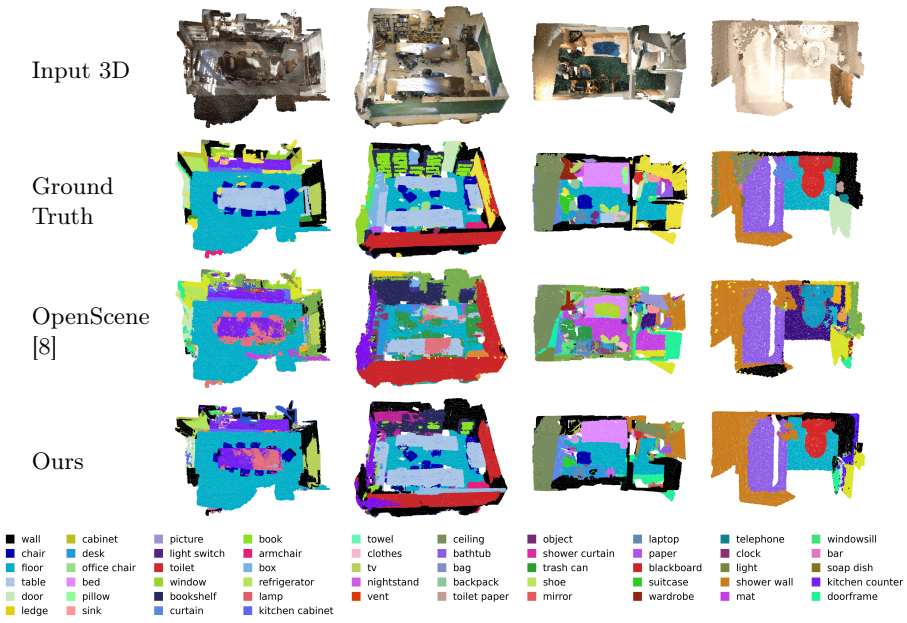

Fig. 1: Qualitative results from our model and OpenScene on zero-shot semantic segmentation. We visualize the segmentation results on the validation set of ScanNet200 [10]. We observe that our model can predict coherent masks with accurate semantic labels compared to OpenScene for both head and tail categories.

chair on the left of the table in column 1). Besides, we observe that our model can also correctly segment objects that were missed by human annotators. For example, in column 3, there is a lamp in the bottom right corner. The lamp was missed by the human annotator but can be correctly segmented and recognized by our method.

Visualization Results of Challenging Categories. In Fig. 3, we provide fine-grained visualizations of the challenging categories on the Scannet200 [10] and Replica [12] datasets. We observe that our model can predict accurate masks and category labels for small objects (e.g., book) and rare categories (e.g., vase and bench). This demonstrates the generalization ability of our method towards challenging and tailed categories.

Visualization Results of Model Ablations. In Fig 4, we visualize the segmentation results of our ablated models on the validation set of Scannet200 [10]. We observe that our model with learned 3D geometric masks can predict coherent masks with consistent semantic labels, compared to the model with 2D salient masks only. For example, the ablated model with salient masks only cannot predict accurate masks for the ceiling class in column 1. On the other hand, our model with learned geometric masks can predict accurate class boundaries for ceiling.

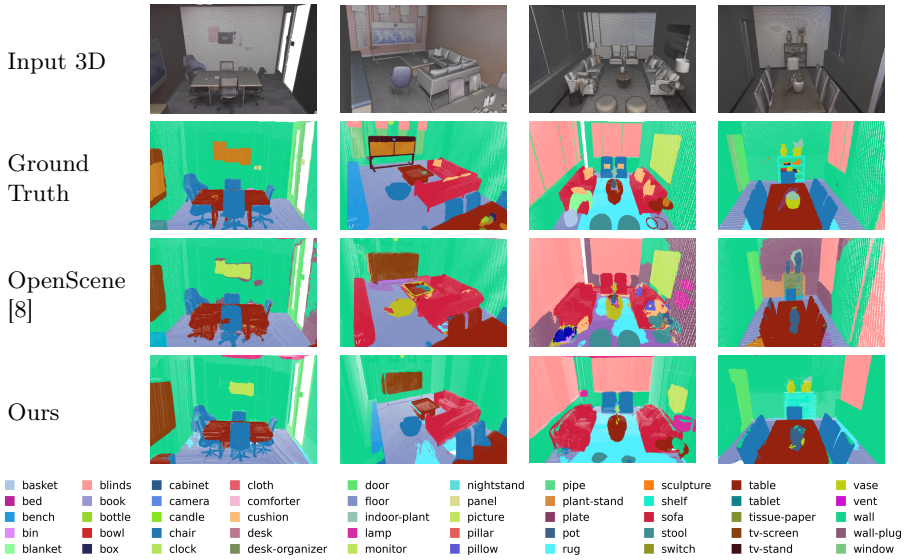

Fig. 2: Qualitative results from our model and OpenScene on zero-shot semantic segmentation. We visualize the segmentation results on the Replica dataset [12]. We observe that our model makes accurate mask predictions compared to OpenScene for both head and tail categories. we also find that our model can correctly segment objects that were missed by human annotators (lamp in column 3).

3 Limitations and Future Work

In this paper, we take the first step in leveraging frozen representations from large text-to-image diffusion models for open-vocabulary 3D scene understanding. Our model establishes new state-of-the-art in zero-shot 3D semantic segmentation and visual grounding tasks. Our method also demonstrates outstanding generalization ability towards unseen datasets and novel text queries. It opens a new direction for how to effectively leverage generative text-to-image models for other 3D scene understanding tasks in the future.

There are several limitations of the proposed model. First, while our model achieves better performance compared to OpenScene in small objects, it still misclassified some small and rare categories (e.g., rail). Second, we observe that the model can be easily confused by fine-grained categories that with similar semantic meaning. For example, the model sometimes wrongly classifies points of windowsill to the window class. In future work, it will be interesting to design models that can accurately distinguish between fine-grained categories in the open-vocabulary setting.

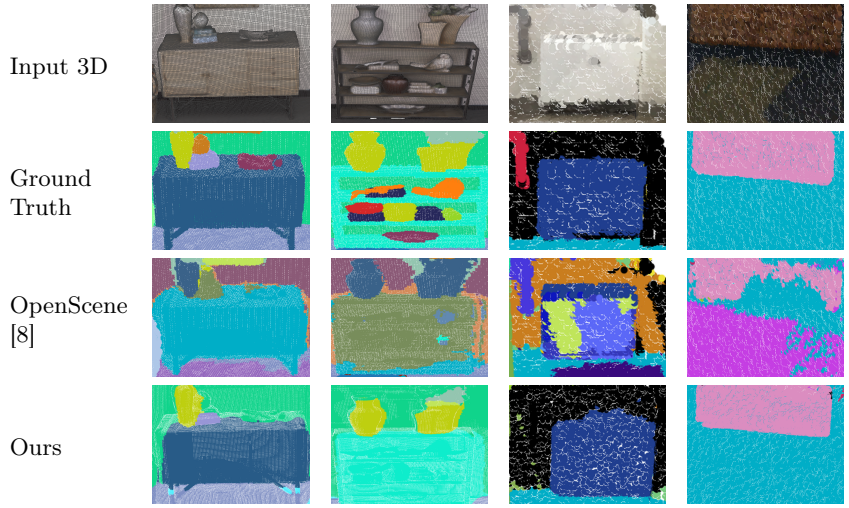

Fig. 3: Qualitative results from our model and OpenScene on challenging categories. We visualize the segmentation results on the Replica [12] and ScanNet200 [10] datasets. We observe that our model can predict coherent masks with accurate semantic labels compared to OpenScene for challenging categories, such as book (in column 1), vase (in column 1&2), radiator (in column 3), and bench (in column 4).

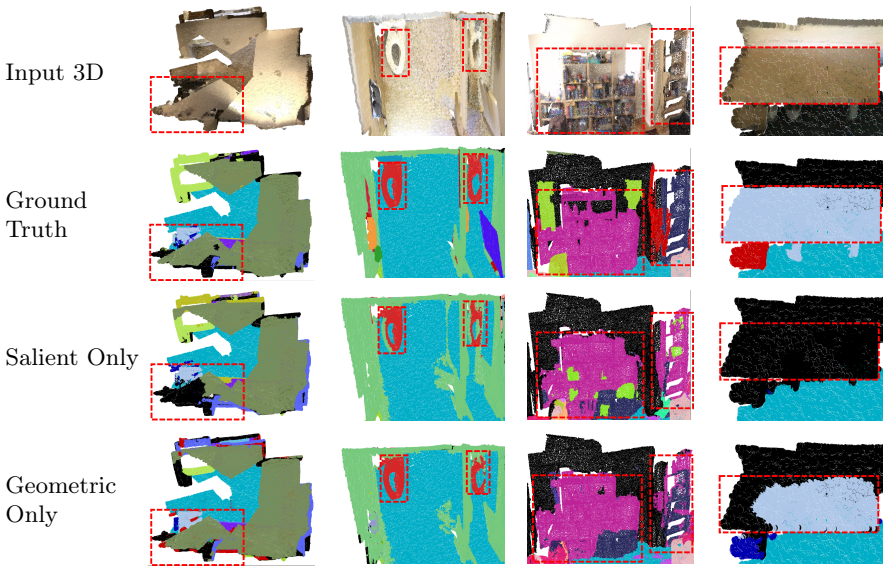

Fig. 4: Qualitative results from our model with salient-only and geometric-only masks. We visualize the segmentation results of our ablated models on the Scannet200 dataset [10]. We observe that our model with learned 3D geometric masks can predict accurate class boundaries compared to the model with 2D salient masks only.

References

1. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention mask transformer for universal image segmentation. In: CVPR (2022)
2. Choy, C., Gwak, J., Savarese, S.: 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: CVPR (2019)
3. Jatavallabhula, K., Kuwajerwala, A., Gu, Q., Omama, M., Chen, T., Li, S., Iyer, G., Saryazdi, S., Keetha, N., Tewari, A., Tenenbaum, J., de Melo, C., Krishna, M., Paull, L., Shkurti, F., Torralba, A.: Conceptfusion: Open-set multimodal 3d mapping. In: Robotics science and systems (2023)
4. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
5. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. In: ICLR (2022)
6. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollár, P.: Microsoft coco: Common objects in context. arXiv preprint arXiv:1405.0312 (2015)
7. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. NeurIPS **32** (2019)
8. Peng, S., Genova, K., Jiang, C.M., Tagliasacchi, A., Pollefeys, M., Funkhouser, T.: Openscene: 3d scene understanding with open vocabularies. In: CVPR (2023)
9. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
10. Rozenberszki, D., Litany, O., Dai, A.: Language-grounded indoor 3d semantic segmentation in the wild. In: ECCV (2022)
11. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. In: NeurIPS (2022)
12. Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., Engel, J.J., Mur-Artal, R., Ren, C., Verma, S., et al.: The replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
13. Takmaz, A., Fedele, E., Sumner, R.W., Pollefeys, M., Tombari, F., Engelmann, F.: Openmask3d: Open-vocabulary 3d instance segmentation. arXiv preprint arXiv:2306.13631 (2023)
14. Zhou, B., Zhao, H., Puig, X., Xiao, T., Fidler, S., Barriuso, A., Torralba, A.: Semantic understanding of scenes through the ade20k dataset. arXiv preprint arXiv:1608.05442 (2018)