GeoGuide: Hierarchical Geometric Guidance for Open-Vocabulary 3D Semantic Segmentation

Supplementary Material

A. Overview

In this supplementary material, we first provide additional implementation details in Sec. B. Next, we present detailed experimental results and ablation studies in Sec. C and additional analyses. We provide more qualitative visualizations across multiple datasets to demonstrate the superiority of our method in Sec. D. Finally, in Section E, we discuss current limitations and potential directions for future research.

B. Implementation Details

B.1. Network Architecture

We employ Sonata [10] as our 3D encoder to extract 3D geometric features for all datasets. We initialize the encoder with pre-trained weights provided by [10] and freeze all parameters during training and inference. To bridge the modality gap between 3D geometric features and 2D vision-language features, we introduce a lightweight 3D-2D adapter consisting of a two-layer Multi-Layer Perceptron (MLP). This adapter is the only trainable component of the feature extraction backbone, which maps 3D features into the vision-language embedding space. We maintain this identical architecture across the ScanNet v2 [3], Matterport3D [2], and nuScenes [1] datasets.

B.2. Training Configuration

Table 1 details the training hyperparameters for each dataset. Our training framework is built upon OpenScene [7] baseline, adapted to incorporate our geometry-guided learning objectives. Specifically, we implement the framework using PyTorch and employ the Adam [6] optimizer with an initial learning rate of 10^{-4}. A Linear learning rate scheduler is applied to ensure stable convergence. To prevent over-regularization of the lightweight adapter, the weight decay is set to 0. The voxel sizes are set to 2cm for the indoor datasets ScanNet [3] and Matterport3D [2] and 5cm for the outdoor dataset nuScenes [1]. During the training phase, we only train the 3D-2D adapter for cross-modal feature alignment, the MLP within the Uncertainty-based Superpoint Distillation module, and the linear projection layer in the Instance-level Mask Reconstruction module. For all three datasets, we use a batch size of 4 and train for 50 epochs. All experiments are conducted on a single NVIDIA A6000 GPU. The hyperparameters λ_1, λ_2, and λ_3 in the loss function are all set to 1.0. In terms of training efficiency, our GeoGuide framework enables convergence in approximately 20 epochs ($\sim$12 hours), whereas the OpenScene [7] baseline requires about 60 epochs ($\sim$16 hours) under the same hardware configuration.

Table 1. Training configurations for different datasets.

ScanNet v2/Matterport3D		**nuScenes**	
Config	**Value**	**Config**	**Value**
optimizer	Adam [6]	optimizer	Adam [6]
scheduler	Linear	scheduler	Linear
base lr	10^{-4}	base lr	10^{-4}
weight decay	0	weight decay	0
batch size	4	batch size	4
epochs	50	epochs	50
voxel size	2cm	voxel size	5cm
$\lambda_1/\lambda_2/\lambda_3$	1	$\lambda_1/\lambda_2/\lambda_3$	1

B.3. Multi-view Feature Fusion

Our multi-view fusion pipeline follows the protocol established in OpenScene [7]. We utilize 2D open-vocabulary models to extract image features and establish geometric correspondences between 3D points and 2D pixels via camera intrinsics and extrinsics. This process projects pixel-level features onto the 3D point cloud, creating 3D-aligned 2D features that serve as the distillation targets. Specifically, for ScanNet [3], we sample one frame every 20 images for feature extraction. For Matterport3D [2] and nuScenes [1], we use all available images to maximize scene coverage. For indoor datasets ScanNet [3] and Matterport3D [2] where depth maps are available, we perform occlusion testing to ensure that a pixel corresponds to only one visible 3D surface point. Specifically, for each 3D point, we project it onto the image plane using camera intrinsics and extrinsics to obtain the corresponding pixel location, then compute the depth difference between the point's camera-space depth and the pixel's depth value D. The pixel feature is assigned to the point only if this difference is below threshold $\tau_{\text{depth}} = 0.2D$, which effectively filters out occluded points while accounting for sensor noise.

B.4. Superpoint and Instance Mask Generation

For indoor datasets ScanNet v2 [3] and Matterport3D [2], we leverage the provided mesh data to generate superpoints. Following SAS [5], we apply the graph-based segmentation algorithm [4] on mesh normals to partition the scene into geometrically homogeneous regions. This algorithm groups adjacent mesh faces with similar normals into su-

perpoints, effectively preserving local geometric structures while maintaining computational efficiency. For the outdoor dataset nuScenes [1], the scene characteristics differ significantly: the point cloud is sparse, irregularly distributed, and dominated by the "road" category. Consequently, superpoint grouping offers limited benefits. We therefore treat each individual point as a superpoint, simplifying computation while allowing the framework to operate at the finest geometric granularity. For instance mask generation, we employ the class-agnostic instance segmentation method from SAI3D [11] to obtain instance masks for all three datasets.

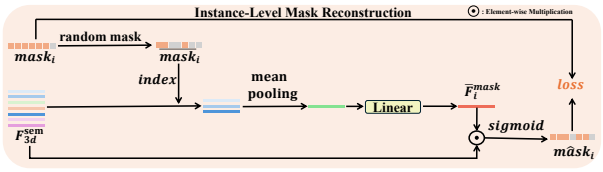

Figure 1. Illustration of the **Instance-level Mask Reconstruction (IMR)** module. We randomly mask a portion of an instance and reconstruct the complete mask from visible points, enforcing intra-instance semantic consistency.

B.5. Details of Instance-Level Mask Reconstruction

Figure 1 illustrates the detailed pipeline of our Instance-level Mask Reconstruction (IMR) module. Given an instance mask $mask_i$ obtained from the class-agnostic instance segmentation method [11], we first randomly mask out a portion of points to create an incomplete mask $\overline{mask_i}$. This incomplete mask is then used to index the corresponding semantic features F_{3d}^{sem} of the remaining visible points. These indexed features are aggregated through mean pooling to obtain a compact instance-level representation, which is subsequently transformed by a linear projection layer to produce the mask feature $\overline{F}_i^{mask}$. To reconstruct the complete instance mask, we compute the cosine similarity between $\overline{F}_i^{mask}$ and the features of all points in the scene F_{3d}^{sem}, followed by a sigmoid activation to obtain the predicted mask $\widehat{mask}_i$. The reconstruction loss is computed by comparing the predicted mask with the original complete mask $mask_i$. This mask reconstruction objective encourages the model to learn consistent semantic representations for all points belonging to the same instance, thereby enforcing intra-instance semantic consistency and enabling the network to recover complete instance structures from partial observations.

B.6. Inference Details

During inference, we discard the three proposed hierarchical guidance modules used during training. To extract text embeddings for open-vocabulary segmentation, we apply simple yet effective prompt engineering. Each category name "XX" is expanded into the template "a XX in a scene". This is consistent with OpenScene [7]. In our experiments, we report results from the pure 3D model without applying the 2D-3D ensemble strategy proposed in OpenScene [7], which fuses 2D predictions with 3D outputs at test time. Additionally, we do not employ the time-consuming second-stage self-training process used in SAS [5] and GGSD [9], which iteratively refines predictions using pseudo-labels. Despite this, our single-stage distillation approach achieves superior performance, highlighting the efficacy of our hierarchical geometry guidance.

For the nuScenes dataset, which contains ambiguous categories (e.g., "manmade," "vegetation"), we adopt the fine-grained label mapping strategy from OpenScene [7]. As detailed in Table 2, we first predict over this expanded label set of specific category names, then map predictions back to the original 16 categories. This strategy improves the model's ability to distinguish semantically similar but visually distinct objects, leading to more accurate predictions. For the indoor datasets, we directly utilize the predefined label names provided within the datasets without any modification, as they are already well-defined and unambiguous.

Table 2. Label mappings for nuScenes 16 classes.

nuScenes Labels	**Pre-defined Labels**
barrier	barrier, barricade
bicycle	bicycle
bus	bus
car	car
construction vehicle	bulldozer, excavator, concrete mixer, crane, dump truck
motorcycle	motorcycle
pedestrian	pedestrian, person
traffic cone	traffic cone
trailer	trailer, semi trailer, cargo container, shipping container, freight container
truck	truck
driveable surface	road
other flat	curb, traffic island, traffic median
sidewalk	sidewalk
terrain	grass, grassland, lawn, meadow, turf, sod
manmade	building, wall, pole, awning
vegetation	tree, trunk, tree trunk, bush, shrub, plant, flower, woods

C. More Experimental Results

C.1. Ablation Study on USD Module

In this section, we investigate the feature aggregation strategies within the Uncertainty-based Superpoint Distillation

Table 3. Ablation study on feature aggregation strategies in the USD module. We compare the effectiveness of using Weighted Pooling (✓) versus Mean Pooling (✗) for processing 2D and 3D semantic features.

2D	3D	mIoU	mAcc
✗	✗	55.81	67.23
✗	✓	55.63	66.95
✓	✓	56.13	67.55
✓	✗	**56.44**	**68.31**

(USD) module. Specifically, we analyze whether to apply uncertainty-based weighted pooling or standard mean pooling for the 2D semantic features and 3D semantic features when generating superpoint representations. The results are reported in Table 3. We observe that applying weighted pooling to 2D features significantly outperforms mean pooling. Comparing the first row (mean pooling for both) with the last row (weighted pooling for 2D, mean pooling for 3D), the mIoU improves from 55.81% to 56.44%. This validates our core motivation that 2D predictions contain noise and occlusion errors; uncertainty-based weighting effectively suppresses these unreliable signals while enhancing discriminative features. However, applying weighted pooling to 3D features tends to degrade performance (e.g., 55.63% in the second row vs. 55.81% in the first row). We hypothesize that the pre-trained 3D features already encode robust and dense geometric structures. Imposing learned weights on these features might disrupt the inherent geometric continuity and structural integrity of the 3D priors. Consequently, the optimal configuration, which we adopt in our final model, is to use weighted pooling for 2D semantic features to filter noise, while using mean pooling for 3D features to preserve complete geometric information.

C.2. Ablation Study on IMR Module

In this section, we further analyze the impact of the masking rate in the Instance-level Mask Reconstruction (IMR) module on the ScanNet v2 dataset. The results are summarized in Table 4. The experiment indicates that a masking rate of 0.6 yields the best performance, with mIoU of 56.82% and mAcc of 68.91%. This suggests that masking approximately 60% of the instance points provides the right balance: it creates sufficient reconstruction challenge to force the model to learn robust instance-level features, while retaining enough visible points for the model to identify the instance. Both extremely low (0.0) and high (0.8) masking rates lead to suboptimal performance. With 0.0 masking (no reconstruction task), the model achieves 56.51% mIoU, which is 0.31% lower than the optimal setting. This confirms that the reconstruction objective provides a valuable training signal. However, masking 80% of points also degrades performance to 56.58% mIoU, likely because the remaining 20% visible context is insufficient for reliable reconstruction, leading to noisy gradients that destabilize training.

Table 4. Ablation study of masking rate in the IMR module.

Masking Rate	**mIoU (%)**	**mAcc (%)**
0.0	56.51	68.74
0.2	56.39	68.43
0.4	56.44	68.39
0.6	**56.82**	**68.91**
0.8	56.58	68.49

Table 5. Ablation study on the IIRC module. We analyze the impact of geometric similarity constraints at the Superpoint (SP) level and Instance Mask level.

SP similar.	Mask similar.	mIoU	mAcc
✗	✓	56.63	68.22
✓	✗	56.96	69.06
✓	✓	**57.12**	**69.30**

C.3. Ablation Study on IIRC Module

We further evaluate the effectiveness of different granularity levels in the Inter-Instance Relation Consistency (IIRC) module for geometric relation modeling on the ScanNet v2 dataset. We examine the impact of applying similarity constraints at the superpoint level (local) and the Instance Mask level (global). The results are summarized in Table 5.

The results demonstrate that both constraints contribute to performance improvements, but they operate at different scales. Using only instance mask constraints yields an mIoU of 56.63%, while using only superpoint constraints achieves an mIoU of 56.96%. The superpoint-level constraint proves slightly more effective individually, likely because superpoints provide denser and more fine-grained geometric guidance across the scene. Most importantly, the best performance of 57.12% mIoU is achieved when both constraints are employed simultaneously. This indicates that local and global geometric relations are complementary. The superpoint constraints ensure local geometric consistency within object parts, while the mask constraints preserve the global topological relationships between instances. By jointly optimizing both, our method effectively aligns the semantic feature space with the hierarchical 3D geometric structure.

C.4. Detailed Results on Semantic Segmentation

We present comprehensive per-category semantic segmentation results for ScanNet v2 [3], Matterport3D [2], and nuScenes [1] datasets. Results for OpenScene [7] and SAS [5] are reproduced using their officially provided weights. As shown in Tables 7, 8, and 9, our method

achieves the best performance in the majority of categories across all three datasets. The segmentation task requires the model to effectively capture fine-grained semantic information across diverse object categories. Our approach significantly enhances geometry-semantic consistency modeling by integrating hierarchical geometric priors, which play a crucial role in improving segmentation performance. This hierarchical geometry guidance allows the network to better preserve 3D geometric structures and spatial relationships during the 2D-to-3D knowledge distillation process, leading to state-of-the-art performance in open-vocabulary 3D semantic segmentation.

C.5. Sensitivity Analysis of Instance Mask Quality

To further evaluate the robustness of our proposed framework, we conduct a sensitivity analysis regarding the quality of the instance masks utilized in the Instance-level Mask Reconstruction (IMR) module. Table 6 presents the segmentation performance on the ScanNet v2 dataset when employing different instance mask sources, including SAI3D [11], Mask3D, and Ground Truth (GT) masks.

As demonstrated in the results, the performance gap between using predicted masks (SAI3D, Mask3D) and GT masks is negligible (less than 1.0 mIoU). This observation validates our core design motivation: the IMR module utilizes instance masks primarily as coarse structural constraints to enforce intra-instance semantic consistency and prevent the degradation of 3D geometric priors, rather than relying on them for strict, pixel-perfect supervision.

Furthermore, we deliberately prioritize SAI3D, a class-agnostic instance segmentation method, to avoid any potential domain prior leakage that might arise from fully supervised methods like Mask3D. Ultimately, these findings confirm that GeoGuide is highly robust to variations in mask quality and effectively leverages structural priors without becoming bottlenecked by upstream segmentation errors.

Table 6. Sensitivity analysis of the IMR module to different instance mask qualities. The negligible performance gap indicates that our method utilizes masks primarily as coarse structural constraints and is robust to mask variations.

Mask Source	**SAI3D** [11]	**Mask3D** [8]	**GT**
mIoU / mAcc	56.8 / 68.9	57.0 / 69.3	57.6 / 69.7

D. Qualitative Analysis

We provide extensive qualitative visualizations to demonstrate the improvements of our GeoGuide framework over the OpenScene [7] baseline across three datasets.

Visualization on ScanNet v2. We present qualitative results of semantic segmentation on the ScanNetV2 validation set, comparing OpenScene's predictions and our GeoGuide method. As shown in Fig. 2, GeoGuide is able to better segment complete instances with more accurate boundaries and improved semantic consistency across object regions.

Visualization on Matterport3D. We present the qualitative results of semantic segmentation on the Matterport3D validation set, comparing OpenScene's predictions and our GeoGuide method. As shown in Fig. 3, GeoGuide demonstrates superior capability in segmenting complete instances, particularly for complex indoor scenes with multiple object categories and varying scales.

Visualization on nuScenes. Visual comparisons with OpenScene [7] on semantic segmentation in nuScenes [1] are shown in Fig. 4. Our method shows improved performance in outdoor scenarios, particularly in distinguishing between road surfaces, vehicles, and surrounding objects.

E. Discussions

In this section, we discuss the limitations of our work and potential directions for future research. Our primary goal is to effectively integrate geometric priors from pretrained 3D models to enhance open-vocabulary 3D semantic segmentation. To address the key issue of geometry degradation during 2D-to-3D knowledge distillation, we propose hierarchical geometry-guided modules that enforce consistency at the intra-superpoint, intra-instance, and inter-instance levels. While our approach effectively preserves 3D geometric structures and achieves significant performance improvements across multiple benchmarks, certain challenges remain. Specifically, we observe suboptimal segmentation results in scenarios with complex geometric structures or ambiguous object boundaries, where our geometry-guided modules may not fully capture the intricate spatial relationships. We attribute this limitation to the inherent modality gap between frozen geometric features from pretrained 3D backbones and learnable semantic features distilled from 2D models, which may lead to inconsistent learning signals during training. To address this, a promising future direction is to explore more adaptive integration mechanisms that can better harmonize geometric priors with semantic knowledge across different scene complexities, leading to more robust and generalizable segmentation performance.

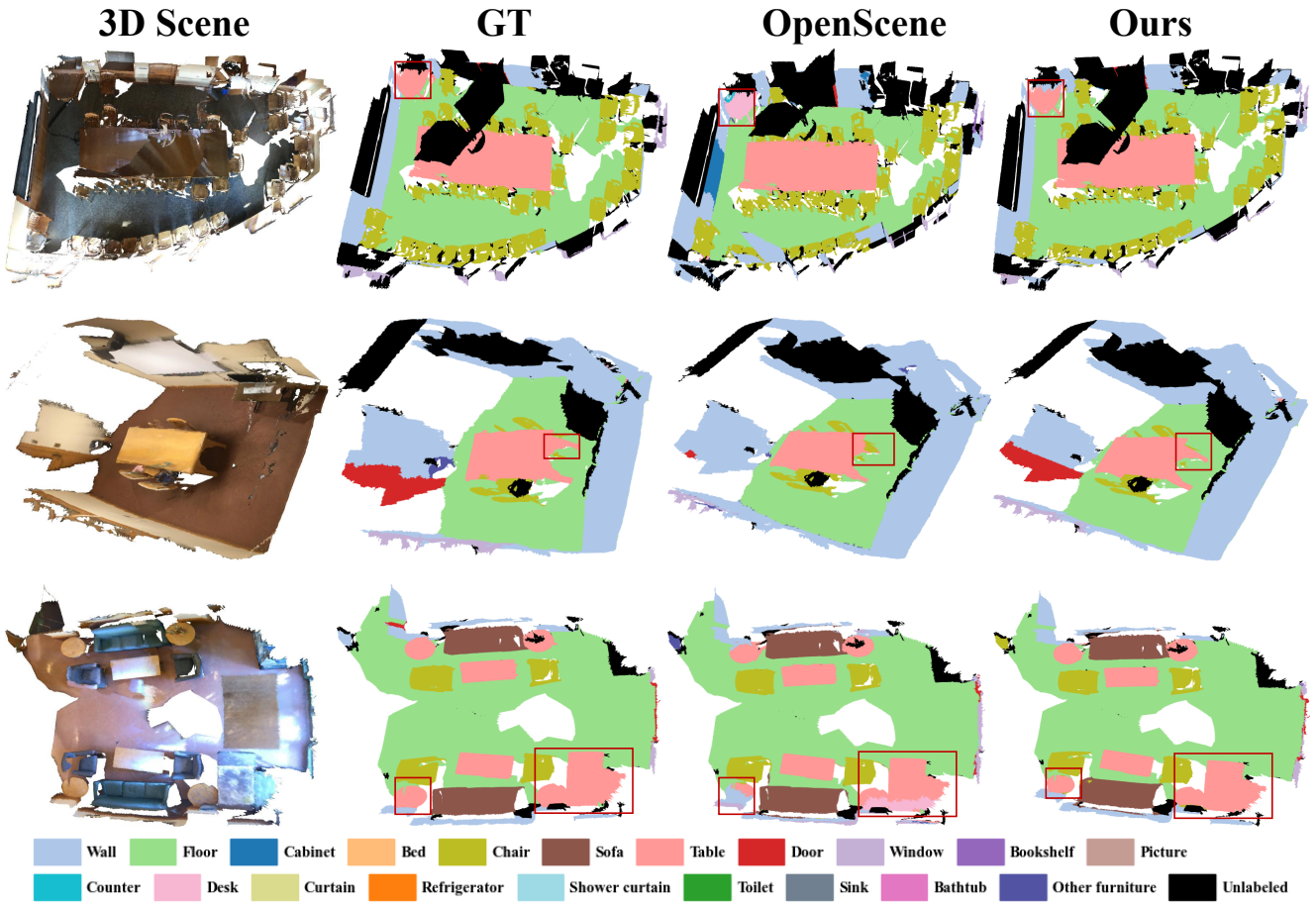

Figure 2. Qualitative comparison on ScanNet V2 validation set. Our method produces more complete and accurate instance segmentations compared to OpenScene.

Table 7. Full quantitative results on **ScanNet V2** validation set. We report the mIoU for each category. SAS* denotes using the same 2D features as SAS [5]. † denotes results reproduced by us using officially provided weights. **LS: LSeg, OS: OpenSeg**.

Method	**mIoU**	wall	floor	cabinet	bed	chair	sofa	table	door	window	shower curtain
SAS(stage2)† [5]	61.9	76.3	94.1	48.9	76.4	85.4	78.6	60.5	47.8	55.8	63.4
Ours (SAS*)	**64.8**	**80.5**	93.5	**52.3**	**77.2**	81.0	72.7	**63.4**	**60.4**	**62.5**	**74.1**
OpenScene(LS)† [7]	52.8	74.0	89.3	46.7	72.3	73.5	67.0	54.7	44.4	48.4	0.0
Ours (LS)	**59.8**	**80.3**	**92.9**	**49.6**	**76.4**	**79.0**	64.6	**60.6**	**59.1**	**61.9**	0.0
OpenScene(OS)† [7]	46.0	62.8	80.0	28.4	70.5	52.5	51.2	37.1	38.4	49.0	32.0
Ours(OS)	**53.4**	**71.5**	**88.1**	**37.4**	**75.5**	**59.4**	**54.4**	**41.5**	**49.7**	**58.9**	**63.0**

Method	picture	counter	desk	curtain	refrig-erator	toilet	sink	bathtub	book-shelf	other-furniture
SAS(stage2)† [5]	11.9	46.8	52.7	70.1	49.9	86.0	56.9	84.7	66.8	23.8
Ours (SAS*)	**19.7**	**49.7**	52.3	**75.2**	**52.0**	78.3	55.0	**87.6**	**69.2**	**29.3**
OpenScene(LS)† [7]	17.0	35.8	45.9	57.1	43.7	78.7	48.9	66.0	66.3	26.9
Ours (LS)	**23.3**	**50.8**	**52.1**	**63.7**	**51.4**	**80.2**	**56.8**	**86.8**	**69.4**	**31.9**
OpenScene(OS)† [7]	15.7	42.5	26.7	61.5	38.0	65.9	22.7	75.5	63.9	7.4
Ours(OS)	**23.4**	**47.4**	**31.6**	**69.1**	29.5	**72.3**	**22.8**	**88.0**	**64.4**	**11.6**

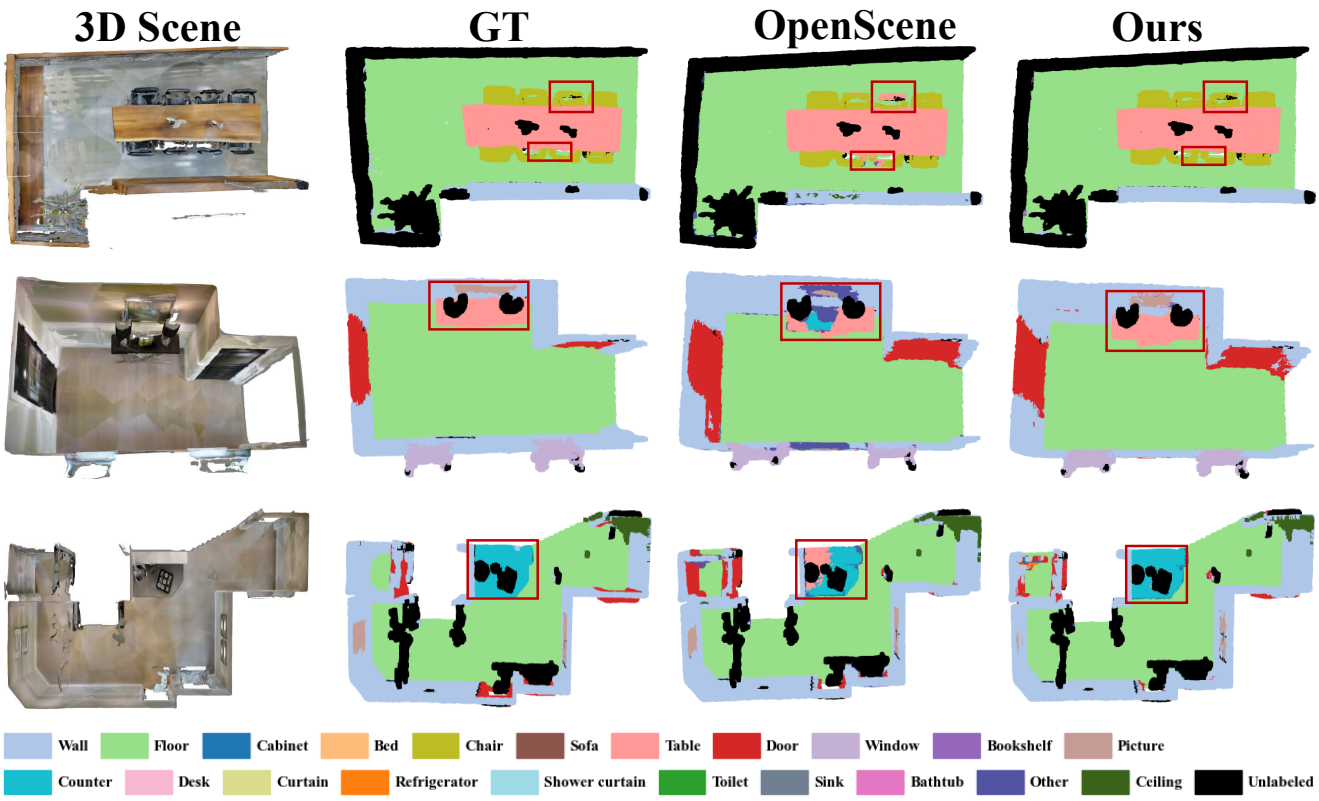

Figure 3. Qualitative comparison on Matterport3D validation set. GeoGuide achieves better instance completeness and boundary accuracy.

Table 8. Full quantitative results on **Matterport3D** test set. We report the mIoU for each category. SAS* denotes using the same 2D features as SAS [5]. † denotes results reproduced by us using officially provided weights. **LS: LSeg, OS: OpenSeg**.

Method	**mIoU**	wall	floor	cabinet	bed	chair	sofa	table	door	window	bookshelf
SAS(stage2)† [5]	48.6	90.6	94.7	46.8	89.3	79.7	76.0	56.6	52.0	41.8	44.8
Ours (SAS*)	**51.9**	85.4	**97.2**	**55.3**	**91.8**	79.4	**81.9**	55.4	**56.5**	**62.3**	**64.6**
OpenScene(LS)† [7]	41.9	83.2	95.1	40.8	81.5	72.9	70.0	48.3	51.4	39.6	25.2
Ours (LS)	**47.8**	**85.9**	**97.5**	**51.2**	**90.9**	**78.2**	**82.5**	**55.9**	**54.9**	**57.6**	**61.5**
OpenScene(OS)† [7]	41.1	62.2	94.4	57.8	85.3	52.6	84.3	50.8	80.9	43.4	14.4
Ours (OS)	**47.7**	**70.1**	**96.0**	**72.5**	**92.8**	49.9	**87.7**	**54.1**	**81.9**	**60.0**	**50.8**

Method	picture	counter	desk	curtain	refrig-erator	shower	toilet	sink	bathtub	other	ceiling
SAS(stage2)† [5]	33.3	47.9	12.3	68.9	27.3	82.7	87.1	34.2	67.4	10.7	94.5
Ours (SAS*)	**36.5**	43.7	**25.4**	**78.3**	**72.7**	75.8	84.3	**50.5**	**77.7**	7.5	**98.0**
OpenScene(LS)† [7]	34.6	33.1	0.0	73.1	0.0	0.0	80.0	39.5	71.4	37.8	95.7
Ours (LS)	**35.7**	**44.2**	**15.3**	**75.7**	**64.7**	0.0	**83.4**	**48.9**	**75.9**	11.6	**98.1**
OpenScene(OS)† [7]	32.2	36.9	25.1	73.8	23.5	28.0	84.7	53.3	74.3	3.9	97.2
Ours (OS)	**37.3**	**39.6**	**37.2**	**77.4**	**63.2**	**76.6**	**91.3**	**60.0**	**84.8**	**4.4**	**98.6**

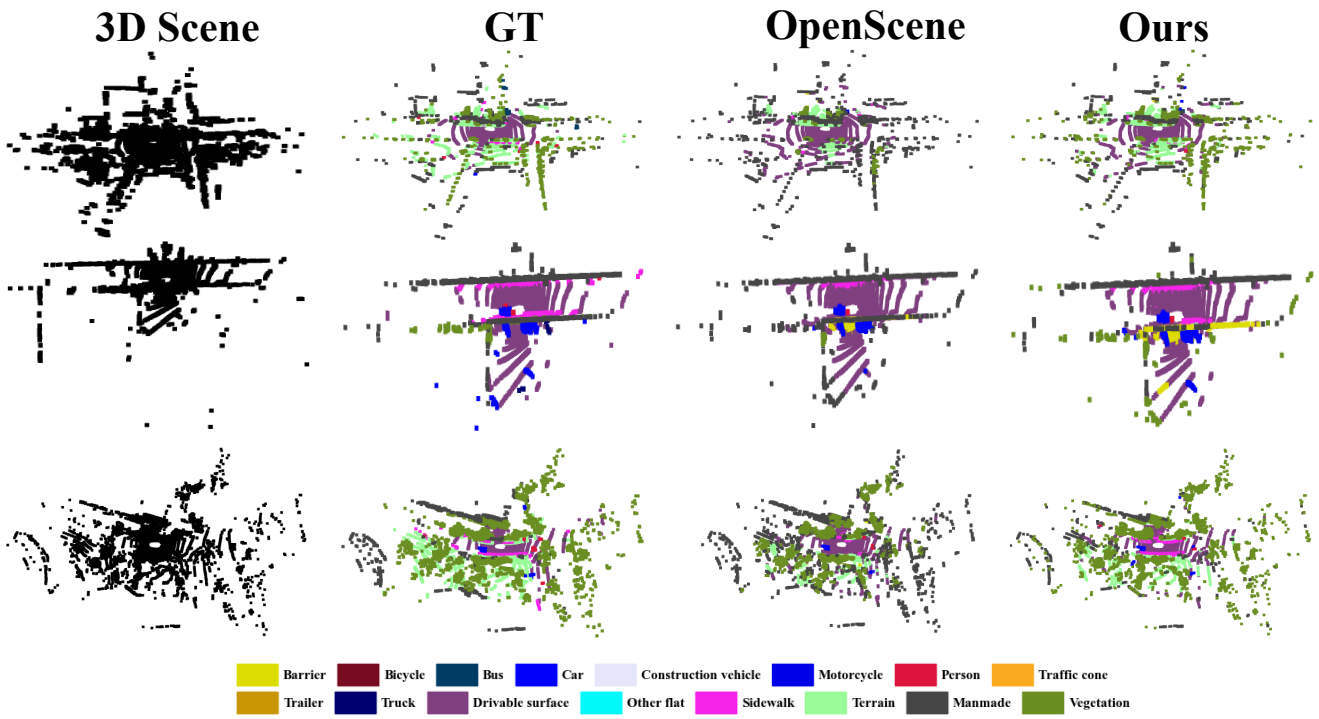

Figure 4. Qualitative comparison on nuScenes dataset. Our method demonstrates improved semantic segmentation in outdoor driving scenarios.

Table 9. Full quantitative results of mIoU on **nuScenes** validation set. SAS* denotes using the same 2D features as SAS [5]. † denotes results reproduced by us using officially provided weights. **LS: LSeg, OS: OpenSeg**.

Method	**mIoU**	barrier	bicycle	bus	car	construction vehicle	motorcycle	person	traffic cone
SAS(stage2)† [5]	48.2	16.7	0.0	79.8	71.6	30.1	64.3	68.3	21.4
Ours (SAS*)	**50.3**	**21.3**	0.0	**80.9**	**74.2**	25.4	**66.1**	**69.7**	**26.6**
OpenScene(OS)† [7]	42.9	16.9	0.0	72.6	75.7	21.6	61.1	60.7	20.8
Ours (OS)	**47.5**	**19.6**	0.0	**78.8**	**80.3**	**22.3**	**65.2**	59.4	**26.3**
OpenScene(LS)† [7]	37.8	0.04	0.0	56.5	79.2	0.0	49.8	49.8	0.0
Ours (LS)	**40.4**	**3.6**	0.0	**59.7**	**82.9**	0.0	**56.7**	45.5	0.0

Method	trailer	truck	drivable surface	other flat	sidewalk	terrain	manmade	vegetation
SAS(stage2)† [5]	10.1	56.3	89.0	0.0	51.3	60.3	68.9	82.2
Ours (SAS*)	**12.9**	**59.4**	86.7	0.0	**54.1**	58.6	**69.1**	**84.5**
OpenScene(OS)† [7]	19.2	43.6	78.7	0.0	31.7	51.4	64.7	68.4
Ours (OS)	**24.7**	**49.3**	74.3	0.0	**39.8**	**52.7**	**70.4**	**73.3**
OpenScene(LS)† [7]	0.0	55.7	78.2	0.0	28.3	56.2	72.5	76.6
Ours (LS)	0.0	54.8	76.9	0.0	**32.4**	**57.8**	**76.9**	**80.3**

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 1, 2, 3, 4

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1, 3

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1, 3

[4] Pedro F Felzenszwalb and Daniel P Huttenlocher. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2):167–181, 2004. 1

[5] Zhuoyuan Li, Jiahao Lu, Jiacheng Deng, Hanzhi Chang, Lifan Wu, Yanzhe Liang, and Tianzhu Zhang. Sas: Segment any 3d scene with integrated 2d priors. *arXiv preprint arXiv:2503.08512*, 2025. 1, 2, 3, 5, 6, 7

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[7] Songyou Peng, Kyle Genova, Chiyu Jiang, Andrea Tagliasacchi, Marc Pollefeys, Thomas Funkhouser, et al. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 815–824, 2023. 1, 2, 3, 4, 5, 6, 7

[8] Jonas Schult, Francis Engelmann, Alexander Hermans, Or Litany, Siyu Tang, and Bastian Leibe. Mask3d: Mask transformer for 3d semantic instance segmentation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 8216–8223. IEEE, 2023. 4

[9] Pengfei Wang, Yuxi Wang, Shuai Li, Zhaoxiang Zhang, Zhen Lei, and Lei Zhang. Open vocabulary 3d scene understanding via geometry guided self-distillation. In *European Conference on Computer Vision*, pages 442–460. Springer, 2024. 2

[10] Xiaoyang Wu, Daniel DeTone, Duncan Frost, Tianwei Shen, Chris Xie, Nan Yang, Jakob Engel, Richard Newcombe, Hengshuang Zhao, and Julian Straub. Sonata: Self-supervised learning of reliable point representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22193–22204, 2025. 1

[11] Yingda Yin, Yuzheng Liu, Yang Xiao, Daniel Cohen-Or, Jingwei Huang, and Baoquan Chen. Sai3d: Segment any instance in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3292–3302, 2024. 2, 4