Semantic Fusion for 3D Gaussian Splatting

Tarun Gangadhar Vadaparthi

Abstract

Recently, 3D Gaussian Splatting (3DGS) has become a cutting-edge neural radiance field representation, allowing complex scenes to be rendered in real time. Standard 3DGS lacks semantic interpretability despite its effectiveness and visual fidelity because individual Gaussian splats encode appearance but not semantic meaning. This work presents a pipeline for *Semantic Gaussian Splatting*, in which multi-view 2D segmentation is used to fuse semantic labels into the 3D representation. The Segment Anything Model (SAM) is used to segment a set of novel views of a Gaussian-rendered scene, resulting in dense per-view masks. A majority-vote aggregation across per-view masks is then used to assign a semantic label to each Gaussian after it has been projected into every camera view. A semantically annotated Gaussian point cloud with consistent class-level coloring can be re-rendered as the end product. The approach effectively separates large structures (walls, floor) and object-like regions (bookshelves, furniture) without the need for explicit 3D supervision, according to experiments conducted on a playroom dataset. This proof-of-concept highlights a path toward object-aware neural rendering, semantic editing, and physics-based scene interaction by bridging the gap between 2D foundation models and 3D radiance fields.

1 Introduction

With the advent of Neural Radiance Fields (NeRF) and their offspring, which allow for photorealistic novel view synthesis, neural scene representations have advanced quickly. Among these, 3D Gaussian Splatting (3DGS) has drawn interest lately due to its capacity to render a scene in real time and with high visual fidelity by representing it as a set of anisotropic Gaussian primitives. Formally, a 3DGS scene is made up of millions of Gaussians that are rasterized to create continuous radiance fields and parameterized by position, covariance, opacity, and color. Despite its rendering power, this representation is only geometric and photometric; it doesn't convey any semantic information about what the splats represent.

For downstream tasks like object-centric editing, scene decomposition, compression, or physics-based simulation, semantic comprehension is crucial. For instance, interactive functions like recoloring, removal, or material simulation would be made possible by the ability to differentiate between splats that belong to the floor, walls, furniture, or objects. However, the lack of 3D annotated datasets makes it impractical to train 3DGS with direct semantic supervision. On the other hand, 2D vision foundation models, like the Segment Anything Model (SAM), lack direct 3D consistency but offer strong segmentation priors in image space. By presenting a pipeline for *Semantic Gaussian Splatting*, this work explores how to integrate these paradigms. The main idea is to project Gaussians into several new views, use a 2D model to segment each view, and then use multi-view voting to fuse the labels back into 3D. This method makes use of geometry and segmentation consistency without requiring extra training of the Gaussian representation. The technique makes semantic-aware rendering possible by combining 2D semantic priors with 3D Gaussian fields, paving the way for real-time neural rendering systems to comprehend scenes at a higher level.

2 Methodology

The suggested framework combines 2D segmentation masks from several novel views to add a semantic labeling layer to 3D Gaussian Splatting (3DGS). Rendering, segmentation, projection, and semantic fusion are the four steps that make up the pipeline. Below is a formal description of each step.

2.1 Scene Representation

A Gaussian point cloud of a reconstructed scene is given, with each splat parameterized as follows:

$$\mathcal{G}_i = \big(\mu_i, \Sigma_i, \mathbf{c}_i, \alpha_i\big),$$

with center $\mu_i \in \mathbb{R}^3$, covariance $\Sigma_i \in \mathbb{R}^{3\times3}$ regulating anisotropic spread, color $\mathbf{c}_i \in [0,1]^3$, and opacity $\alpha_i \in [0,1]$. By projecting Gaussians into image space and adding up their contributions, rendering is accomplished through differentiable rasterization.

2.2 Novel View Rendering

The projection of a 3D point $\mathbf{x}$ into homogeneous image coordinates for each camera view v, defined by intrinsics $K_v \in \mathbb{R}^{3\times3}$ and extrinsics $(R_v, \mathbf{t}_v)$, is as follows:

$$\mathbf{u}_v \sim K_v\big(R_v\mathbf{x} + \mathbf{t}_v\big), \quad \mathbf{u}_v = \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}.$$

The pixel location at which the Gaussian center μ_i is rasterized in this case is $\mathbf{u}_v$. The result is a set of V rendered RGB images $\{I_v\}_{v=1}^{V}$.

2.3 2D Semantic Segmentation

The Segment Anything Model (SAM) processes each image I_v to generate a set of binary masks $\{M_{v,j}\}_{j=1}^{J_v}$ with corresponding scores. Only the largest k masks above a minimum area threshold $A_{\min}$ are kept in order to control complexity:

$$\mathcal{M}_v = \{M_{v,j} \mid \text{area}(M_{v,j}) > A_{\min},\ j \le k\}.$$

This creates a label map $L_v : \Omega \to \{0, \ldots, C_v\}$, where background is represented by 0 and the retained regions are indexed by $1 \ldots C_v$.

2.4 Semantic Voting in 3D

To lift labels into 3D, each Gaussian $\mathcal{G}_i$ is projected into all views. Let $\mathbf{u}_{i,v}$ represent the projected pixel for $\mathcal{G}_i$ in view v. The label that goes with it is:

$$\ell_{i,v} = L_v(\mathbf{u}_{i,v})\,.$$

$\ell_{i,v} = 0$ if $\mathbf{u}_{i,v}$ is behind the camera or outside the image plane. Votes are accumulated by each Gaussian across all views:

$$\mathbf{h}_i(c) = \sum_{v=1}^{V} \mathbb{1}\big[\ell_{i,v} = c\big], \quad c \in \{0, 1, \ldots, C\}.$$

The final assigned semantic label is given by:

$$\hat{\ell}_i = \arg\max_c \mathbf{h}_i(c).$$

2.5 Semantic Rendering

Colors are now taken from a palette $\mathcal{P} = \{\mathbf{p}_c\}_{c=0}^{C}$ associated with class indices during semantic rendering rather than from $\mathbf{c}_i$. Next, the rendered image is:

$$I_{\text{sem}}(\mathbf{u}) = \sum_i \alpha_i(\mathbf{u})\, \mathbf{p}_{\hat{\ell}_i},$$

where the cumulative contribution of Gaussian i at pixel $\mathbf{u}$ is represented by $\alpha_i(\mathbf{u})$. As a result, the scene is consistently visualized semantically from various perspectives.

3 Results

The publicly accessible "Playroom" dataset was used to test the suggested Semantic Gaussian Splatting pipeline. Eight new rendered views, each with a resolution of 256×256, and a Gaussian reconstruction of the scene made up the input.

3.1 Label Fusion Statistics

SAM segmentation generated nine to fifteen different masks for every view. The fused 3D label histogram showed $C = 16$ candidate semantic classes following area filtering and merging across all views. The majority of splats were found in four dominant clusters, which included a floor region, large wall segments, and smaller furniture-like regions (desk, bookshelf). In terms of numbers, the background class (label 0) had roughly 1.7×10^6 splats, whereas the next most common object-level classes had 1×10^3 to 1×10^4 splats.

3.2 Semantic Rendering

Figure 1 contrasts the semantic-colored rendering generated by our pipeline with the original photometric rendering. The ability of the voting process to transfer 2D semantics into the 3D representation was demonstrated by the consistent identification and coloring of large structural elements (floor, walls) across various views and the distinct coloring of smaller object regions.

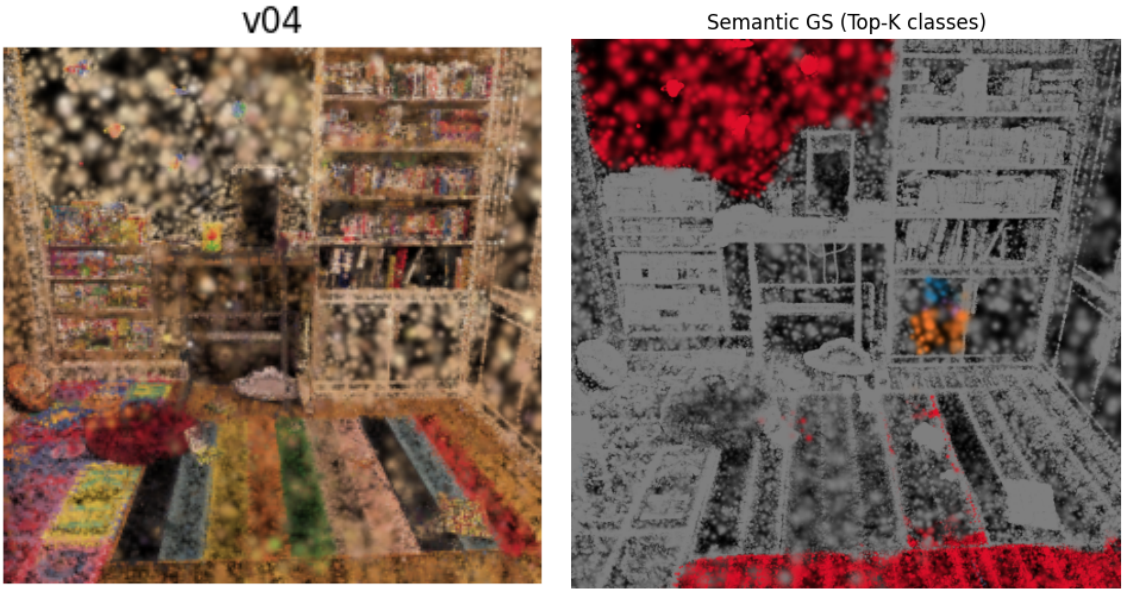

Figure 1: Left: original appearance-based 3DGS rendering. Right: semantic re-rendering with fused class labels.

4 Discussion

The findings show that, without further training, semantic labels can be effectively incorporated into a Gaussian Splatting representation. The projection-and-voting mechanism offers a simple method of adding semantic structure to 3DGS while guaranteeing consistency across views. Crucially, this was accomplished without the need for costly 3D annotations by utilizing only 2D supervision from foundation models.

Nevertheless, a number of restrictions were noted. First, SAM frequently creates large, region-based masks that designate entire floor patches or walls as distinct semantic classes. These segments are technically consistent, but they don't follow the object-centric semantics that editors prefer. Second, because they are not as visible from several angles, smaller objects (like toys and books) are underrepresented in the fused label histogram. Third, the method is susceptible to occlusion and imbalance in viewpoints because the labels are solely based on voting: splats that are visible in fewer views are given less accurate semantic assignments.

Notwithstanding these drawbacks, the semantic re-renderings validate the viability of this strategy and its potential for use in subsequent stages. A route toward editing, selective rendering, or interactive simulation within a Gaussian representation is suggested by the separation of background from object-like classes.

5 Future Work

There are still a number of ways to enhance Semantic Gaussian Splatting.

- **Models Aware of Categories:** By incorporating semantic segmentation models (like DeepLab and Mask2Former) that have been trained on extensive labeled datasets, SAM could be replaced with category-specific labels, allowing for the recognition of explicit objects like shelves, desks, and chairs.
- **Hierarchical Label Fusion:** More stable per-splat labels could result from weighting contributions by view confidence, area size, or visibility using probabilistic or Bayesian fusion, which goes beyond simple voting.
- **Object-Centric Decomposition:** Large background regions may be divided into smaller, more meaningful semantic objects using post-processing techniques like connected-component analysis in 3D label space.
- **Combining Editing and Physics:** Semantic labels offer a natural starting point for physics-aware Gaussian Splatting, which allows splats from materials like "floor" or "water" to be simulated with the right dynamics. **Efficiency and Scalability:** Extending this approach to real-world datasets will require optimizing the multi-view projection and voting step to handle larger scenes (millions of splats).

All things considered, this project lays the groundwork for adding semantics to Gaussian splatting, bridging the gap between interpretable scene representations and photorealistic rendering.

6 Conclusion

A Semantic Gaussian Splatting pipeline that combines 3D radiance field representations with 2D foundation segmentation models was presented in this work. The approach effectively enriches Gaussian splats with semantic information by rendering multiple novel views, implementing per-view segmentation, and fusing labels back into 3D through a voting mechanism. Large structural elements (floor, walls) and smaller object-like regions could be reliably annotated and visualized in 3D, according to experiments conducted on the Playroom dataset. The method shows a promising path for object-aware neural rendering, despite some remaining drawbacks, such as the dominance of background regions and decreased fidelity for small objects. This framework lays the groundwork for further studies in interpretable 3D scene understanding, physics-based simulation, and semantic editing.

References

[1] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, *3D Gaussian Splatting for Real-Time Radiance Field Rendering*. ACM Transactions on Graphics (SIGGRAPH), 2023.

[2] 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*. In Proceedings of ECCV, 2020.

[3] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T.-Y. Lin, P. Dollár, and R. Girshick, *Segment Anything*. arXiv preprint arXiv:2304.02643, 2023.

[4] B. Cheng, A. Schwing, and A. Kirillov, *Masked-attention Mask Transformer for Universal Image Segmentation*. In Proceedings of CVPR, 2022.