Supplementary Material for LASS3D: Language-Assisted Semi-Supervised 3D Semantic Segmentation with Progressive Unreliable Data Exploitation

Jianan Li[1,2] and Qiulei Dong *[1,2,3]

[1] School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences
[3] Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences
lijianan211@mails.ucas.ac.cn, qldong@nlpr.ac.cn

As introduced in the main paper, we propose a method named LASS3D for handling the semi-supervised 3D semantic segmentation task. In the supplementary material, we provide some details on the language-vision models used in LASS3D and additional experimental results.

1 Preliminaries

In this section, we give a brief introduction to the language-vision models (LVM) used in LASS3D, including KOSMOS-2 [5], GroundedSAM [7], and CLIP [6].

KOSMOS-2 is a transformer-based multi-modal large language model, which is trained through the next-word prediction task. It can understand multi-modal input, follow instructions, perceive object descriptions (*e.g.*, bounding boxes), and ground language to the visual world. We show some examples of the image-level captions generated by KOSMOS-2 in Fig. 1. As seen from this figure, KOSMOS-2 can capture most of the major objects in the given images. Thus, we use KOSMOS-2 to generate image-level captions in LASS3D.

GroundedSAM combines the open-set object detector GroundingDINO [4] and the open-world segmentation model SAM [3]. It can effectively tackle the open-set segmentation challenge by dividing it into two main components: open-set detection, and prompt segmentation. We show some examples of the segmentation results output by GroundedSAM in Fig. 2. As seen from this figure, GroundedSAM can segment the major entities in the given images and classify the segmented entities. Thus, we use GroundedSAM to produce the entity-level captions in LASS3D.

CLIP is trained on massive web-crawled image-text datasets. It consists of modality-specific (*i.e.*, image and text) encoders that produce embeddings for each modality, and the corresponding embeddings of each modality are aligned by minimizing the contrastive loss. In LASS3D, the text encoder of CLIP is

* Corresponding author

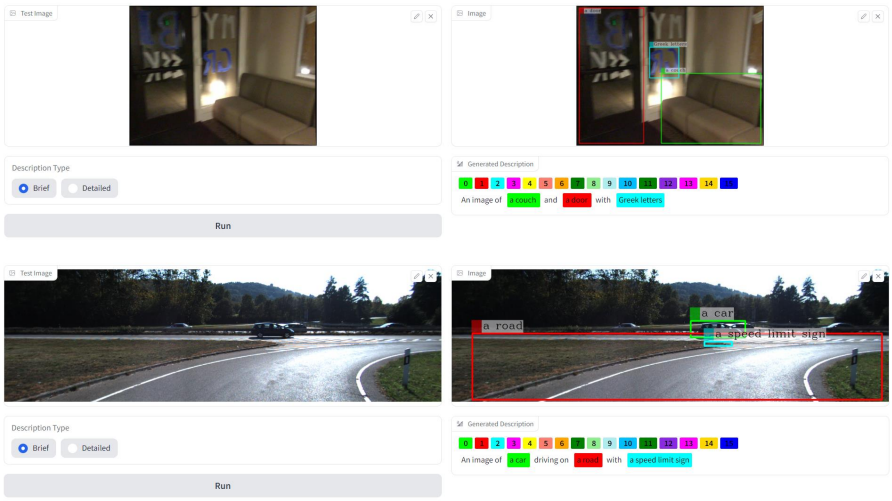

Fig. 1: Examples of the image-level captions generated by the demo of KOSMOS-2 [5]. The images are from the ScanNet [2] (top) and SemanticKITTI [1] (bottom) datasets.

Fig. 2: Examples of the segmentation results output by GroundedSAM [7]. The images are from the ScanNet [2] (left) and SemanticKITTI [1] (right) datasets.

leveraged to extract text embeddings from the image-level and entity-level captions generated by KOSMOS-2 and GroundedSAM, which are then used to inject semantic information into 3D features.

2 Additional Experiments

As mentioned in Sec. 1, we use KOSMOS-2 [5] and GroundedSAM [7] to generate image-level and entity-level captions respectively. Here, we replace GroundedSAM with GRiT [8] to generate the entity-level captions. GRiT can localize

Table 1: Ablation of the language-vision models for generating entity-level captions.

Model	1%	10%	20%	50%
GRiT [8]	51.5	57.8	58.0	60.1
GroundedSAM [7]	**54.7**	**62.8**	**63.1**	**63.5**

all the presented objects in the given images and generate free-form descriptions for each of them. The corresponding results on SemanticKITTI [1] are reported in Tab. 1. As seen from this table, the model trained with the captions generated by GroundedSAM achieves better results. Probably because GroundedSAM is a segmentation model, which can provide pixel-text pairs. While GRiT is a detection model, which only provides region-text pairs. Considering that we use the images as the bridge to connect the text data and point clouds, pixel-text pairs can construct more accurate text-point pairs. In addition, we show some detection results output by GRiT in Fig. 3. As seen from this figure, GRiT may output many wrong predictions, which may confuse the model training and worsen the segmentation performance. Thus, we use GroundedSAM to produce entity-level captions, which can provide more accurate and fine-grained captions.

In addition, we conduct experiments with 5% and 40% labeled data on SemanticKITTI. The results reported in Tab. 2 show that LASS3D still outperforms LiM3D.

Table 2: Results on SemanticKITTI with more labeled ratios.

	5%	40%
LiM3D	59.5	63.3
LASS3D	**61.5**	**64.2**

References

1. Behley, J., Garbade, M., Milioto, A., Quenzel, J., Behnke, S., Stachniss, C., Gall, J.: Semantickitti: A dataset for semantic scene understanding of lidar sequences. In: ICCV. pp. 9297–9307 (2019)
2. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR. pp. 5828–5839 (2017)
3. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: ICCV. pp. 4015–4026 (2023)
4. Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al.: Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 (2023)

5. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 (2023)
6. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021)
7. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., Zeng, Z., Zhang, H., Li, F., Yang, J., Li, H., Jiang, Q., Zhang, L.: Grounded sam: Assembling open-world models for diverse visual tasks (2024)
8. Wu, J., Wang, J., Yang, Z., Gan, Z., Liu, Z., Yuan, J., Wang, L.: Grit: A generative region-to-text transformer for object understanding. arXiv preprint arXiv:2212.00280 (2022)

Fig. 3: Examples of the detection results output by GRiT [8]. The images are from the ScanNet [2] (the first two rows) and SemanticKITTI [1] (the last row) datasets.