# Supplementary Material for "MOSAIC-GS: Monocular Scene Reconstruction via Advanced Initialization for Complex Dynamic Environments"

## 1. Implementation details

### 1.1. Dynamic Objects Tracking and Scene Flow Extraction

To reduce initialization time, we extract dynamic object masks and their temporal tracks using SAM2 [9] only on a subset of frames: up to 150 frames for iPhone DyCheck [2] sequences and 30 frames for the NVIDIA Dynamic Scene datasets [10, 11]. Sufficient temporal spacing between sampled frames is important, as too-small gaps may cause true object motion to be mistaken for noise during epipolar error thresholding. Subsampling naturally enforces this separation. We use a Sampson epipolar error threshold of $\tau_{\text{epi}} = 3$ and a segmentation-confidence threshold of $\tau_{\text{mask}} = 0.8$, both determined empirically.

All frames are still used for per-point tracking and scene-flow estimation, which improves trajectory accuracy with minimal additional computational cost. For the DyCheck [2] and NVIDIA (Gaussian Marbles) [11] datasets, point trajectories are extracted using BootsTAPIR [1]. However, for the original NVIDIA dataset [10], following MoSca [6], we use CoTracker [3], which provides more stable tracking under sparse temporal sampling than BootsTAPIR [1] (see Table 1).

| Tracker | PSNR↑ | SSIM↑ | LPIPS↓ |
|---|---|---|---|
| BootsTAPIR | 26.22 | 0.865 | 0.062 |
| CoTracker | **26.26** | **0.866** | **0.060** |

Table 1. Quantitative evaluation of novel view synthesis results obtained with different tracking methods on the original NVIDIA dataset [10].

### 1.2. Trajectories Encoding

For trajectory encoding, we use a polynomial degree of $d_{\text{pol}} = 3$ and a Fourier degree of $d_{\text{Fourier}} = 32$ for long sequences in DyCheck [2]. For the NVIDIA (Gaussian Marbles) dataset [2], which contains shorter videos, a reduced degree of $d_{\text{Fourier}} = 24$ provides a more compact representation without sacrificing accuracy. For the original NVIDIA dataset [10], which offers only 12 training frames, we further decrease the degree to $d_{\text{Fourier}} = 4$. This ensures that the number of temporal samples exceeds the number of unknown Poly-Fourier coefficients, which produces an overdetermined system that can be solved robustly using least squares.

### 1.3. Initialization of Static Regions

We sample points from static regions every 20th frame, which provides broad scene coverage without excessive oversampling. For the original NVIDIA dataset [10], which contains only 12 spatially distributed frames, we instead sample points from every second frame.

Our sampling follows a slightly modified version of the probability-based strategy proposed by Meuleman *et al.* [8]. For each pixel $(x, y)$ in frame $\mathbf{I}_t$, we compute the Laplacian-of-Gaussian magnitude $\text{LoG}(x, y)$. The sampling probability is then defined as

$$P(x,y) = \begin{cases} \text{LoG}(x,y), & \text{if } \mathbf{M}_t(x,y) = 0, \\ 0, & \text{if } \mathbf{M}_t(x,y) > 0, \end{cases} \tag{1}$$

where $\mathbf{M}_t$ denotes the combined segmentation mask for frame $t$, with the value 0 indicating static regions.

### 1.4. Photometric Optimization

We perform photometric optimization for 20,000 iterations. The loss function combines photometric losses L1 and SSIM with $\lambda_{\text{ssim}} = 0.2$ and depth regularization loss with $\lambda_{\text{depth}} = 0.2$. Although our method supports spherical harmonics, in our experiments we set the degree to 0, using only RGB values for Gaussian colors. We found that this provides sufficient representation capacity on monocular benchmarks while remaining more compact than higher-degree spherical harmonics.

The model is trained using the Adam [5] optimizer with separate learning rates for different Gaussian parameters. For deformation parameters, we use an initial learning rate of $5 \times 10^{-6}$ for the mean offsets, decaying exponentially to $5 \times 10^{-8}$, and $1 \times 10^{-6}$ for rotation quaternions, decaying to $1 \times 10^{-8}$. All other optimization settings follow the original Gaussian Splatting scheme [4].

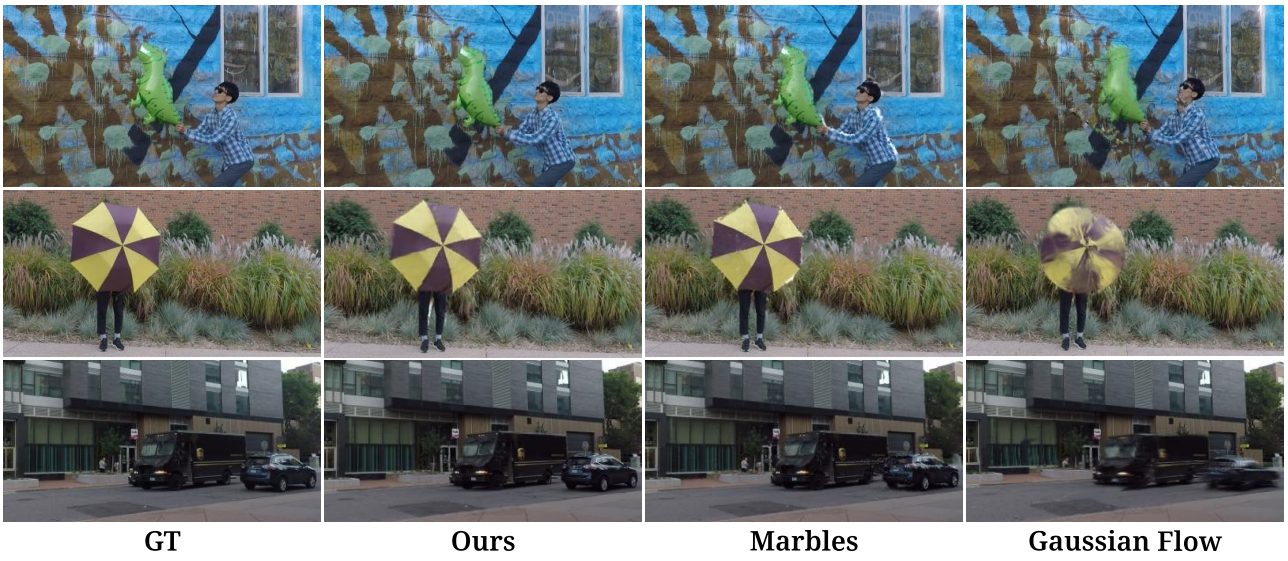


Figure 1. **Qualitative comparison on the NVIDIA (Gaussian Marbles) dataset [11].** MOSAIC-GS produces smoother, more coherent object structures and cleaner contours, whereas Gaussian Marbles exhibits grainy artifacts and irregular, patchy boundaries in dynamic regions.

For the NVIDIA datasets [10, 11], scenes are rescaled to the $[-1, 1]$ cube. For DyCheck [2], we use the refined camera poses provided by MoSca [6] without additional scaling. All experiments were conducted on a single NVIDIA RTX 4090 GPU with 24 GB of memory.

## 2. Additional Experiments

In addition to the experiments presented in the main paper, we include a qualitative comparison of MOSAIC-GS with prior methods on the NVIDIA (Gaussian Marbles) dataset [11], as illustrated in Fig. 1.

We also provide a per-scene quantitative comparison across both variants of the NVIDIA datasets [10, 11], summarized in Table 2. Additionally, the *video_results* folder in our supplementary materials contains videos of MOSAIC-GS reconstruction results captured from the evaluation cameras for scenes from the NVIDIA (Gaussian Marbles) [11] and DyCheck [2] datasets.

Figure 2. **Illustration of two key limitations of MOSAIC-GS. Left:** Insufficient scene flow refinement leads to missing fine facial details when the person turns away from the camera. **Right:** Splitting a single object into separate instances prevents the occluded hand from inheriting motion cues from the visible body, leading to incorrect motion estimation.

## 3. Limitations and Future Work

As discussed in the main paper, MOSAIC-GS has two primary limitations: (1) reliance on accurate initialization, and (2) insufficient scene flow refinement in certain cases.

First, our method depends on the quality of initial scene flow estimation and segmentation masks. If dynamic regions are not correctly identified during initialization, photometric optimization may fail to capture their motion accurately, leading to reconstruction artifacts. Because the approach relies on external models for optical flow, segmentation, and point tracking, it may also inherit their limitations. However, ongoing advances in these areas are likely to directly benefit our method.

Second, when dynamic objects are completely invisible in certain training frames, the system may struggle to reconstruct fine details of their appearance or motion. The examples for this type of limitation are presented in Fig. 2. This issue mainly stems from two factors: (1) limited precision of scene flow refinement during initialization, and (2)

| *NVIDIA (origin)* [10] | | | | | | | | *Reported: PSNR↑ / LPIPS↓* |
|---|---|---|---|---|---|---|---|---|
| | Balloon1 | Balloon2 | Jumping | Playground | Skating | Truck | Umbrella | Mean |
| Gaussian Flow [7] | 20.98 / 0.199 | 23.15 / 0.152 | 23.13 / 0.153 | 18.42 / 0.170 | 26.57 / 0.099 | 25.92 / 0.075 | 22.60 / 0.189 | 22.97 / 0.148 |
| MoSca [6] | **23.58** / 0.10 | **27.80** / 0.05 | **25.01** / 0.09 | **24.25** / 0.05 | **33.41 / 0.03** | **27.83** / 0.08 | 25.17 / 0.09 | **26.72** / 0.07 |
| Ours | 23.21 / **0.090** | 27.77 / **0.040** | 24.44 / **0.085** | 24.14 / **0.049** | 31.73 / 0.032 | 27.24 / **0.060** | **25.27 / 0.066** | 26.26 / **0.060** |
| *NVIDIA (Gaussian Marbles)* [11] | | | | | | | | *Reported: PSNR↑ / LPIPS↓* |
| | Balloon1 | Balloon2 | Jumping | Playground | Skating | Truck | Umbrella | Mean |
| Gaussian Flow [7] | 22.24 / 0.068 | 21.72 / 0.116 | 19.28 / 0.123 | 17.05 / 0.134 | 26.14 / 0.039 | 24.39 / 0.092 | 22.41 / 0.095 | 21.89 / 0.095 |
| Gaussian Marbles [11] | 24.09 / 0.041 | **23.84** / 0.077 | 20.20 / **0.100** | 17.48 / 0.127 | **27.83 / 0.030** | 27.30 / 0.049 | **25.04 / 0.059** | 23.68 / 0.069 |
| Ours | **24.44 / 0.036** | 23.76 / 0.077 | **20.70** / 0.101 | **17.54 / 0.126** | 27.69 / 0.032 | **27.82 / 0.044** | 24.59 / 0.066 | **23.79** / 0.069 |

Table 2. **Per-scene quantitative results on NVIDIA datasets.** MOSAIC-GS achieves the lowest LPIPS scores on the original NVIDIA dataset [10] and establishes a new state-of-the-art on the NVIDIA (Gaussian Marbles) dataset [11].

segmentation of a single object into multiple instances (e.g., human body parts segmented separately). The first issue could be mitigated by introducing rigidity constraints during photometric optimization, while the second opens an interesting direction for future work on leveraging correlations among dynamic instances to infer the motion of unobserved parts from related visible ones.

## References


[1] Carl Doersch, Pauline Luc, Yi Yang, Dilara Gokay, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ignacio Rocco, Ross Goroshin, João Carreira, and Andrew Zisserman. BootsTAP: Bootstrapped training for tracking-any-point. *Asian Conference on Computer Vision*, 2024. 1

[2] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Dynamic novel-view synthesis: A reality check. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2

[3] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 18–35, Berlin, Heidelberg, 2024. Springer-Verlag. 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), July 2023. 1

[5] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *International Conference on Learning Representations*, 12 2014. 1

[6] Jiahui Lei, Yijia Weng, Adam W. Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6165–6177, 2025. 1, 2, 3

[7] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21136–21145, 2024. 3

[8] Andreas Meuleman, Ishaan Shah, Alexandre Lanvin, Bernhard Kerbl, and George Drettakis. On-the-fly reconstruction for large-scale novel view synthesis from unposed images. *ACM Transactions on Graphics*, 44(4), 2025. 1

[9] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 1

[10] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5335–5344, 2020. 1, 2, 3

[11] Colton Stearns, Adam W Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 1, 2, 3