Cross-Modal and Uncertainty-Aware Agglomeration for Open-Vocabulary 3D Scene Understanding

Supplementary Material

Jinlong Li[1,†] Cristiano Saltori[1] Fabio Poiesi[2] Nicu Sebe[1]
[1] University of Trento [2] Fondazione Bruno Kessler

Abstract

This supplementary material provides additional details and analysis to support the main paper, as follows:

- *In Sec. 1, we describe the choice of distillation losses for various 2D foundation models.*
- *In Sec. 2, we conduct more experiments in terms of combining both Lseg [1] and OpenSeg [3] to supervise the 3D model training.*
- *In Sec. 3, we illustrate more comparisons with AMRADIO [6].*
- *In Sec. 4, we provide parameters evolution in terms of the deterministic uncertainty estimation during the training.*
- *In Sec. 5, we provide more visualizations.*
- *In Sec. 6, we describe the potential limitation and future improvements.*

1. Distillation Loss Choice

In this section, we conduct more ablative studies to explain the distillation loss combination in our main paper. We first ablate three losses, including cosine similarity loss, L1 loss, and MSE (L2) loss, to explore how each one influences the 3D distilled model's open-vocabulary performance over the baseline model. As shown in Table 1, we can find that when supervising the 3D model training with Lseg that has been aligned with language spaces, cosine similarity loss strikes the best. In contrast, both L1 and L2 loss significantly deteriorates the open-vocabulary semantic segmentation performance. We surmise that L1 and L2 loss cannot force the 3D model to align with the language space, meaning degraded semantic information learning, however, it may help with learning geometric awareness. Regarding the pretraining mechanisms in terms of Lseg, DINOv2, and Stable Diffusion 2D foundation models, we opt to equip the 3D model with cosine similarity loss for Lseg and Stable Diffusion supervision since they have been trained with text embeddings. Moreover, we conducted additional analyses to study the effect of the 3D distillation model when attempting to use L1 or L2 loss to supervise from the DINOv2. As shown in Table 2, we can observe that adopting L1 loss for DINOv2 achieves the best one where we deduce this is due to the similar feature scales with Lseg's and L2 loss will decrease the supervised signals for the 3D model from feature embeddings which contains scales lower than 1.0. Therefore, we choose to use cosine similarity loss for Lseg and Stable Diffusion while L1 loss for DINOv2 in our main paper.

Table 1. Ablation: loss ablation about Cosine Similarity Loss, L1 Loss and MSE (L2) Loss.

Baseline_{Lseg}	$+CosineLoss2$	$+L1Loss$	$+L2Loss$	mIoU ↑	mAcc ↑
✓	✓			**51.4**	**62.3**
✓		✓		46.6	57.0
✓			✓	48.4	61.7

Table 2. Ablation: loss ablation for DINOv2 supervision when attempting with Cosine Similarity Loss, L1 Loss or MSE (L2) Loss.

Baseline_{Lseg}	$+SD$	$+DINOv2-Cosine$	$+DINOv2-L1$	$+DINOv2-L2$	mIoU ↑	mAcc ↑
✓	✓	✓			51.4	62.3
✓	✓		✓		**52.7**	**62.6**
✓	✓			✓	51.7	63.3

Table 3. Ablation: additional studies when introducing AMRADIO distillation.

Baseline_{Lseg}	$+DINOv2$	$+SD$	$+AMRADIO$	$+Unc$	mIoU ↑	mAcc ↑
✓					51.4	62.3
✓	✓				51.7	63.3
✓	✓	✓			52.7	62.6
✓	✓	✓		✓	53.5	64.2
✓	✓	✓	✓		52.2	62.6
✓	✓	✓	✓	✓	53.0	63.8

[†]Corresponding author: tyronejinlongli@gmail.com.

Table 4. Experimental results on ScanNetV2 and Matterport3D in terms of *val* on linear probing evaluation. Upperbound-full sup. denotes the fully-supervised upperbounding results while Baseline init. means initialize the model from our baseline model and then perform linear probing evaluation. Adding AMRADIO for comparision.

Type	Method	ScanNetV2 mIoU	ScanNetV2 mAcc
Upperbound-fully sup.	MinkowskiNet [2]	68.9	77.4
Baseline init.	MinkowskiNet [2]	54.4	64.7
Concat (Lseg+DINOv2+StableDiffusion)	3-heads concat	62.1	72.7
Concat (Lseg+DINOv2+StableDiffusion+AMRADIO)	4-heads concat	62.1	72.9

Table 5. Ablation: study on different open-vocabulary 2D semantic segmentation model supervision.

Lseg	OpenSeg	Lseg+OpenSeg	mIoU ↑	mAcc ↑
✓			**51.4**	62.3
	✓		44.0	62.8
		✓	48.5	62.8

2. Comparisons with AMRADIO

As the very recent work AMRADIO [6] proposes to distill knowledge from various 2D foundation models (including CLIP, DINOv2, and SAM) which act as teachers, we also employ extra experiments to ablate whether combining supervisions from AMRADIO model can help boost the 3D model distillation. As shown in Table 3, after extracting the 2D multi-view posed image embeddings and projecting them into the corresponding 3D space, we train the 3D model with feature supervisions from Lseg, DINOv2, Stable Diffusion and AMRADIO simultaneously. However, we discover this training will bring a slight performance drop, from 52.7% mIoU in our main paper to 52.2% mIoU in terms of the open-vocabulary semantic segmentation evaluation on ScanNetV2 *val* set. Furthermore, we also extend the proposed deterministic uncertainty estimation for this study, but no further improvement can be obtained. We analyze this is because AMRADIO has been constructed as a student model to acquire knowledge from 2D foundation teacher models, like CLIP, DINOv2, and SAM, resulting in there are no extra informative language or geometric knowledge from the student model, AMRADIO, even causing noisy supervision for the 3D model distillation. Additionally, a linear probing experiment is conducted in Table 4, there are no further improvements from the 3D model which is distilled from four 2D foundation models in parallel.

3. Lseg and OpenSeg Supervision Combination

In this section, we provide additional studies about combing both Lseg and OpenSeg 2D models, being trained to align with the text embeddings for the 2D image open-vocabulary semantic segmentation task. As shown in Table 5, a significant performance drop can be observed in terms of open-vocabulary semantic segmentation validation on ScanNetV2 *val* set, though both Lseg and OpenSeg have been aligned with the language space on dense-level supervisions. Since Lseg and OpenSeg present different mask segmentation results for pixels, especially encountering complicate contexts from rapidly changing posed images, this leads to confusion for the 3D model training, attempting to "cluster" the same local regions into different 'clusters'. Hence, we prefer to utilizing Lseg in our main paper, which is also based on the CLIP.

4. Parameter σ Evolution

As we validate in Table 6 in our main paper, we find that employing Auto-Weighting learning results in a trivial solution, we then provide the evolutions of parameters in terms of deterministic uncertainty estimation σ_i to help with better capturing the learning using our method CUA-O3D during the distillation in Fig. 3.

5. More Visualizations

In this section, we demonstrate more clustering results for ScanNetV2 and Matterport3D to further clarify the motivation of agglomerating heterogeneous and complementary feature supervisions from various 2D foundation models in Fig. 1 and Fig. 2, while open-vocabulary semantic segmentation visualizations are displayed in Fig. 4 and Fig. 5.

6. Limitations and Future Improvements

Our method CUA-O3D inspires three aspects about the potential limitations and future improvements, which can serve as starting points for future research. Firstly, the 3D model distilled from 2D Vision-Language Models shows significantly lower performance than the fully supervised baselines, necessitating more explorations to boost the performance and bridge this gap. Secondly, though our distilled model presents essential improvements over the baseline, there is still no in-depth study about the alignment between visual embedding (3D) and text embedding. Thirdly, how to naturally transform the backbone architectures from 2D to 3D within the multi-view 3D scene understanding shall be interesting, which inherits the strong generalizability and zero-shot capacities from 2D foundation models, while also exploring weakly/semi-supervised setting [4, 5, 8, 9] under the in-context learning or inducing video sequences data [7, 10] to help the model better understand the 3D scene to further enhance the 3D model leaves interesting insights.

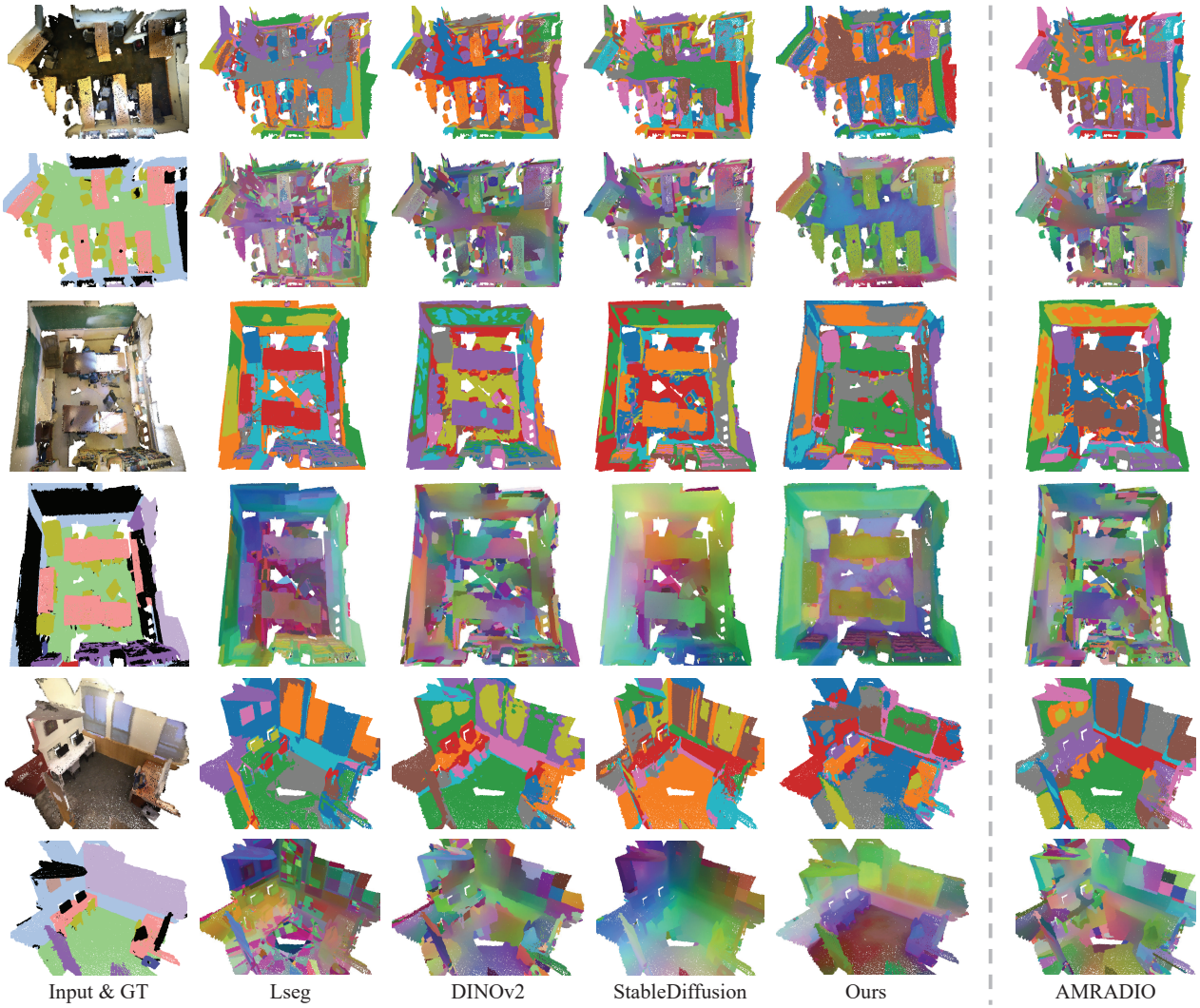


Figure 1. Clustering visualizations of ScanNetV2 from various 2D foundation models.

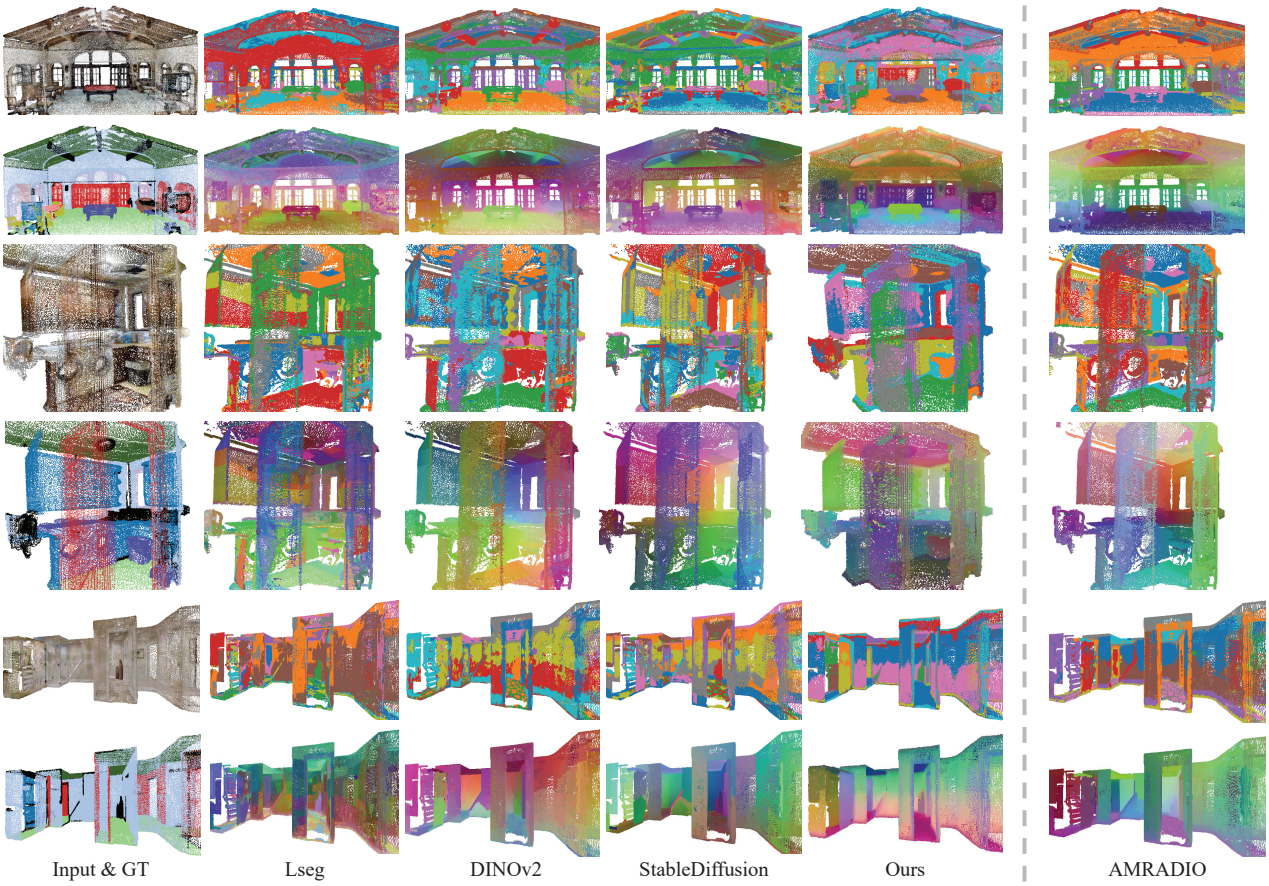
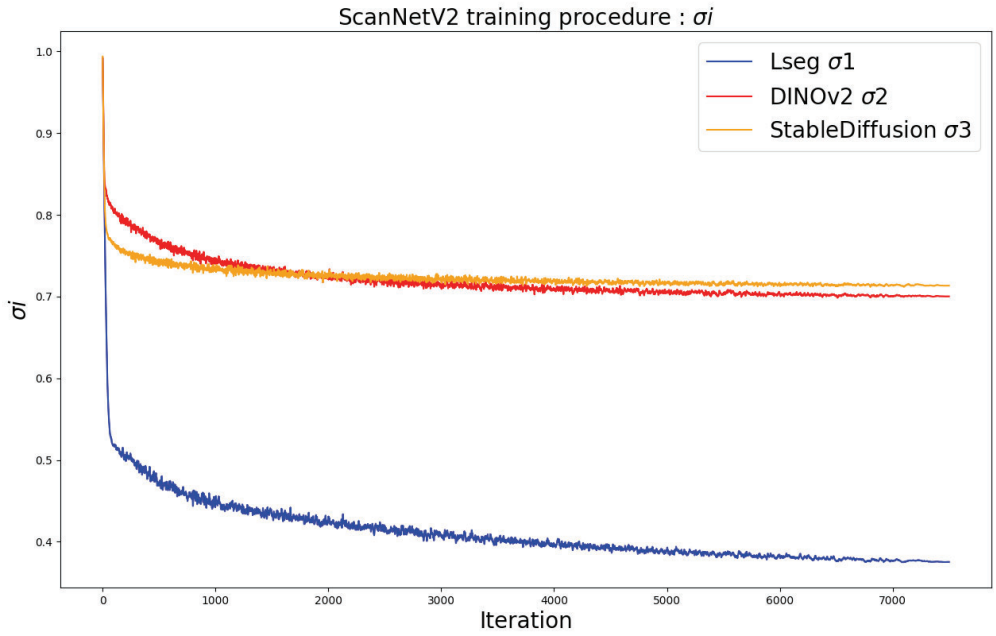


Figure 2. Clustering visualizations of Matterport3D from various 2D foundation models.



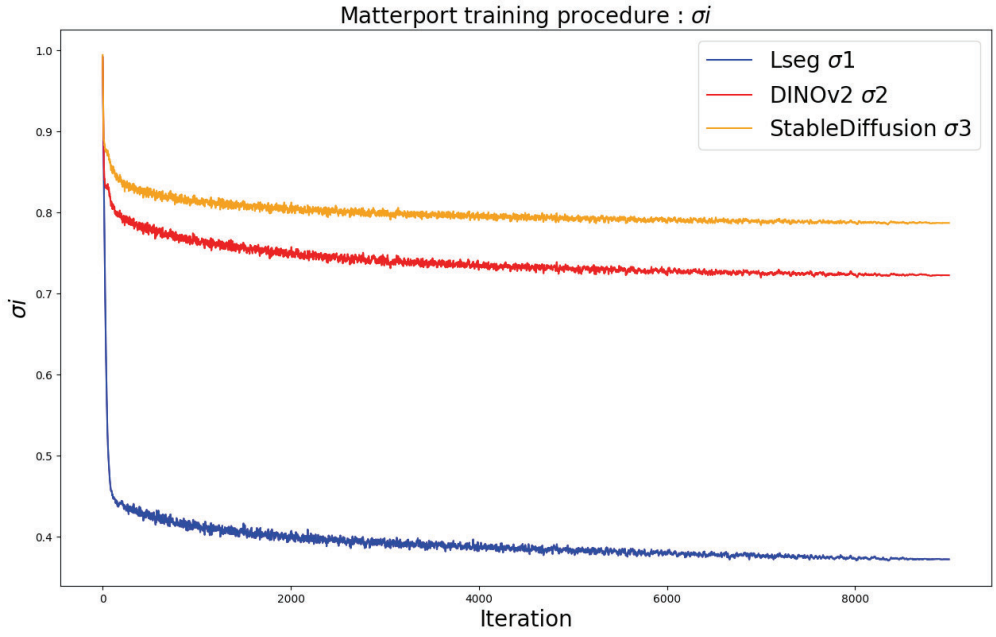


Figure 3. Evolutions of parameters in terms of deterministic uncertainty estimation σ_i.

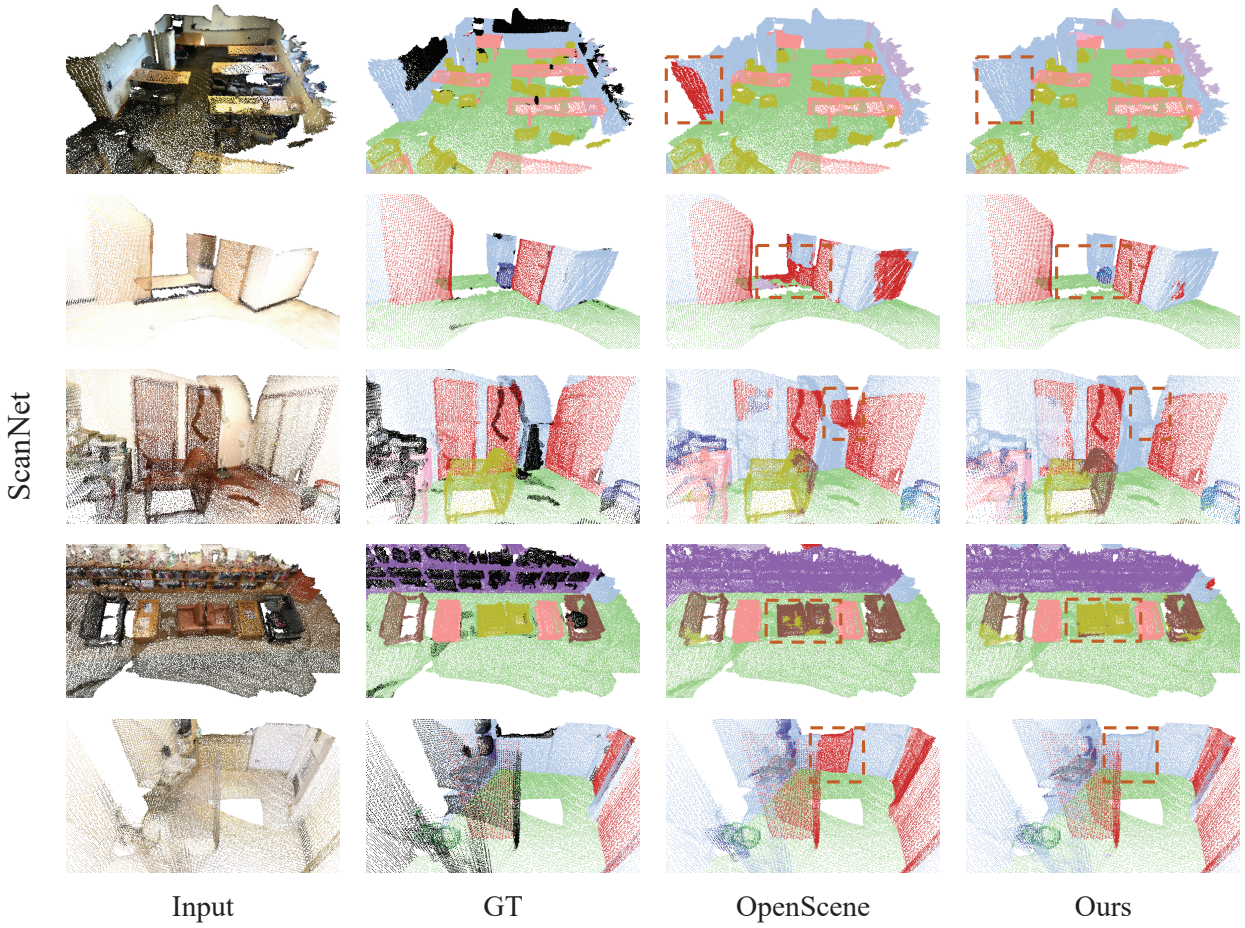


Figure 4. Open-vocabulary semantic segmentation visualizations of ScanNetV2.

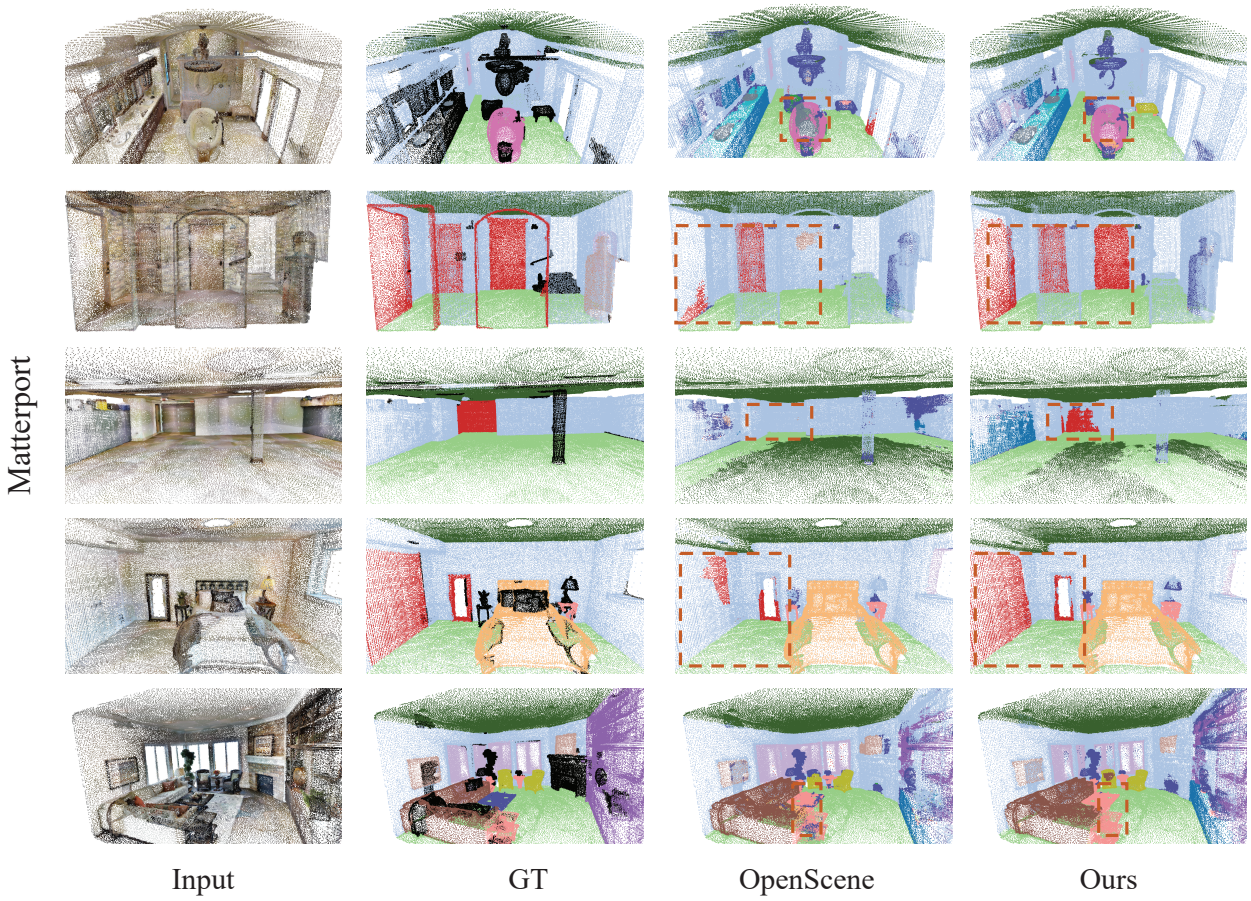


Figure 5. Open-vocabulary semantic segmentation visualizations of Matterport3D.

References

[1] L. Boyi, W. Kilian, B. Serge, K. Vladlen, and R. Rene. Language-driven semantic segmentation. In *ICLR*, 2022. 1

[2] C. Choy, J. Gwak, and S. Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *ICCV*, 2019. 2

[3] G. Ghiasi, X. Gu, Y. Cui, and T.-Y. Lin. Scaling open-vocabulary image segmentation with image-level labels. In *ECCV*, 2022. 1

[4] Jinlong Li, Zequn Jie, Xu Wang, Xiaolin Wei, and Lin Ma. Expansion and shrinkage of localization for weakly-supervised semantic segmentation. *Advances in neural information processing systems*, 35:16037–16051, 2022. 2

[5] Jinlong Li, Zequn Jie, Xu Wang, Yu Zhou, Xiaolin Wei, and Lin Ma. Weakly supervised semantic segmentation via progressive patch learning. *IEEE Transactions on multimedia*, 25:1686–1699, 2022. 2

[6] Mike Ranzinger, Greg Heinrich, Jan Kautz, and Pavlo Molchanov. Am-radio: Agglomerative vision foundation model reduce all domains into one. In *CVPR*, pages 12490–12500, 2024. 1, 2

[7] Yan Shu, Peitian Zhang, Zheng Liu, Minghao Qin, Junjie Zhou, Tiejun Huang, and Bo Zhao. Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024. 2

[8] Qi Zang, Shuang Wang, Dong Zhao, Yang Hu, Dou Quan, Jinlong Li, Nicu Sebe, and Zhun Zhong. Generalized source-free domain-adaptive segmentation via reliable knowledge propagation. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5967–5976, 2024. 2

[9] Pingping Zhang, Jinlong Li, Kecheng Chen, Meng Wang, Long Xu, Haoliang Li, Nicu Sebe, Sam Kwong, and Shiqi Wang. When video coding meets multimodal large language models: A unified paradigm for video coding. *arXiv preprint arXiv:2408.08093*, 2024. 2

[10] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 2