3D Gaussian Splatting With Semantic Information Based on Semantic Class Labels

Masaya Takabe
Graduate School of FSE,
Waseda University
Tokyo, Japan
masaya.ta@asagi.waseda.jp

Taiju Watanabe
Graduate School of FSE,
Waseda University
Tokyo, Japan
lvpurin@fuji.waseda.jp

Hiroshi Watanabe
Graduate School of FSE,
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—3D Gaussian Splatting has recently emerged as a promising method for real-time 3D reconstruction. However, it does not leverage semantic information, limiting its utility in tasks such as editing and segmentation. We propose an efficient approach that incorporates semantic class labels directly into each 3D Gaussian by learning from segmentation results. This reduces memory and training time compared to prior methods relying on high-dimensional feature maps. Experiments show that our method maintains competitive rendering and segmentation performance while greatly reducing computational complexity.**

***Index Terms*—3D gaussian splatting, 3D scene representation, distillation, semantic segmentation.**

I. Introduction

Recently, 3D scene representation has advanced rapidly in computer vision and graphics, supporting applications such as AR and VR. Although NeRF [1], which implicitly represents the 3D scene, achieves photorealistic results by optimizing a neural radiance field, it requires dozens of hours of training. In contrast, 3D Gaussian Splatting [2] represents scenes explicitly with 3D Gaussians, enabling faster rendering. It provides a good trade-off between quality and rendering speed, making it suitable for real-time applications. Moreover, adding semantic information is crucial for applications such as scene editing and object-aware rendering, and also enables high-level understanding of 3D scenes for downstream tasks. Feature 3D Gaussian Splatting [3] supports these tasks but stores large feature maps per Gaussian, increasing memory consumption and training time. To address this, we propose learning class labels directly from segmentation results, offering an efficient solution.

II. Related Works

A. 3D Gaussian Splatting

3D Gaussian Splatting (3DGS) [2] initializes 3D Gaussian positions based on point clouds obtained by Structure from Motion (SfM) and renders 2D images using parameters such as position p, rotation q, scaling s, opacity α, and spherical harmonics h. To project the 3D Gaussians into images, the pixel RGB values are calculated by volumetric rendering. The parameters are optimized based on a loss between the rendered image and the ground truth image. Adaptive Density Control is used during training to clone or split Gaussians, improving detail representation while maintaining computational efficiency.

B. Semantic Segmentation on 3D Scene

In Feature 3D Gaussian Splatting [3], each 3D Gaussian is augmented with a feature map f to enable segmentation of the reconstructed 3D scene. The feature maps are rendered into 2D space through volumetric rendering, along with RGB images. Ground-truth feature maps, extracted from the intermediate layers of a segmentation model such as LSeg [4] or SAM [5], are used to calculate a loss between rendered feature maps. Although this method achieves rich semantic representations, it suffers from increased memory usage and longer training time due to storing high-dimensional feature maps for each 3D Gaussian.

III. Proposed Method

To reduce training time and memory usage, this study proposes a more intuitive method that introduces a class label

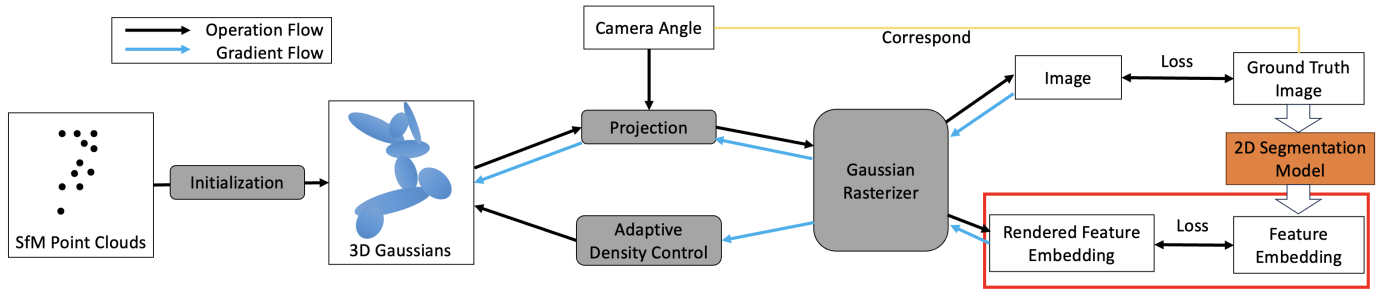

Fig. 1. Training process of the proposed method. In the red box, the model learns class labels from segmentation results, allowing each 3D Gaussian to be classified into a specific class.

979-8-3315-2415-9/25/$31.00 ©2025 IEEE

TABLE I
QUANTITATIVE EVALUATION OF RENDERED IMAGES

Method	Deep Blending			Tanks&Temples		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
3DGS	**29.488**	**0.899**	**0.247**	**23.637**	**0.845**	**0.175**
F-3DGS	29.282	0.892	**0.247**	23.587	0.844	0.180
Ours	29.345	0.893	0.248	23.559	0.843	0.180

Fig. 2. Reconstruction images. From left to right: 3DGS, Feature 3D Gaussian Splatting, and ours.

l as a new parameter for each 3D Gaussian. These labels are learned from the ground truth labels output by segmentation models such as LSeg [4] or SAM [5], allowing each Gaussian to be classified into a specific class.

The label L_j for each pixel is calculated as:

$$L_j = \sum_{i \in \mathcal{N}} l_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \tag{1}$$

where $\mathcal{N}$ is the set of 3D Gaussians within pixel j.

The loss function combines photorealistic reconstruction loss and a label loss based on cross entropy, and is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{rgb}} + \lambda \mathcal{L}_{\text{label}}, \tag{2}$$

where $\mathcal{L}_{\text{rgb}}$ and $\mathcal{L}_{\text{label}}$ denote the reconstruction loss for rendered RGB images and the cross entropy loss for class labels. λ is the hyperparameter for balancing losses. In our experiments, λ is set to 0.05. This formulation allows efficient optimization through backpropagation, enabling high-efficiency learning with reduced computational cost. The structure of our proposed method is shown in Fig. 1.

IV. EXPERIMENTS

The model is trained and evaluated using four scenes from the Tanks&Temples [6] and Deep Blending datasets [7], covering both indoor and outdoor environments. Ground truth labels are obtained using the LSeg. The shape of the label is $n \times H \times W$ (the number of classes n, height H, width W).

Unlike conventional Feature 3D Gaussian Splatting, which relies on different intermediate features from each segmentation model, the proposed method uses a generic class label format, relaxing model dependency and enabling flexible and efficient segmentation on 3DGS.

We show the quantitative results using PSNR, SSIM and LPIPS in Table I and visual comparisons in Fig. 2. The results indicate that the 3DGS achieves the highest quality across most metrics on both datasets. However, our approach also produces high-quality images comparable to those generated by 3DGS and Feature 3D Gaussian Splatting.

The quantitative results of the segmentation performance evaluated with mIoU and accuracy are shown in Table II, and visual comparisons are shown in Fig. 3. Although our method

TABLE II
QUANTITATIVE EVALUATION OF SEGMENTATION AND TRAINING TIME

Method	Deep Blending			Tanks&Temples		
	mIoU↑	Accuracy↑	Time↓	mIoU↑	Accuracy↑	Time↓
F-3DGS	0.213	**0.735**	34:43:56	0.409	**0.892**	22:28:34
Ours	**0.281**	0.731	**18:49:08**	**0.426**	0.864	**09:35:46**

Fig. 3. Results of segmentation on reconstructed 3D scenes. From left to right: Ground Truth image, Feature 3D Gaussian Splatting, and ours.

shows slightly lower accuracy than Feature 3D Gaussian Splatting on both datasets, it achieves higher mIoU. Moreover, our method reduces the training time by approximately half. Since our method directly learns class labels as parameters for each 3D Gaussian, it can sometimes assign different labels to the same object when the object belongs to multiple possible classes, and the segmentation model cannot resolve the ambiguity. These label inconsistencies inherited from the segmentation model degrade the accuracy of the learned semantics.

V. CONCLUSION

This paper proposes a lightweight method for incorporating semantic information into 3DGS by directly learning class labels from segmentation results. The experimental results demonstrated that the proposed method maintains reconstruction and segmentation performances comparable to conventional methods while achieving approximately half the training time. These results indicate that the proposed method enhances the editability and scene understanding capabilities of 3DGS, making it a promising approach for real-time 3D scene representation.

REFERENCES

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," Communications of the ACM, vol. 65, no. 1, pp. 99–106, 2021.

[2] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering." ACM Trans. Graph., vol. 42, no. 4, pp. 139–1, 2023.

[3] S. Zhou, H. Chang, S. Jiang, Z. Fan, Z. Zhu, D. Xu, P. Chari, S. You, Z. Wang, and A. Kadambi, "Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21676–21685, 2024.

[4] B. Li, K. Q. Weinberger, S. Belongie, V. Koltun, and R. Ranftl, "Language-driven semantic segmentation," arXiv preprint arXiv:2201.03546, 2022.

[5] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo et al., "Segment anything," in Proceedings of the IEEE/CVF international conference on computer vision, pp. 4015–4026, 2023.

[6] A. Knapitsch, J. Park, Q.-Y. Zhou, and V. Koltun, "Tanks and temples: Benchmarking large-scale scene reconstruction," ACM Transactions on Graphics (ToG), vol. 36, no. 4, pp. 1–13, 2017.

[7] P. Hedman, J. Philip, T. Price, J.-M. Frahm, G. Drettakis, and G. Brostow, "Deep blending for free-viewpoint image-based rendering," ACM Transactions on Graphics (ToG), vol. 37, no. 6, pp. 1–15, 2018.