VoteSplat: Hough Voting Gaussian Splatting for 3D Scene Understanding

Supplementary Material

1. Training Details

To prevent background Gaussian primitives from participating in the voting process, the transmission rate for color differs from that for voting. The transmission threshold T_i follows the original 3DGS and is set to 0.0001, while the voting transmission threshold $\hat{T}_i$ is set to 0.01. The learning rate for the offset vector Δp is 0.0001, while other Gaussian primitive parameters and learning rates remain consistent with the original 3DGS. The vote loss is designed to be independent of other Gaussian properties to avoid affecting appearance and structural integrity.

For depth distortion loss, instead of following 2DGS, which transforms voting points into NDC space before extracting depth, we directly obtain depth in camera space. The CUDA implementation of 3D vote blending, projection, and depth distortion ensures maximum training efficiency. Compared to the original 3DGS, the additional computational cost is negligible.

From a shooting's perspective, when scene instances are large, the vote loss weight is set higher, whereas for smaller instances, it is scaled down by an order of magnitude.

2. Feature Visualization of Gaussian Grouping

Figure 1 presents the feature visualization of Gaussian Grouping, showing that it fails to distinguish features between instances.

When recording the training time for Gaussian Grouping, $\mathcal{L}_{3d}$ was omitted, as it requires excessive GPU memory, exceeding the capacity of an NVIDIA RTX 3090, which leads to program crashes. The training time of Gaussian Grouping in Table 2 is obtained without $\mathcal{L}_{3d}$.

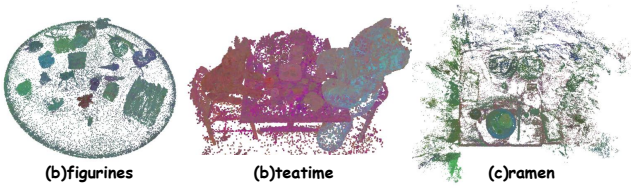

Figure 1. Feature visualization of gaussian grouping.

3. Hierarchical Segmentation

Figure 2 illustrates hierarchical 3D votes, with the left side showing the first level and the right side depicting a finer-grained second level. Initially, the 3D votes for the bear are clustered around a single center, but at the second level, they separate into two distinct centers, i.e., one for the upper body and one for the lower body.

The second row presents the rendering results of the clustered 3D votes, further confirming that the bear is accurately divided into two distinct parts, i.e., upper and lower body.

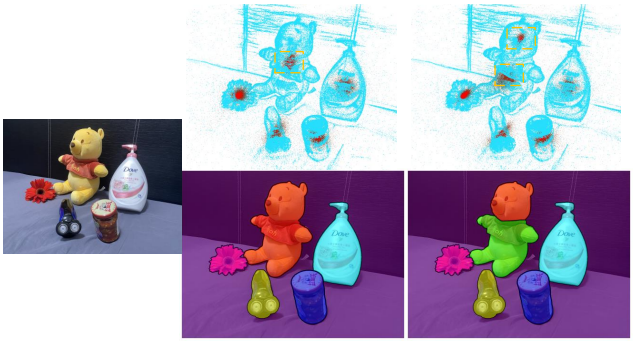

Figure 2. Hierarchical Segmentation Results. Using SAM, the bear is segmented into upper and lower parts, with a 2D vote assigned to each. After training, VoteSplat generates a 3D vote for each component. The bottom row displays the rendering results using the corresponding Gaussian primitives.

4. The Order of Blending and Projection

As stated in Section 3.3, "*blending is performed in 3D space before projection, ensuring stable voting.*" Here, we examine the opposite approach, i.e.,

$$\tilde{\mathbf{V}}_{\mathbf{i}}^{\mathbf{2d}} = \mathbf{H}\mathbf{V}_{\mathbf{i}}^{\mathbf{3d}}, \tag{1}$$

$$\tilde{\mathbf{V}}^{\mathbf{2d}} := \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} \tilde{\mathbf{V}}_{\mathbf{i}}^{\mathbf{2d}}. \tag{2}$$

In this case, distant 3D votes may be far from the instance center in 3D space, yet after perspective projection, they can appear better aligned with 2D votes. This misalignment makes convergence toward the instance center less reliable, as votes may cluster at a farther point while still reducing the loss, leading to inaccurate localization.

5. Quantitative ablation study.

We conducted ablation experiments on the 3D-OVS dataset, with quantitative results presented in Table 1. The results demonstrate that each component of our method contributes to the overall segmentation performance. In particular, allowing background Gaussians to participate in the voting process leads to a significant degradation in segmentation accuracy, highlighting the importance of explicitly filtering them out.

Case	Eq. 7	L_d	mIoU(%)↑	mAcc.↑
#1	✓		76.04	0.88
#2		✓	58.72	0.72
#3	✓	✓	**85.66**	**0.96**

Table 1. Quantitative ablations result on the snacks scene of the 3D-OVS dataset.

6. 3D Vote Visualization of Other Scenes In the 3D-OVS Dataset