Identity-aware Language Gaussian Splatting for Open-vocabulary 3D Semantic Segmentation

Supplementary Material

1. Introduction

In this supplementary material, we provide additional explanations of the proposed method. First, details of the data preprocessing and the training procedure are described in Section 2. Further ablation study on the progressive mask expanding scheme is presented in Section 3. Finally, additional qualitative results with the discussion are presented in Section 4.

2. Implementation Details

2.1. Data Preprocessing

Following the previous method [6], we use SAM [4] to segment objects in images and CLIP to extract language features for each segmented region. To reduce memory costs, we use an autoencoder to compress high-dimensional CLIP features into 3 dimensions. CLIP features derived from each segment mask are assigned to their corresponding pixels. This process results in a CLIP feature map with dimensions (3,H,W). Additionally, we use DEVA [1] to generate coherent identity labels. DEVA is a zero-shot tracker that ensures consistent identity labels for objects across multiple views. These processed data are used to train language and identity embeddings.

2.2. Training

Implementation details of end-to-end training. Our method is based on the official implementations of 3DGS [2] and Gaussian Grouping [8]. Specifically, we augment each Gaussian with both language and identity embeddings. While we use the same type of identity embeddings in [8], the proposed method utilizes them differently to maintain multi-view consistency of language features. In contrast to previous approaches [6, 7] that pre-train a 3DGS [2] algorithm and then incorporate language embeddings in a separate stage, we design an end-to-end framework where Gaussian attributes, language embeddings, and identity embeddings are jointly trained. Our setting for learning rates is shown in Table 1. Both language and identity embeddings are rasterized through a differentiable rasterizer. Rasterized identity features pass through MLP, which applies the softmax function to produce per-class probability values. Identity embeddings are trained by using the cross-entropy between per-class probability values and coherent identity labels. Language embeddings are trained by using L1 loss between rasterized language feature maps and CLIP features. In subsection 3.2 of the manuscript, we set $M = 800$ and $N = 5$ for the identity-aware semantic consistency loss. Here, M represents the total number of Gaussian components selected to model the identity distribution. N denotes the number of pair-wise comparisons performed for each Gaussian.

Parameter	Learning Rate
Position	1.6×10^{-4}
Opacity	5.0×10^{-2}
Scaling	5.0×10^{-3}
Rotation	1.0×10^{-3}
Identity embedding	2.5×10^{-3}
Language embedding	2.5×10^{-3}

Table 1. Learning rates for Gaussian attributes and each embedding.

Effect of outlier filtering. In the proposed method, we apply the outlier filtering scheme, which stabilizes the optimization process of identity embeddings by providing reliable identity labels. During training, the outlier filtering scheme automatically excludes any incorrect identity labels that occur in certain views. As a result, identity embeddings can be trained with labels accurately assigned from other views. As shown in Fig. 1, the initial rasterized identity map contains noise, which gradually decreases as iterations progress.

Similarity margin	mIoU (%)	Boundary IoU (%)
5%	74.4	69.2
10%	**80.5**	**76.0**
15%	75.0	70.6
20%	74.2	69.6
25%	66.8	61.6

Table 2. Performance analysis of the proposed method based on changes in progressive mask expanding on the LERF dataset (the best result are shown in bold).

3. Ablation study

Progressive mask expanding. In subsection 3.3 of the manuscript, we describe a progressive mask expansion approach, where each neighboring segment is added when its cosine similarity differs from the similarity of the seed segment by less than 10% of the seed value. We compare the similarity margin of {5%, 10%, 15%, 20%, 25%} on the

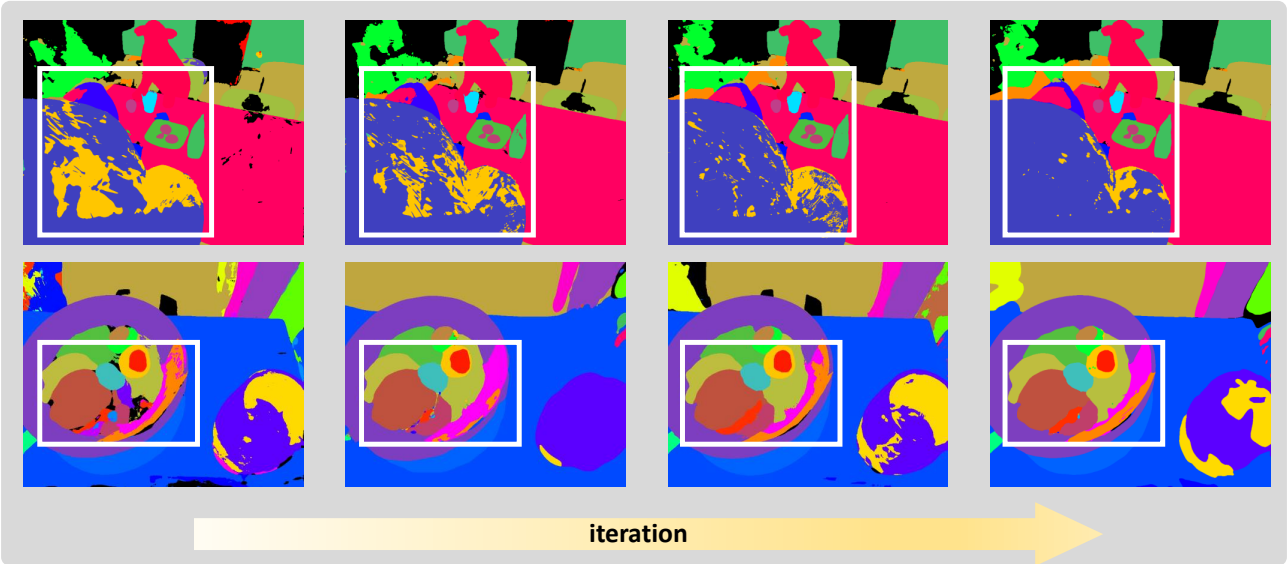

Figure 1. Visualization of rasterized identity feature map during training.

LERF [3] dataset and evaluate both mIoU and boundary IoU. As shown in Table 2, the 10% threshold achieves the best overall performance on the LERF [3] dataset. Therefore, we adopt 10% as our default similarity margin.

4. More Results and Limitations

4.1. Qualitative Results

Additional examples of open-vocabulary 3D semantic segmentation are shown in Figs. 2 and 3. As can be seen, the proposed method successfully generates semantic masks in novel views. Note that our method consistently produces accurate segmentation results in both simple and complex scenarios.

4.2. Limitations

Our method leverages coherent identity labels generated by a zero-shot tracker, which may introduce inaccuracies of labels in complex scenes. Although the outlier filtering scheme helps mitigate this issue, the quality of identity tracking impacts overall performance. The current implementation focuses on static scenes, thus it may face challenges in dynamic environments where objects move or change appearance over time. The extension of our approach to handle temporal consistency in dynamic scenes remains for future work.

References

[1] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking Anything with Decoupled Video Segmentation. In *Proc. Int. Conf. Comput. Vis.*, pages 1316–1326, 2023.

[2] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4):1–14, 2023.

[3] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. LERF: Language Embedded Radiance Fields. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 19729–19739, 2023.

[4] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment Anything. In *Proc. Int. Conf. Comput. Vis.*, pages 4015–4026, 2023.

[5] Kunhao Liu, Fangneng Zhan, Jiahui Zhang, MUYU XU, Yingchen Yu, Abdulmotaleb El Saddik, Christian Theobalt, Eric Xing, and Shijian Lu. Weakly Supervised 3D Open-vocabulary Segmentation. *Adv. Neural Inform. Process. Syst.*, 36:53433–53456, 2023.

[6] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3D Language Gaussian splatting. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, pages 20051–20060, 2024.

[7] Yansong Qu, Shaohui Dai, Xinyang Li, Jianghang Lin, Liujuan Cao, Shengchuan Zhang, and Rongrong Ji. GOI: Find 3D Gaussians of Interest with an Optimizable Open-vocabulary Semantic-space Hyperplane. In *Proc. ACM Int. Conf. Multimedia*, pages 5328–5337, 2024.

[8] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian Grouping: Segment and Edit Anything in 3D Scenes. In *Proc. Eur. Conf. Comput. Vis.*, pages 162–179, 2023.

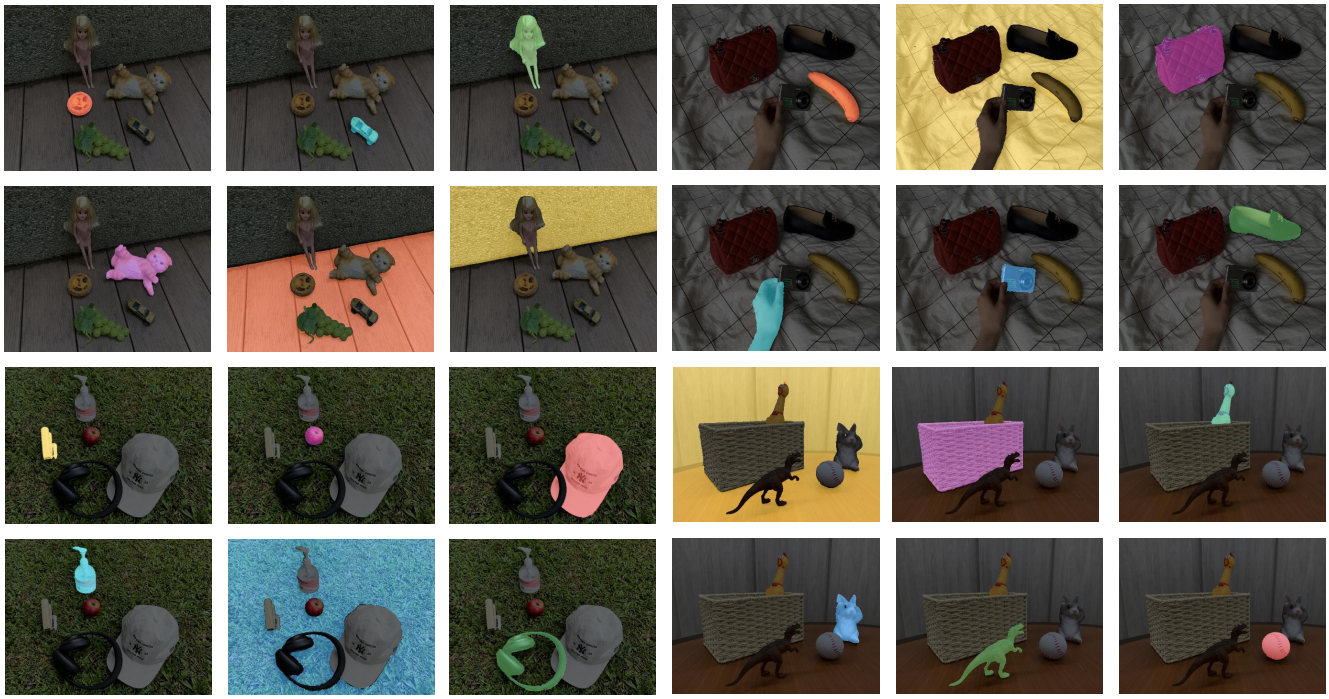

Figure 2. More results of open-vocabulary 3D semantic segmentation by the proposed method on the 3D-OVS [5] dataset.

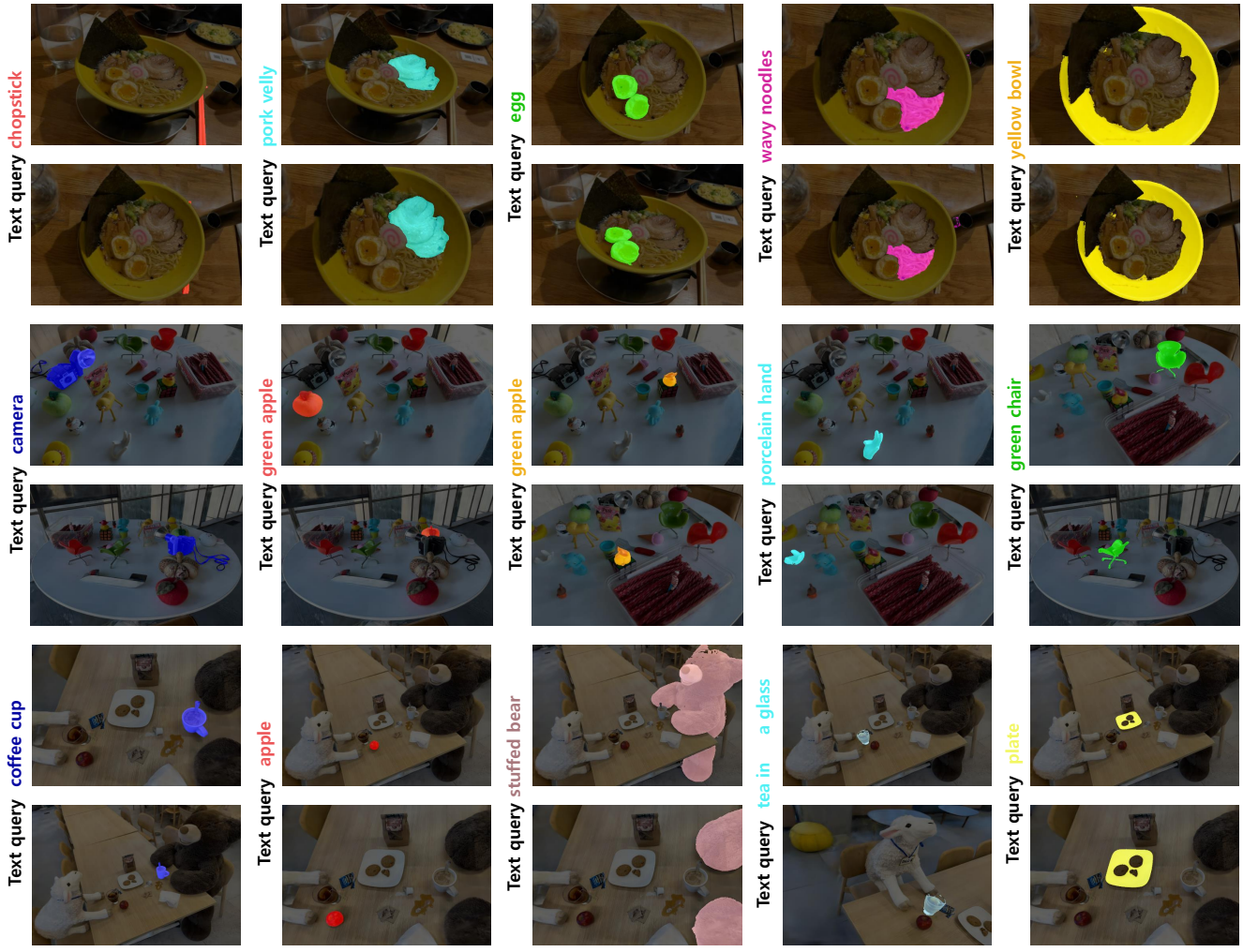

Figure 3. More results of open-vocabulary 3D semantic segmentation by the proposed method on the LERF [3] dataset.