Dynamic 3D Gaussian Splatting Object Reinitialization

Juan F. Atehortúa Paredes
Massachusetts Institute of Technology
Cambridge, MA
ate@mit.edu

Abstract

3D Gaussian Splatting techniques have recently emerged as a promising approach to address the tasks of novel-view synthesis and six degree-of-freedom (6-DOF) tracking of dense scene elements in a dynamic scene, most notably the method presented by Luiten et al. [?]. Though this method produces state of the art results in both performance and quality for these tasks, it exhibits a significant limitation in that it cannot handle new objects entering into the scene due to the fact that all of the scene's gaussians are initialized from the first frame. This project aims to address this by introducing an additional step in the training pipeline to identify new objects via identifying relative dips in the reconstruction losses of segmentation masks, initialize new gaussians for them via monocular depth estimation maps, and optimize the scene reconstruction to effectively initialize objects into the scene.

1. Related Work

The project builds from 'Tracking by Persistent Dynamic View Synthesis' [?] which proposes a dynamic scene novel-view synthesis method by initializing a scene parametrized by 200-300k dynamic 3D Gaussians which move and rotate over time to optimize the reconstruction of input images via differentiable rendering. This method relies on 3D Gaussian Splatting [?], a novel approach to volumetric rendering which boasts a significant performance boost over previous radiance field methods by leveraging scene sparsity and properties of Gaussians to analitically compute the rendering of a ray instead of sampling along it.

Image segmentation is a task that involves dividing a digital image into multiple segments to identify and categorize different objects and boundaries within an image to facilitate tasks like object detection, recognition, and scene understanding. For our project, we will use segmentation masks generated by Segment Anything [?] to identify where in the scene the PSNR reconstruction loss is particularly bad.

Monocular depth estimation is the task of estimating the depth map of a scene from a single image. We will use the Depth Anything model [?] to retrieve depth maps for the scene during object reinitialization, and leverage them alongside the pre-existing scene gaussians to initialize new gaussians that represent the new objects in the scene.

A tangential approach to improving static 3D Gaussian Splatting is the work of Chung et al. [?] in using depth maps to improve the quality of initialized gaussians. Indeed we will use a similar insight in the construction of our method to initialize the gaussians of new objects.

2. Motivation

The tasks of dynamic 3D world modelling and dense tracking are easily justifiable in the context of robotics, augmented/virtual reality, and autonomous driving by providing a reconstrusction on where everything in a scene is and how it moves. Pertaining to generative AI, the method would enable us to seamlessly generate high-fidelity assets for virtual reality and video games from captured video. The main limitation described in the original paper is that all the scene gaussians are initialized from the first frame, and as such the method cannot handle new objects entering the scene. By addressing this, we can effectively extend the method to broadly handle more complex scenes.

3. Methodology

The proposed method comprises of a new pipeline on top of the existing dynamic 3D Gaussian Splatting method. It is comprised by:

- Identifying New Objects: At each non-initial time step in the training process, we use the Segment Anything model to generate segmentation masks $\mathcal{M}_c$ associated with camera c for the ground truth

input images, ignoring those that comprise a very small part of the image. We then compute the PSNR loss between the ground truth and rendered images restricted to each mask. Provided we have high variance (hyperparameter), we identify such masks that exhibit a PSNR loss between the input image restricted to the mask $I|m$ and the rendered image restricted to the same $\tilde{I}|_m$ more than two standard deviations below the mean for object reinitialization, calling such subset $\mathcal{M}_c^*$. Concisely,

$$\mathcal{M}_c^* = \{m \in \mathcal{M}_c | \text{PSNR}(I|_m, \tilde{I}|_m) < \mathbb{E} - 2\sigma, \sigma^2 > v_{\min}\} \tag{1}$$

Where $v_{\min} = 2$ seems to be a good value experimentally.

If all of these subsets turns out to be empty, then there are no new objects identified in the scene and we can proceed with the original method. Henceforth, we will abuse notation and extend any operator f on a mask $m \in \mathcal{M}$ to the set of masks $\mathcal{M}$ by defining $f(\mathcal{M}) = \cup_{m \in \mathcal{M}} f(m)$.

- Initializing New Gaussians: For each camera c with non-empty $\mathcal{M}_c^*$, we retrieve an estimated depth map $\tilde{D}_c$ from the Depth Anything model and a sparse dense map D_c from reprojecting the scene gaussian means to the camera. To make use of scale-invariant depth estimation models, we additionally solve for a scale factor s_c such that

$$\mathbb{E}[D_c \backslash \mathcal{M}_c^*(D_c)] = s_c \cdot \mathbb{E}[\tilde{D}_c \backslash \mathcal{M}_c^*(\tilde{D}_c)] \tag{2}$$

That is, we restrict the depth maps to the masks' complement where we expect both maps to match so that we can then regularize the scale factor based on the average value of the maps in this region. Once we have the scale-adjusted dense depth map, we can sample it uniformly at $\mathcal{M}_c^*$ to our desired density and reproject it to yield new means μ_c^* that can be used in tandem with the color values of the input images to initialize new gaussians for the scene. More explicitly,

$$\mu_c^* = s_c * \pi_c^{-1}\left(\{m_i\}_{i=1}^{N_c} \sim \text{Uniform}(\tilde{D}_c | \mathcal{M}_c^*)\right) \tag{3}$$

where π_c is the projection function for camera c and N_c is the desired number of new gaussians to initialize. Since the success of the gaussian splatting training heavily depends on how we initialize the gaussians, this step is crucial. Experimentally we find setting N_c as follows yields good results:

$$N_c = |\mu_s| \cdot \frac{|\mathcal{M}_c^*|}{|H \cdot W|} \cdot \frac{2}{|\mathcal{C}|} \tag{4}$$

Where μ_s are the existing scene gaussian means, (H, W) is the resolution of the images, and $|\mathcal{C}|$ is the number of cameras. That is, we roughly expect half the cameras to see the new object, and we want the amount of new gaussians to be proportional to both the relative space that the object occupies in the scene and the amount of gaussians already initialized.

- Optimizing New Gaussians: After initialization, we opt to optimize these new gaussians by running the usual 3D Gaussian Splatting training loop, but with the non-newly initialized gaussians frozen for a predetermined amount of iterations.

Since the PanopticSports dataset used in the original paper does not exhibit any scene with new objects entering, we will create an adversial synthetic dynamic scene using Kubric [?] by modifying the MoVI-A dataset worker script to have multiple cameras in a dome, just like PanopticSports.

4. Analysis

We use the generated scene to contrast the original method described in [?] with the proposed addendum. As one might expect, the original method does not handle the objects entering the scene well.

5. Limitations and Future Work

The keen-eyed reader might have already picked up that the proposed method does not depend on the specific models we are using for segmentation and depth estimation. As such, it is exciting to see how the improvements in tackling these tasks will make a downstream impact on the performance of the method provided. Indeed it would be interesting to see how the method performs with well segmented images and LiDAR depth maps.

Additionally, we have just tested the method on one synthetic adversial scene, testing its robustness on real data and more complex scenes is a natural next step to better gain insights on how to improve this method.

The broad idea of identifying the space in the scene where there are new objects and initializing gaussians in them leveraging the context around it seems to have some merit to it, but the details of how we achieve this could very well be refined, or even replaced by more sophisticated insights. Hopefully, this paper lays the groundwork to ask the right questions to improve dynamic scene reconstruction and novel-view synthesis. For instance, could we find a more clever way to sample from the depth maps to initialize new gaussians?

Maybe we can go beyond just determining the gaussian's mean and color, and we can use the geometric context to determine some of the rest of the gaussian's parameters at initialization, as the quality of the gaussian splatting heavily depends on the quality of the initializations. Perhaps we can modify modern techniques in SfM like Dust3r [?] to initialize the gaussians in a more informed way.

Figuring out how to gracefully handle objects that take more than one frame to fully enter the scene would also be an interesting direction, as we currently reinitialize such objects every timestep naively.

On a final note regarding future threads to pull when thinking about this method, notice how we didn't leverage the stereo framework that the multiple cameras provide. This was done deliberately in hopes of keeping the method relevant if future advances in gaussian splatting allow for single/few image scene reconstruction. However, it might be the case that we can more efficiently achieve each step of the proposed method leveraging stereo techniques in extracting the new object's geometric data.

6. Conclusion