Supplementary Material of
All in One: Visual-Description-Guided Unified Point Cloud Segmentation

Zongyan Han[1], Mohamed El Amine Boudjoghra[2], Jiahua Dong[1],
Jinhong Wang[1], Rao Muhammad Anwer[1]
[1]Mohamed Bin Zayed University of Artificial Intelligence, Abu Dhabi, UAE
[2]Technical University of Munich, Germany
{zongyan.han, jiahua.dong, jinhong.wang, rao.anwer}@mbzuai.ac.ae
Mohamed.boudjoghra@tum.de

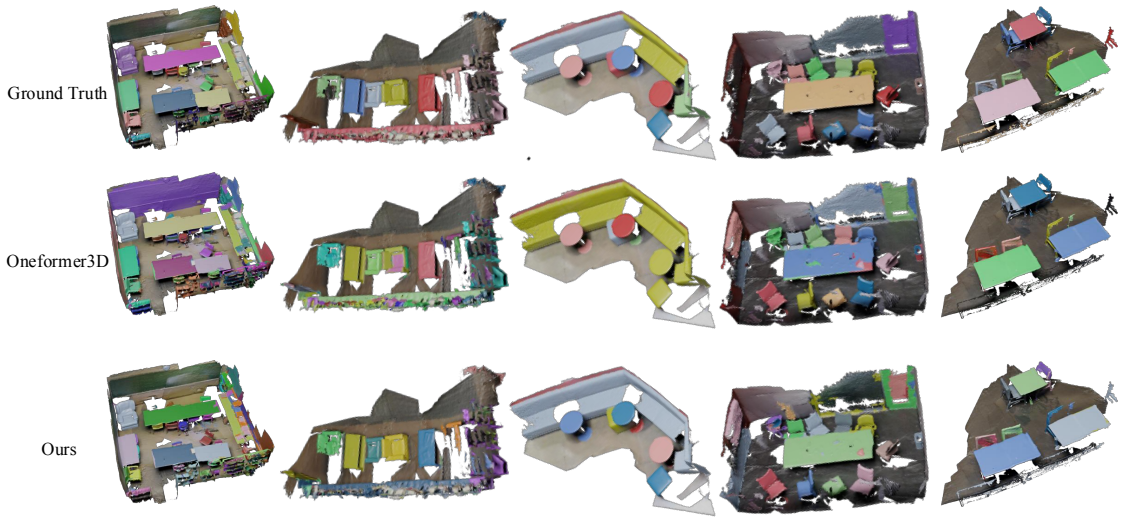

Figure 1. The qualitative results of the instance segmentation on ScanNet.

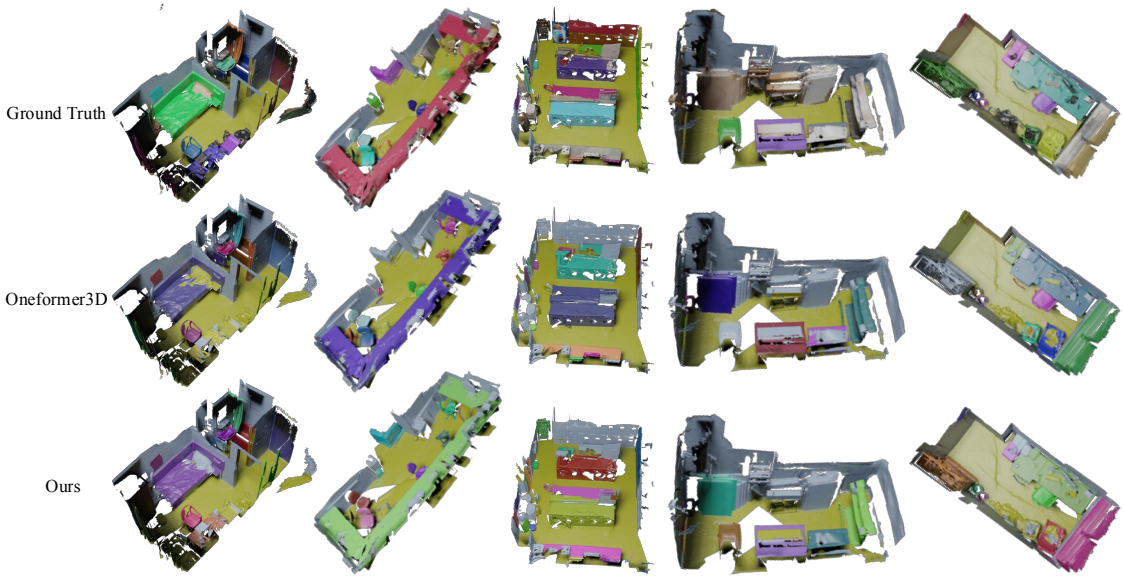

Figure 2. The qualitative results of the panoptic segmentation on ScanNet.

1. Qualitative Results

Fig.1 presents qualitative comparisons of instance segmentation results on ScanNet, where VDG-Uni3DSeg is evaluated against the baseline OneFormer3D[1]. Our approach exhibits a strong ability to distinguish individual object instances, particularly excelling in cluttered environments with complex spatial arrangements. By leveraging LLM-generated descriptions and reference images, our method enhances class-specific feature learning, resulting in more consistent instance segmentation.

Fig. 2 further showcases the panoptic segmentation results on ScanNet. VDG-Uni3DSeg produces more coherent and consistent segmentation across both stuff and thing categories. Notably, our approach significantly enhances the consistency of stuff regions, effectively reducing misclassification in large homogeneous areas while preserving precise instance differentiation. These qualitative results underscore the robustness of our method in complex 3D scenes.

2. Ablation Study

About Spatial Enhancement Module (SEM) Full attention over all points is prohibitively expensive, especially for large-scale point clouds. To address this, our SEM adopts a fixed-size subset sampling strategy, which significantly reduces computational cost while still allowing each point to access contextual information. We present results on ScanNet using different sampling sizes in Table R1. Increasing the size leads to improvements. But larger sizes will cause increased computational cost. These results, together with the ablation study, validate that the module is both effective and necessary.

Table R1. Effect of different sample size on ScanNet.

size	mAP_{50}	mAP	$mPrec_{50}$	$mRec_{50}$	mIoU	PQ
16	77.0	56.9	85.3	76.5	75.0	70.1
64	77.0	57.9	**86.0**	76.1	75.9	70.8
128	**78.5**	**59.3**	85.7	**78.3**	**76.2**	**71.5**

About internet images We collect 20 images per class and select the top-5 most relevant ones based on their CLIP similarity with corresponding class name. We also test with 2 random image sets on S3DIS. As shown in Table R2, CLIP-selected images perform better overall, which shows that high-quality images are more beneficial, especially those paired multi-view images. The internet images offers a lightweight and easily accessible alternative, in contrast to methods [2–4] that use paired multi-view images. Our approach is the first to leverage a small number of unpaired general images and textual descriptions for enhancing 3D point cloud segmentation. We agree, however, that bridging the remaining gap to models that leverage paired multi-view inputs remains a challenging problem. In future work, we plan to explore stronger image–point-cloud alignment techniques, enabling more effective use of easily collected general images to enhance 3D scene understanding, including in open-vocabulary settings.

Table R2. Comparision with random images on S3DIS.

Method	mAP_{50}	mAP	$mPrec_{50}$	$mRec_{50}$	mIoU	PQ
rand1	72.5	58.5	79.3	72.9	**71.7**	62.0
rand2	72.1	57.6	**81.9**	**75.2**	71.3	63.5
Ours	**74.1**	**60.1**	81.0	73.9	71.5	**66.3**

References

[1] Maxim Kolodiazhnyi, Anna Vorontsova, Anton Konushin, and Danila Rukhovich. Oneformer3d: One transformer for unified point cloud segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20943–20953, 2024. 2

[2] Abhijit Kundu, Xiaoqi Yin, Alireza Fathi, David Ross, Brian Brewington, Thomas Funkhouser, and Caroline Pantofaru. Virtual multi-view fusion for 3d semantic segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*, pages 518–535. Springer, 2020. 2

[3] 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.

[4] Damien Robert, Bruno Vallet, and Loic Landrieu. Learning multi-view aggregation in the wild for large-scale 3d semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5575–5584, 2022. 2