Joint Semantic and Rendering Enhancements in 3D Gaussian Modeling with Anisotropic Local Encoding

Supplementary Material

Jingming He[1] Chongyi Li[2] Shiqi Wang[1] Sam Kwong[3]
[1]City University of Hong Kong, Hong Kong SAR, China
[2]Nankai University, Tianjin, China [3]Lingnan University, Hong Kong SAR, China
jingmhe3-c@my.cityu.edu.hk lichongyi@nankai.edu.cn shiqwang@cityu.edu.hk samkwong@ln.edu.hk

In this supplementary material, we provide further details of the method and results. Specifically, in Sec. 1, we introduce the concepts and rendering preocess of 3D Semantic Gaussian Splatting. Sec. 2 introduces the details on frustum-based sampling methods used in constructing Anisotropic 3D Gaussian Chebyshev descriptors. Sec.3 provides additional ablation studies on the ScanNet and Replica datasets. Sec.4 reports open vocabulary segmentation results on the LERF-Mask dataset. Sec. 5 displays more experimental setups and parameter configurations across our used datasets. Sec. 6 shows results on a per-scene basis for the Replica and ScanNet datasets.

1. Preliminaries on 3D Semantic Gaussian Splatting

In 3D Gaussian Splatting (3DGS) [6], the scene is modeled as a collection of anisotropic 3D Gaussians. Each Gaussian is parameterized by a center $\boldsymbol{\mu} \in \mathbb{R}^3$, an anisotropic covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$, a color vector $\mathbf{c} \in \mathbb{R}^3$, and an opacity value $\alpha \in [0, 1]$. To ensure that $\boldsymbol{\Sigma}$ remains positive semi-definite during optimization, it is factorized as:

$$\boldsymbol{\Sigma} = \mathbf{R}\,\mathbf{S}\,\mathbf{S}^\top\,\mathbf{R}^\top, \tag{1}$$

where $\mathbf{R} \in \mathbb{R}^{3\times3}$ is a rotation matrix (often represented by a unit quaternion) and $\mathbf{S} \in \mathbb{R}^{3\times3}$ is a diagonal scaling matrix whose entries describe the standard deviations along the principal axes. The density function of a Gaussian at a point $\mathbf{x} \in \mathbb{R}^3$ is then given by

$$G(\mathbf{x}) = \exp\!\big(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\big). \tag{2}$$

For rendering, the Gaussians are projected onto the image plane. Given a camera projection matrix $\mathbf{P}$ and an approximated Jacobian $\mathbf{J} \in \mathbb{R}^{2\times3}$ at $\boldsymbol{\mu}$, the projected covariance is computed as:

$$\boldsymbol{\Sigma}' = \mathbf{J}\,\boldsymbol{\Sigma}\,\mathbf{J}^\top, \tag{3}$$

and the projected center as $\boldsymbol{\mu}' = \mathbf{P}[\boldsymbol{\mu}^\top, 1]^\top$. The color of a pixel at position $\mathbf{u}$ is then obtained by alpha blending the contributions of all Gaussians sorted along the ray:

$$C(\mathbf{u}) = \sum_{i=1}^{N} \mathbf{c}_i\,\alpha_i\,T_i, T_i = \prod_{j=1}^{i-1}(1-\alpha_j), \tag{4}$$

where N is the number of Gaussians affecting the pixel. To enhance scene understanding and support downstream tasks such as semantic segmentation, recent works [4, 11, 13–15] propose augmenting each 3D Gaussian with an additional semantic feature vector $\mathbf{f} \in \mathbb{R}^D$, leading to an extended parameter set:

$$\Theta = \{\boldsymbol{\mu}, \mathbf{R}, \mathbf{S}, \alpha, \mathbf{c}, \mathbf{f}\}. \tag{5}$$

The rendered semantic feature map is computed analogously to color blending:

$$F_s(\mathbf{u}) = \sum_{i=1}^{N} \mathbf{f}_i\,\alpha_i\,T_i, \tag{6}$$

with T_i defined as above. The rendered feature $F_s(\mathbf{u})$ is then supervised through feature distillation methods [15], direct pixel-wise 2D supervision [13, 14], or contrastive learning with 2D segmentation masks [4].

2. Frustum-Based Sampling for Anisotropic 3D Gaussian Chebyshev Descriptors

To reduce computational load and ensure that subsequent local shape encoding is guided by view-specific semantic cues, we restrict the processing spatial domain to the camera's current viewing frustum. Let the camera be characterised by its position $\mathbf{o} \in \mathbb{R}^3$, orientation matrix $R_{\text{cam}} \in SO(3)$ (with the view axis defined as $v_{\text{cam}} = R_{\text{cam}}\mathbf{e}_z$, where $\mathbf{e}_z = [0, 0, 1]^\top$), field-of-view angle θ, and near/far distances $d_{\min}$ and $d_{\max}$. For any point $p \in \mathbb{R}^3$, set

Task	Segmentation (IoU)			Rendering (PSNR)		
Scene	Feature 3DGS [15]	Semantic Gaussians [5]	Ours	Feature 3DGS [15]	Semantic Gaussians [5]	Ours
scene0050_02	46.2	47.1	50.4	25.47	25.50	25.98
scene0144_01	67.7	72.0	73.9	27.95	27.91	28.31
scene0221_01	57.1	56.9	60.1	29.26	29.19	29.42
scene0300_01	58.1	60.7	62.7	25.45	25.34	25.96
scene0354_00	52.3	55.2	58.3	26.93	26.94	27.45
scene0389_00	55.0	57.9	61.2	28.12	28.08	28.67
scene0423_02	65.4	69.7	72.6	29.57	29.53	29.91
scene0427_00	68.9	71.5	74.0	28.68	28.53	28.83
scene0494_00	65.0	67.2	71.6	29.19	29.16	29.45
scene0616_00	57.8	57.0	61.8	19.04	18.92	19.66
scene0645_02	52.2	54.7	58.9	22.80	22.82	23.31
scene0693_00	68.2	71.8	75.2	31.25	31.20	31.44
Mean	59.5	61.8	**65.1**	26.98	26.93	**27.37**

Table 1. Detailed Results by Scene on the ScanNet Dataset.

Task	Segmentation (IoU)		Rendering (PSNR)	
Method	Feature3DGS [15]	Ours	Feature3DGS [15]	Ours
room0	82.0	84.2	35.18	35.57
room1	74.7	78.1	37.97	38.21
office3	82.3	84.5	37.74	38.07
office4	73.4	77.7	33.92	34.47
Mean	78.1	**81.1**	36.20	**36.58**

Table 2. Detailed Results by Scene on the Replica Dataset.

Method	mIoU	OA	PSNR	FPS
baseline	59.2	75.1	26.98	143
Ours (LEM w/o AGCD)	60.8	76.4	27.10	159
Ours (LEM w/o TLE)	63.1	78.3	27.29	179
Ours (AGP w/o SG)	64.0	79.3	27.37	149
Ours (AGP w/o AD)	62.9	77.8	26.93	196
Ours (ASHP w/o SG)	64.2	79.3	27.39	155
Ours	64.3	79.7	27.41	184
Ours (w/ CSKT)	65.1	80.5	27.46	191

Table 3. Ablation study results on the ScanNet dataset.

Method	mIoU	OA	PSNR	FPS
baseline	78.2	94.3	37.01	148
Ours (LEM w/o AGCD)	78.6	94.8	37.13	163
Ours (LEM w/o TLE)	80.1	95.3	37.28	186
Ours (AGP w/o SG)	80.8	95.8	37.32	153
Ours (AGP w/o AD)	79.8	95.0	36.92	195
Ours (ASHP w/o SG)	80.9	95.8	37.29	158
Ours	81.1	96.1	37.38	189

Table 4. Ablation study results on the Replica dataset.

Method	**mIoU**	**mBIoU**
DEVA [3]	52.43	49.47
LERF [7]	37.17	29.30
SA3D [1]	24.93	23.33
LangSplat [10]	57.57	53.60
Gaussian Grouping [9]	72.80	67.57
Ours	**75.67**	**71.81**

Table 5. Results of Segmentation on the LERF-Mask Dataset supervised by SAM.

$\mathbf{v} = p - \mathbf{o}$. The point p lies inside the view frustum if $d_{\min} \le \|\mathbf{v}\| \le d_{\max}$ and $\angle(\mathbf{v}, v_{\text{cam}}) \le \theta/2$ (equivalently, $\frac{\mathbf{v} \cdot v_{\text{cam}}}{\|\mathbf{v}\|\,\|v_{\text{cam}}\|} \ge \cos\frac{\theta}{2}$).

Alternatively, if the frustum is given by plane normals $\{n_k\}_{k=1}^{K}$ with offsets d_k, the condition becomes $n_k^{\top} p + d_k \ge 0,\ \forall k$. Discarding the points that violate these tests yields the clipped set $\mathcal{P}$. We then perform iterative farthest-point sampling to obtain a representative centre set $\mathcal{C}$.

For each centre $c^{(i)}$, we define its local neighbourhood as $N(c^{(i)}) = \{p \in \mathcal{P} \mid \|p - c^{(i)}\| \le r\}$, where $r > 0$ is a preset locality radius (independent of $d_{\min\,/\,\max}$).

3. Additional Ablation Results on ScanNet and Replica Datasets

This section reports additional ablation results on both the ScanNet and Replica datasets, as shown in Tables 3 and 4. The experiments include baseline comparisons and variants with individual modules removed. The results are consistent with those observed on the Deep Blending dataset: AGCD contributes significantly to semantic performance (mIoU, OA), while AGP and ASHP with semantic guidance improve rendering efficiency (FPS), with almost no loss in PSNR.

4. Open Vocabulary Segmentation on LERF-Mask Dataset

In this section, we further evaluate our method on the LERF-Mask dataset [13], which provides accurate pixel-level masks for three open-world scenes (figurines, ramen, teatime) originally from the LERF-Localization benchmark [7]. Following Gaussian Grouping [13], we supervise the semantic identity embeddings using SAM masks selected by Grounding DINO [9] with language prompts. As shown in Table 5, our method surpasses existing methods in both mIoU and mBIoU, validating our segmentation improvements.

5. Additional Experimental Details Across Datasets

This section offers more experimental setups for three datasets, including dataset configurations, optimizer settings, and hyperparameters.

The Replica Dataset: Following the experimental setup with previous [8, 15], we select four scenes: room 0, room 1, office 3, and office 4. For each scene, a trajectory generates eighty images, from which every eighth image starting with the third is chosen for experiments, and class = 7 for the mIoU metric. Same with [15], for room 1, the last 2 test images are excluded from the results since these images do not have 7 classes in the image. Optimization is performed using the Adam optimizer with a learning rate of 0.001 over 5000 iterations. In the anisotropic 3D Gaussian Chebyshev descriptor configuration, the maximum order of Chebyshev polynomials (D) is set to 4, the number of rotation angle sets (J) is 10, and the eigen-decomposition uses $K = 5$. The binarization threshold (τ) for Gaussian Pruning and Spherical Harmonics is set at 0.1.

The Deep Blending: Due to the broad variety of classes for open-vocabulary tasks in this dataset, we follow the setting of [12], which maps classes to 21 classes from the COCO dataset. Optimization is conducted using the Adam optimizer with a learning rate of 0.001, across 13,000 iterations. The D is set to 6, with 10 rotation angle sets (J), and $K = 8$. The τ is set at 0.1.

The ScanNet dataset: We conducted experiments on both open-vocabulary and closed-set segmentation. The supervision for open-vocabulary segmentation follows the same method as used with the Replica dataset, and maps the ScanNet classes into 21 classes from the COCO dataset. For closed-set segmentation, we used pseudo labels with cross-entropy loss for supervision, the same as the approach in [12], which employs 2D segmentation masks generated by Mask2Former [2]. During the assessment phase, evaluation metrics were calculated using real ground truth labels. Optimization is carried out using the Adam optimizer with a learning rate of 0.001 over 13,000 iterations (8,000 when CSKT is enabled). Given ScanNet's complex shapes and local structures, the maximum order of Chebyshev polynomials (D) is set to 6, the number of rotation angle sets (J) is 15, and the eigen-decomposition uses $K = 10$. These settings impact the descriptor's dimensionality and sensitivity to directions. The τ is set at 0.08.

We limit AGCD computations only during training within frustum regions, with no extra cost at inference time. The CSKT module reduces training iterations (e.g., ScanNet from 13k to 8k), offsetting the higher per-iteration cost (∼1.3×), ultimately decreasing total training time by 25%. The original 3DGS baselines achieved 26.58 PSNR (ScanNet) and 35.97 (Replica). Using AGP and ASHP, our method reduces the model size: Replica (204.5→48.2 MB), ScanNet (484.7→86.4 MB), Deep Blending (1601.3→154.9 MB).

6. Detailed Per-Scene Results for Replica and ScanNet Datasets

In this section, we present detailed results of our method's performance on a per-scene basis for both the Replica and ScanNet datasets, as shown in Table 1 and Table 2. Our method outperforms competing approaches across almost all scenes.

References

[1] Jiazhong Cen, Zanwei Zhou, Jiemin Fang, Wei Shen, Lingxi Xie, Dongsheng Jiang, Xiaopeng Zhang, Qi Tian, et al. Segment anything in 3d with nerfs. *Advances in Neural Information Processing Systems*, 36:25971–25990, 2023. 2

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 3

[3] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1316–1326, 2023. 2

[4] Seokhun Choi, Hyeonseop Song, Jaechul Kim, Taehyeong Kim, and Hoseok Do. Click-gaussian: Interactive segmentation to any 3d gaussians. In *European Conference on Computer Vision*, pages 289–305. Springer, 2024. 1

[5] Jun Guo, Xiaojian Ma, Yue Fan, Huaping Liu, and Qing Li. Semantic gaussians: Open-vocabulary scene understanding with 3d gaussian splatting. *arXiv preprint arXiv:2403.15624*, 2024. 2

[6] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time

radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[7] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 2, 3

[8] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in neural information processing systems*, 35:23311–23330, 2022. 3

[9] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 2, 3

[10] Minghan Qin, Wanhua Li, Jiawei Zhou, Haoqian Wang, and Hanspeter Pfister. Langsplat: 3d language gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20051–20060, 2024. 2

[11] Jin-Chuan Shi, Miao Wang, Hao-Bin Duan, and Shao-Hua Guan. Language embedded 3d gaussians for open-vocabulary scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5333–5343, 2024. 1

[12] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9043–9052, 2023. 3

[13] Mingqiao Ye, Martin Danelljan, Fisher Yu, and Lei Ke. Gaussian grouping: Segment and edit anything in 3d scenes. In *European Conference on Computer Vision*, pages 162–179. Springer, 2024. 1, 3

[14] Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugs: Holistic urban 3d scene understanding via gaussian splatting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21336–21345, 2024. 1

[15] Shijie Zhou, Haoran Chang, Sicheng Jiang, Zhiwen Fan, Zehao Zhu, Dejia Xu, Pradyumna Chari, Suya You, Zhangyang Wang, and Achuta 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*, pages 21676–21685, 2024. 1, 2, 3