Supplementary Material for "SOVGaussian: Sparse-View 3D Gaussian Splatting for Open-Vocabulary Scene Understanding"

Anonymous submission

More Implementation Details

Training Strategy

Our method is not based on separately training Gaussian geometric parameters and language features but rather employs an end-to-end training process. In the first 3000 iterations, we do not use LAD but apply global normalization to monocular depth cues to supervise the coarse geometric structure of the scene. In subsequent iterations, LAD is employed to focus depth loss on object scale, trusting the depth cue's intra-object ranking accuracy and addressing object scale errors to improve geometry. For the 3DOVS dataset, LOP is applied every 1000 iterations between the 3500th and 9500th iterations. For the DTU dataset, it is applied between the 2000th and 4000th iterations.

Scene	Text Query
scan21	plastic model building with green roof, plastic model building with red roof
scan31	pumpkin, metal can of baked beans, plastic packet of coffee, cardboard box of Lipton green tea, orange cardboard box of fruit tea
scan38	concrete block, wooden plank
scan41	metal bucket with painted patterns, concrete block
scan55	sculpture of rabbit
scan103	plush toy pig, concrete block
scan110	gloden statue

Table 1: Open-vocabulary text prompts for evaluating querying mIOU and localization accuracy on DTU dataset.

DTU Dataset

We select seven complex scenes from the DTU dataset for testing and evaluation. Akin to the annotation style of the 3DOVS dataset, our text queries for objects within these scenes include their categories and specific details, as shown in Table 1. This annotation style aligns more closely with human visual analysis habits, focusing on both object categories and feature details.

Method	PSNR↑	SSIM↑	LPIPS↓
DietNeRF	11.85	0.633	0.314
RegNeRF	18.89	0.745	0.190
MipNeRF	8.68	0.571	0.353
PixelNeRF	16.82	0.695	0.270
MVSNeRF	18.63	0.769	0.197
3DGS	10.99	0.585	0.313
DNGaussian	18.91	0.790	0.176
Ours	**19.50**	**0.791**	**0.171**

Table 2: Quantitative results on the DTU dataset with cross-task methods. For fairness, our method and the baselines use the same resolution of 400 $\times$ 300, trained with the same 3 views, and evaluated on the same test views. It is noteworthy that all results of these methods are obtained through publicly accessible papers or codes.

More Quantitative Results

Our SOVGaussian possesses the capability to create superior geometry for the language field, an advantage that is reflected not only in open-vocabulary querying but also in novel view synthesis. Therefore, in this section, we also quantitatively compare our method with other state-of-the-art methods in terms of metrics for novel view synthesis.

Quantitative Results on DTU Dataset

We conduct comparisons with cross-task methods on the DTU dataset, and the quantitative results are presented in Table 2. We selected current state-of-the-art few-shot methods used solely for novel view synthesis tasks as baselines, which are DietNeRF (Jain, Tancik, and Abbeel 2021), RegNeRF (Niemeyer et al. 2022), MipNeRF (Barron et al. 2021), PixelNeRF (Yu et al. 2021), MVSNeRF (Chen et al. 2021), 3DGS (Kerbl et al. 2023), and DNGaussian (Li et al. 2024). Our observations indicate that our method achieves the best performance in terms of PSNR, SSIM, and LPIPS

Methods	sofa		bed		room		lawn		bench		**Overall**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
3DOVS	12.03	0.426	10.04	0.176	15.36	0.361	10.15	0.054	14.08	0.142	12.33	0.232
LEGaussians	13.47	0.411	14.05	0.340	17.35	0.406	8.74	0.071	14.46	0.244	13.61	0.295
LangSplat	13.89	0.488	13.31	0.374	16.91	0.482	8.47	0.061	13.70	0.178	13.26	0.316
Ours	**20.41**	**0.683**	**18.89**	**0.644**	**20.45**	**0.669**	**11.75**	**0.153**	**16.63**	**0.287**	**17.63**	**0.487**

Table 3: Performance of novel view synthesis on the 3DOVS dataset. We report the PSNR↑ and SSIM↑.

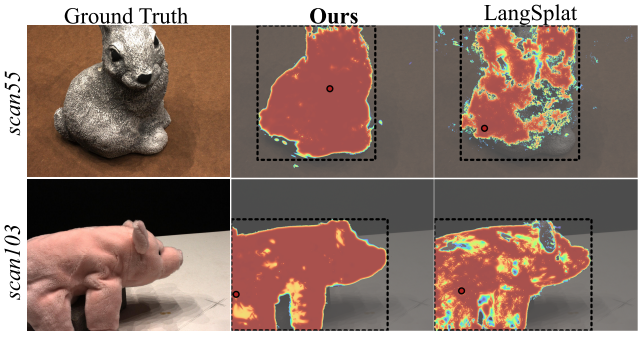

Figure 1: More qualitative comparisons on the DTU dataset. We visualized the relevance map of novel view for each scene.

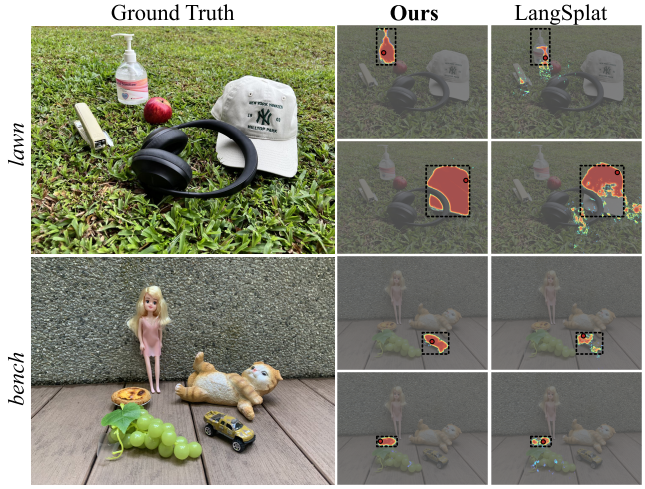

Figure 2: More qualitative comparisons on the 3DOVS dataset. Top to bottom: query words "hand soap", "New York Yankees cap", "mini offroad car", "Portuguese egg tart".

metrics. This demonstrates that our method effectively leverages the intrinsic correlation between the newly introduced language features and scene geometry, improving geometric reconstruction and thereby enhancing generalization to novel views synthesis.

Quantitative Results on 3DOVS Dataset

We present the quantitative analysis comparison of novel view synthesis on the 3DOVS dataset, as shown in Table 3. Compared to the current state-of-the-art approaches, our method produces significantly better synthesis results in novel views. This clearly demonstrates that the introduction of depth cues enhances scene geometry reconstruction, benefiting not only novel view language querying but also the synthesis of RGB images from novel views.

More Qualitative Results

Qualitative Results on DTU Dataset

In Figure 1, we present open-vocabulary querying results on additional scenes from the DTU dataset. It can be observed that although most of these scenes contain only a single main object, LangSplat fails to generate accurate relevancy maps, producing significant noise and voids instead. This highlights a challenging issue with the DTU dataset: whether the language field falls into suboptimal solutions when object material, lighting, and shadows are prominent features, leading to significantly reduced generalization in novel views. In contrast, our method accurately fits the shape of objects, providing complete bounding boxes, demonstrating its effectiveness in addressing the aforementioned issues.

Qualitative Results on 3DOVS Dataset

In Figure 2, we visualize more examples of novel view open-vocabulary querying on the 3DOVS dataset. We observed that for text queries like "hand soap" and "New York Yankees cap", LangSplat's results exhibit drift and fail to provide accurate boundaries, whereas our SOVGaussian achieves higher accuracy and precision, demonstrating its superiority.

References

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, 5855–5864.

Chen, A.; Xu, Z.; Zhao, F.; Zhang, X.; Xiang, F.; Yu, J.; and Su, H. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, 14124–14133.

Jain, A.; Tancik, M.; and Abbeel, P. 2021. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5885–5894.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Trans. Graph.*, 42(4): 139–1.

Li, J.; Zhang, J.; Bai, X.; Zheng, J.; Ning, X.; Zhou, J.; and Gu, L. 2024. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20775–20785.

Niemeyer, M.; Barron, J. T.; Mildenhall, B.; Sajjadi, M. S.; Geiger, A.; and Radwan, N. 2022. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5480–5490.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4578–4587.