Dynamic Gaussians in the Wild: Sparse-view Unposed Dynamic 3D Gaussian Splatting

Juan F. Atehortúa Paredes, Alice Yu
Massachusetts Institute of Technology
77 Massachusetts Ave, Cambridge, MA
ate@mit.edu, alicey@mit.edu

Abstract

Over the past year, the task of novel view synthesis has seen massive improvements on the SOTA with the introduction of 3D Gaussian Splatting (3DGS) based techniques [4], leveraging the Gaussian primitive's algebraic properties to sample rays analytically and optimize a scene in a fully differentiable manner. Though 3DGS boasts rendering times that are more than an order of magnitude faster than alternatives like NeRF variants [10] [1], techniques to train the models have mostly relied on complicated heuristics that require many views with known poses and intrinsics, as well as a sparse point cloud representation of the scene. These priors have been traditionally very hard to obtain, and necessitate very accurate estimations to be able to work; however, recent work by Fan et al. [3] has shown that by leveraging the dense point cloud output and camera estimations of Dust3r [8], splats can be trained seamlessly from sparse views letting Dust3r handle the prior generation. Our contribution is thus to extend these insights to the dynamic case by reimplementing the core ideas of InstantSplat on top of the pipeline proposed by Luiten et al. in their brilliant paper "Dynamic 3D Gaussians: Tracking by Persistent Dynamic View Synthesis" [7], being the first to create an end to end pipeline that can recover dynamic 3d gaussian splat scenes from a low number of videostreams without intrinsics or extrinsics.

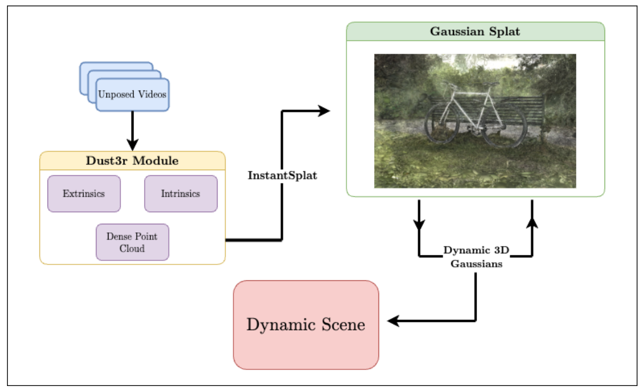

Figure 1. End to end pipeline corresponding to the method proposed: 1. Use Dust3r to generate dense Gaussian Splatting priors from unposed and sparse videos. 2. Implement the ideas of InstantSplat to generate high-quality splats of the first time step. 3. Incorporate dynamic training to successfully yield a dynamic scene representation.

1. Related Work

Many approaches[6][9][2] have recently been proposed to use 3D Gaussian Splatting and NeRF techniques to represent dynamic scenes for the purpose of Novel View Synthesis (NVS). Most notably, the method proposed in 'Dynamic 3D Gaussians'[7] notes the intuitive temporal relationships in the scene representation of adjacent time steps to develop a set of heuristic losses and parameter updates that enable fast training of a splat for a time step based on the spatial attributes of the prior splats. In addition to the scene representation yielded by the method, persistent 6-DOF tracking naturally follows from the paths traced by the means of the Gaussian primitives.

Though this inductive approach gives incredible results that speak for themselves, it unfortunately suffers the same drawback that most dynamic approaches have in tUS Reno (NGVPN50) - SSOhat it is necessary that the initial splat is of high quality, thus requiring an expensive multi-view stereo rig with the cameras calibrated to be able to work.

InstantSplat[3] has recently come out with very promising adjustments to the 3DGS training that allows for high quality splats to be generated from sparse-view, unposed camera images. The crux of the method proposed lies in Dust3r[8], a novel, deep-learning based approach that breaks from traditional SfM methods such as COLMAP to yield a dense multi-view point cloud reconstruction, coupled with estimated camera parameters. All of this from as little as two images. Though maybe Dust3r is not the panacea for stereo reconstruction, as the estimations have some considerable noise, the model performs inference incredibly fast, and the dense reconstruction can then be leveraged to very quickly train a high quality splat, as the authors of InstantSplat explain.

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 reconstruction 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 without the need of an expensive MVS setup. As far as we are aware, we are the first to propose and end to end pipeline from sparse and unposed video streams to dynamic scenes.

3. Methodology

All of the timings provided were tested on an Nvidia RTX 5880.

The keen-eyed reader might have already caught on that to achieve the contribution that we propose, there need not be a complicated solution. Indeed what we propose is to generate a high quality splat using InstantSplat for the first frame of each video stream, then use that as a prior for the Dynamic 3D Gaussians method. Of course, since the InstantSplat code was not made publicly available at the time of making the project, we reimplemented InstantSplat within the already available code for Dynamic 3D Gaussians. We subsequently detail how we accomplished exactly that

3.1. Prior Inference from Dust3r

Given a set of camera video streams $c = \{c_i(t)|i \in N\}$ where N is the total amount of feeds and t the time step, we use the Dust3r model Φ to yield extrinsics $E(0) = \{E^{(c_i)}(0)|c_i \in c\}$, intrinsics $I(0) = \{I^{(c_i)}(0)|c_i \in c\}$ and a dense colored point cloud $P(0)$. Concisely,

$$\Phi(c) = \big(E(0), I(0), P\big) \tag{1}$$

We follow InstantSplat's suggestion to force the resulting intrinsic focals to be consistent with each other.

3.2. Training the Initial Splat

We initialize a Gaussian Rasterizer $\mathcal{R}$ and an Adam optimizer with the following parameter groups for the Gaussians g:

- 3D means (x, y, z) yielded from P.
- RGB colors (r, g, b) also from P.
- Rotation quaternion (qw_t, qx_t, qy_t, qz_t) for the Gaussians (After first time step)

Figure 2. Dust3r inference on 5 views[5], 6 seconds to generate

Figure 3. NVS after initial time step inference (9 seconds)

- 3D size in standard deviations (sx, sy, sz)
- Opacity o as a float for each Gaussian
- Camera poses $\big((qw_c, qx_c, qy_c, qz_c), \vec{x_c}\big)$ parametrized as quaternions and position vector.

We follow roughly the same losses and learning rates that were employed by 'Dynamic 3D Gaussians' and InstantSplat. Like InstantSplat additionally, we opt to not subsample the dense priors and to not use the densify heuristics.

3.3. Dynamic Steps

Once we have the initial Gaussian Splat, we let the existing Dynamic 3D Gaussian method take over. An important note is that due to time constraints we have not implemented way to generate binary segmentation masks for the ground truth images. Though the method still works without them, tracking is quite bad without them. We intend to generate these masks via computing optical flow in the future.

4. Experimental Results

Due to the time and compute constraints, we did not run full ablations with test views and a priori known poses or intrinsics. Nevertheless, we believe that the results speak for themselves and urge the reader to see some dynamic examples we've created in our project webpage. In any case, we mantain that metrics such as PSNR, LPIPS, and SSIM do not very accurately reflect the quality of a splat, even if they are useful for training purpose, as shown above.

We do hope in the future to test the optimized poses and see how much better they are than the initialized ones, especially on dynamic scenes where some cameras are moving.

We note that with the InstantSplat implementation, the initial splat actually trains a little faster than the subsequent. This means that perhaps an online algorithm with it might speed up the creation of dynamic scenes purely for NVS purposes, since we'd lose tracking on reinitialization. Despite this there is a lot of room to speed up the dynamic portion of the code, especially considering that a Dust3r pass is nearly instantaneous relative to the dynamic timesteps.

5. Discussion

Though our proposed method yields promising results, there are some important caveats that we must address.

- **Inherent limitations of Dust3r**: Though Dust3r is the key component to making this training extremely fast and efficient, in it's current form it has some very apparent shortcomings. Most notable, it is incredibly memory inefficient during inference and doesn't scale when provided with many images. It was also only trained on indoor scenes and objects, meaning that challenging depictions such as humans often come out somewhat bad. Images with a high level of symmetry, such as the ones taken for the original PanopticSports dataset that came with the Dynamic 3D Gaussian paper, fail to accurately align themselves, making it a significant failure mode.
- **Reliance on heuristics**: We acknowledge that the big contribution that InstantSplat provided was getting rid of the highly complex rules that COLMAP and the original Gaussian optimization had in favor to the learned representations that Dust3r provides. In contrast to this, we have integrated dynamic training, but we have not addressed the heuristic overhead that allows the method to work. The computation of the dynamic losses is currently the main bottleneck in training time.
- **No Pruning Strategy**: With the removal of densification, we are working start to end with only the Gaussians initialized by the first Dust3r pass. Because of this we end up with a lot of floaters in the scene, and we inherit the failure case from Dynamic 3D Gaussians where we cannot handle new objects coming mid scene.

A natural next step would be to address the limitations noted, particularly improving how we do our dynamics and figuring out a way to prune useless Gaussians. We believe this is on of the first steps towards a polished end to end pipeline to yield high quality scene reconstructions without the need of expensive equipment.

Though we believe the splats speak for themselves, a full ablation of the proposed method is in place to quantify how well it recovers camera parameters. Due to our limitations with compute, we didn't have the time or resources to run our same scenes through the original SfM + 3DGS pipeline.

6. Conclusion

In this paper, we have demonstrated a novel approach to dynamic 3D Gaussian splatting by integrating InstantSplat with Dynamic 3D Gaussians. Our method allows for efficient and accurate reconstruction of dynamic scenes from sparse, unposed video streams without relying on pre-calibrated camera setups. We have showcased the potential of this approach through initial experiments, indicating promising results in both the quality of the reconstructed scenes and the speed of the training process.

While our work marks a significant step forward in simplifying and accelerating dynamic scene reconstruction, several limitations remain. Dust3r, despite its advantages, exhibits memory inefficiencies and struggles with certain challenging scenes. Additionally, the reliance on heuristics and the lack of a pruning strategy for Gaussians are areas that require further improvement. Future work should focus on addressing these limitations and conducting comprehensive evaluations to better quantify the performance gains.

We believe our contributions provide a valuable foundation for future research in dynamic 3D scene reconstruction, and we encourage further exploration to refine and expand upon the methods presented.

7. Individual Contributions

- **Juan Atehortua**: Reimplemented the ideas from InstantSplat on top of the existing Dynamic 3D Gaussians code (Before InstantSplat's code was made public). Ran experiments on low fidelity data sources.
- **Alice Yu**: Implemented visualization tooling used to export our Gaussian splats and their training to video format. Created drafts of final presentation and paper.

References

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields, 2022. 1

[2] Wenjing Bian, Zirui Wang, Kejie Li, Jiawang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. 2023. 1

[3] Zhiwen Fan, Wenyan Cong, Kairun Wen, Kevin Wang, Jian Zhang, Xinghao Ding, Danfei Xu, Boris Ivanovic, Marco Pavone, Georgios Pavlakos, Zhangyang Wang, and Yue Wang. Instantsplat: Unbounded sparse-view pose-free gaussian splatting in 40 seconds. In *CVPR*. IEEE, 2024. 1 The University of Texas at Austin, 2 Nvidia, 3 Xiamen University, 4 Georgia Institute of Technology, 5 Stanford University, 6 University of Southern California, * denotes equal contribution. 1

[4] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1

[5] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, and Zhaoyang Lv. Neural 3d video synthesis from multi-view video, 2022. 2

[6] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. *arXiv preprint arXiv:2312.16812*, 2023. 1

[7] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *3DV*, 2024. 1

[8] Wang, Shuzhe and Leroy, Vincent and Cabon, Yohann and Chidlovskii, Boris and Revaud Jerome. DUSt3R: Geometric 3D Vision Made Easy, 2023. 1

[9] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering, 2023. 1

[10] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images, 2021. 1