Supplementary Material for Lifting by Gaussians: A Simple, Fast and Flexible Method for 3D Instance Segmentation

A. Appendix

In this appendix, we provide further experimental results, including additional 3D segmentation comparisons in Section B, a qualitative comparison with SAGA [1] on rendering of 2D masks at novel views through different scales in Section C.1 and qualitative results on the 3D-OVS [5] dataset in Section C.2. We further show that our method can be used to 3D segment 2DGS fields without modification in Section D. Finally, we show some intuition into our improvements of the Mini-Splatting importance sampling in Section F and conclude by showing an application of our method to lift 2D feature maps, such as DINO, into 3D in Section G.

B. Additional 3D results

We show additional results of extracted 3D objects from our LBG method in Figure 1. Contrastive methods like SAGA require a 3D feature clustering step to extract objects, which is prone to floaters and noise. Gaussian Grouping also requires a 3D clustering step which produces noisy 3D objects. Our 3D objects are more coherent and have cleaner boundaries than other methods due to our simple lifting and mask merging strategy.

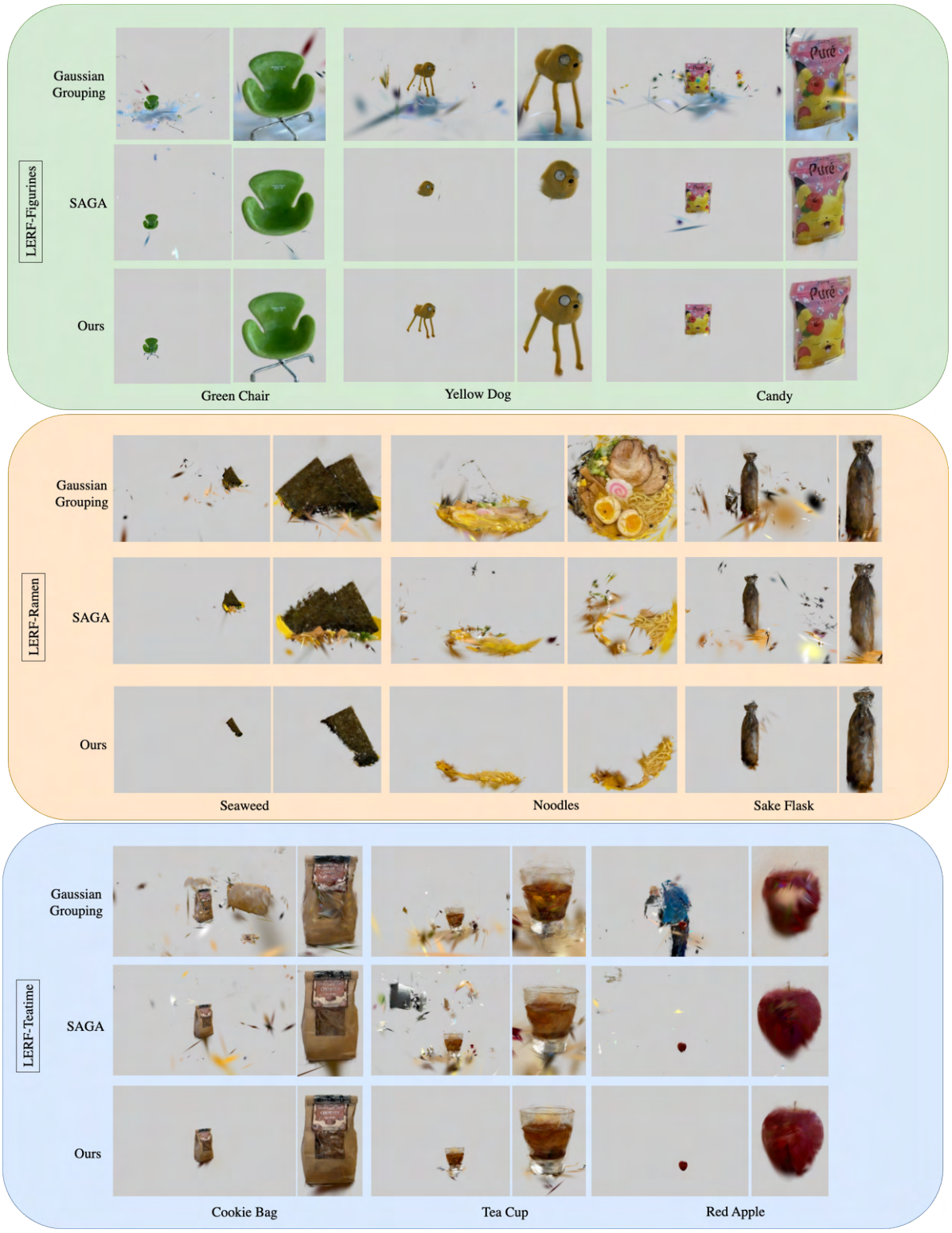

Figure 1. **Additional 3D segmentation results on LERF dataset**.

C. Additional 2D results

C.1. LERF

We show mask novel view synthesis results on the three LERF scenes in Figure 2. Specifically, we compare SAGA and LBG. For SAGA, we show images rendered at three levels: 0.1 (left), 0.5 (middle), and 1.0 (right). For our method, we show object level (left), part level (middle), and subpart level (right).

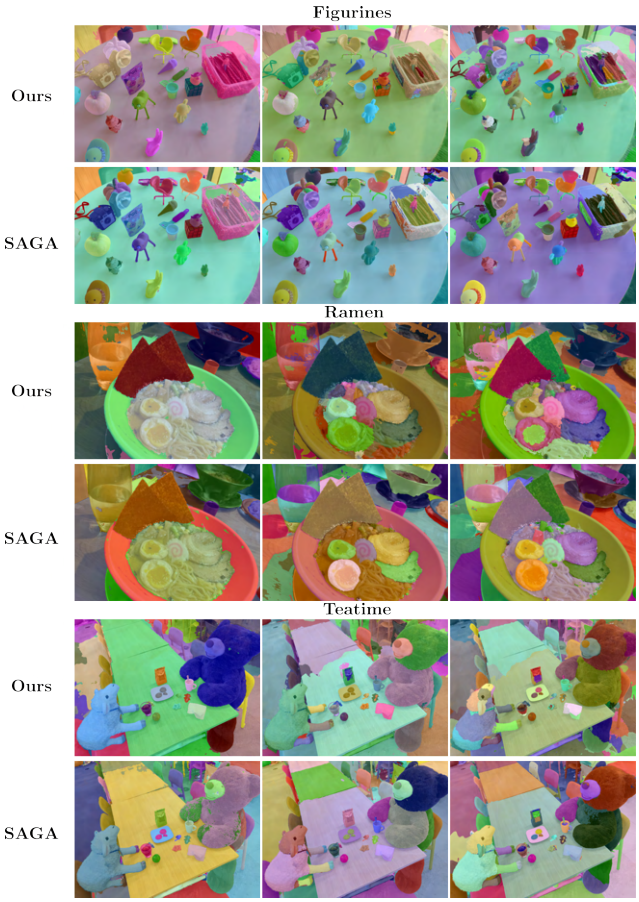

Figure 2. **Additional results on novel view synthesis for 2D instance masks**. For SAGA, we show images rendered at three levels: 0.1 (left), 0.5 (middle), and 1.0 (right). For our method, we show object level (left), part level (middle), and subpart level (right).

Even though our method is not optimized on the task of novel view synthesis for 2D masks, it performs well, especially on the object level. We can see that SAGA often breaks up objects into parts, even on the top level (camera in figurines, bear in teatime). This is largely due to SAGA using metric diagonal measurements to determine scale without associating these scales back to object/part/subpart decompositions. We argue that instead of using such arbitrary scales it is much more intuitive to break a scene into it's logical parts, starting from complete objects.

C.2. 3DOVS

We compare our method for segmentation against prior methods, such as LSEG [3] and OVSeg [4]. The numbers for these methods are taken from [7]. We found that LangSplat evaluations, as described in the paper, led to suboptimal performance due to limited contrast in the learned feature representation. To improve performance, we modified the protocol described in the paper and used a per-scene threshold.

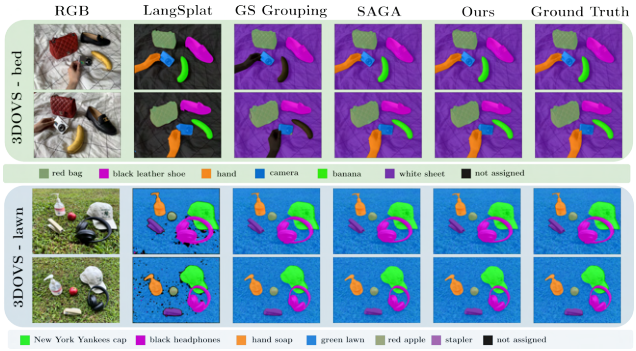

Figure 3. **Qualitative comparison on the 3DOVS dataset.** Black regions are unassigned. In the bed scene, Gaussian Grouping merges hand and banana objects together, resulting in segmentation failure. Similarly, LangSplat fails to segment the white sheet due to low contrast in the feature space. Our method shows cleaner boundaries compared to both baselines.

On the 3DOVS dataset (Fig. 3), our method demonstrates superior performance across the board against most methods and is comparable to SAGA. Notably, LangSplat overlooks the background in the bed scene and exhibits gaps in the lawn scene, attributed to inconsistencies in thresholds and noise within the *subpart* level of the language embeddings. While Gaussian Grouping yields results similar to our method, it often produces less defined boundaries due to tendencies towards over-segmentation. Regions in black are segmentations that were not detected during the evaluation.

D. Applying our method on 2DGS

As our method, LBG can consume any Gaussian Splatting-based reconstruction, we apply our method on a scene reconstructed using 2DGS [2] without any modification. As a consequence of using [2], we can produce meshes of the individual segmented objects. Results are shown in Figure 4.

E. Additional Ablations

We present additional ablation results using our method with Fast-SAM instead of the standard Segment Anything

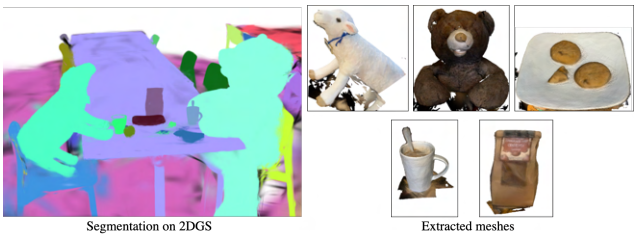

Figure 4. **LBG Segmentation on 2DGS**. 2DGS with colored Gaussians according to instance IDs (left) and individually extracted meshes (right).

Model for mask extraction. While the Fast-SAM model provides results in near real-time, which is desirable for most applications in robotics and AR, Figure 5 shows that the results are much worse. We can see that the segmentation lifted with Fast-SAM masks struggles to keep clean object boundaries. Furthermore, even with our post-processing step that merges adjoining clusters, we can see that the Fast-SAM model still contains many objects that cannot be merged using a purely geometric approach.

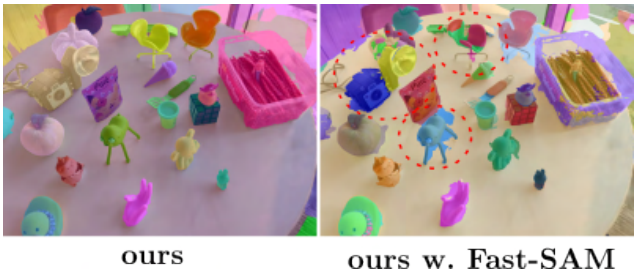

Figure 5. **Additional ablation results**. We show performance of our method using the standard SAM model (left) and Fast-SAM (right).

F. Improvements to Mini-Splatting

We adopted a technique similar to Mini-splatting to remove floaters from Gaussian Splatting reconstructions. 3D Gaussians are removed based on a probability dictated by an importance score. Initially, we found that using opacity contribution as the basis for this score was insufficient, as it assigned small values to floaters. Many floaters, we discovered, resulted from over-fitting to a single view (see Figure 6). To address this, we augmented the probability score by considering the number of views a 3D Gaussian maximally contributes to, through a log multiplier on the number of views. This modification, combined with the pruning and resampling strategy from Mini-splatting, effectively reduces floaters, particularly those caused by single-view over-fitting.

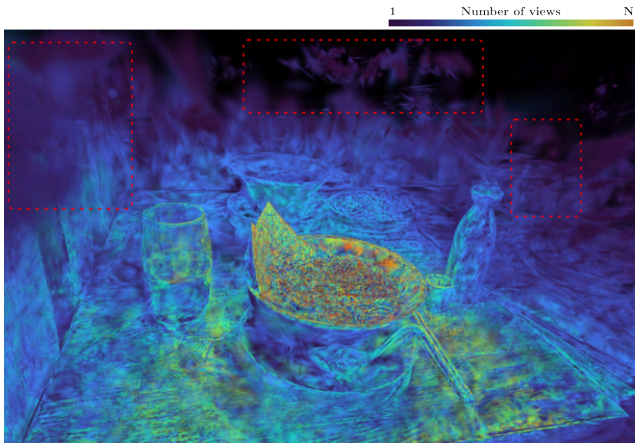

Figure 6. **Visualization of the number of views which see each Gaussian**. Notice how many structured floaters are only seen by a single view, showcasing visual artifacts from single-view over-fitting.

G. DINO Feature lifting

Our approach to lift 2D masks to 3D Gaussian Splatting fields can also lift any 2D foundation model features onto 3D Gaussians using the same strategy. Consequently, we lift DINOv2 [6] features onto a 3DGS field using our LBG approach. We visualize the first 24 PCA components of the features in Figure 7.

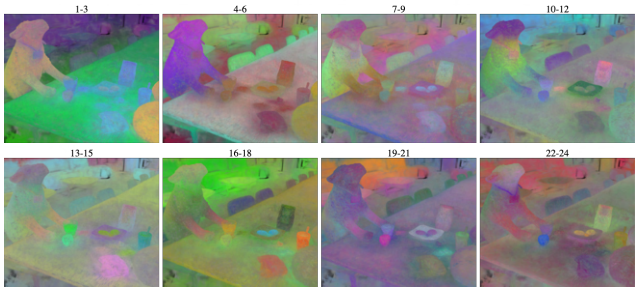

Figure 7. **Lifting DINOv2 features onto Gaussians**. Using our Lifting-by-Gaussians approach, we lift DINOv2 features and visualize the first 24 PCA components.

References

[1] Jiazhong Cen, Jiemin Fang, Chen Yang, Lingxi Xie, Xiaopeng Zhang, Wei Shen, and Qi Tian. Segment Any 3D Gaussians. *arXiv preprint arXiv:2312.00860*, 2023. 1

[2] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*, 2024. 3

[3] Boyi Li, Kilian Q. Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven Semantic Segmen-

tation. In *International Conference on Learning Representations*, 2022. 3

[4] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7061–7070, 2023. 3

[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. *Advances in Neural Information Processing Systems*, 36:53433–53456, 2023. 1

[6] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision, 2023. Publication Title: arXiv:2304.07193. 4

[7] 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. 3