MOSAIC-GS: Monocular Scene Reconstruction via Advanced Initialization for Complex Dynamic Environments

Svitlana Morkva[1] Maximum Wilder-Smith[1] Michael Oechsle[2] Alessio Tonioni[2]
Vaishakh Patil[1] Marco Hutter[1]
[1]ETH Zürich [2]Google
{smorkva,mahutter}@ethz.ch {mwilder,patilv}@leggedrobotics.com
{michaeloechsle,alessiot}@google.com

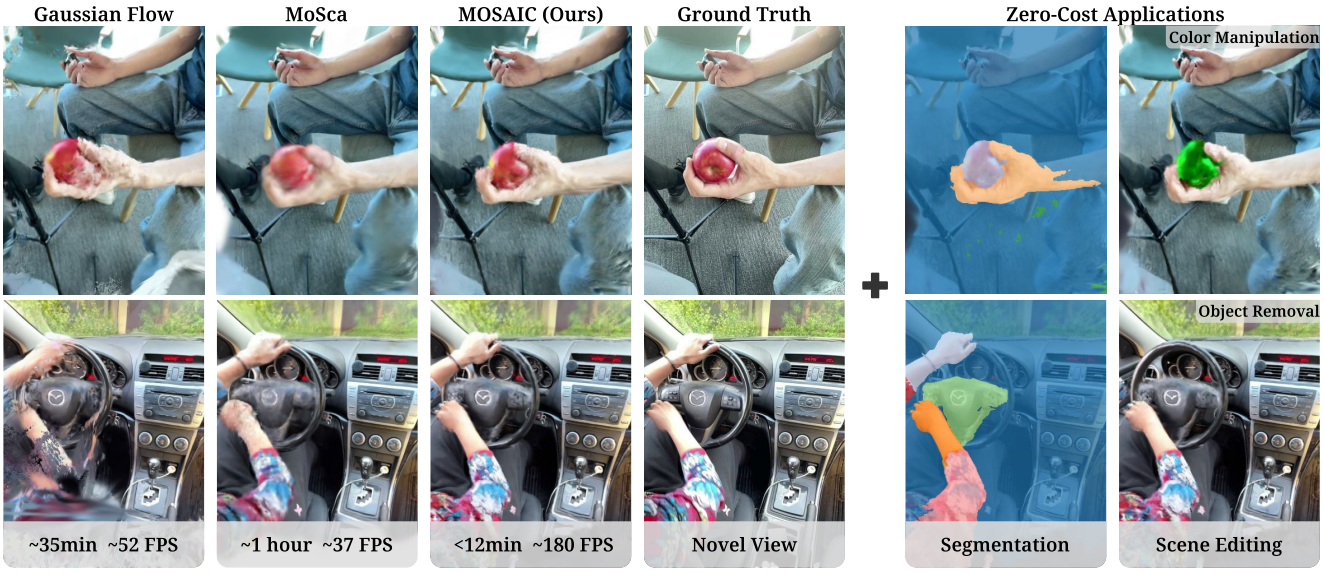

Figure 1. MOSAIC-GS produces high-quality reconstructions of dynamic scenes, capturing fine details in dynamic regions more accurately while significantly reducing training and rendering time compared to prior methods. Beyond reconstruction, our approach enables dynamic scene segmentation and editing, allowing objects to be removed, isolated, or manipulated without any additional computational cost.

Abstract

We present MOSAIC-GS, a novel, fully explicit, and computationally efficient approach for high-fidelity dynamic scene reconstruction from monocular videos using Gaussian Splatting. Monocular reconstruction is inherently ill-posed due to the lack of sufficient multiview constraints, making accurate recovery of object geometry and temporal coherence particularly challenging. To address this, we leverage multiple geometric cues, such as depth, optical flow, dynamic object segmentation, and point tracking. Combined with rigidity-based motion constraints, these cues allow us to estimate preliminary 3D scene dynamics during an initialization stage. Recovering scene dynamics prior to the photometric optimization reduces reliance on motion inference from visual appearance alone, which is often ambiguous in monocular settings. To enable compact representations, fast training, and real-time rendering while supporting non-rigid deformations, the scene is decomposed into static and dynamic components. Each Gaussian in the dynamic part of the scene is assigned a trajectory represented as time-dependent Poly-Fourier curve for parameter-efficient motion encoding. We demonstrate that MOSAIC-GS achieves substantially faster optimization and rendering compared to existing methods, while maintaining reconstruction quality on par with state-of-the-art approaches across standard monocular dynamic scene benchmarks.

1. Introduction

Recent advances in Neural Radiance Fields (NeRFs) [36] and 3D Gaussian Splatting [20] have revolutionized 3D scene reconstruction, enabling highly accurate and photorealistic modeling of complex environments. However, these techniques are limited to static scenes, restricting their applicability in domains that involve dynamic environments, such as robotics, autonomous driving, or digital-twin generation for virtual and augmented reality (VR/AR) applications.

Several recent methods extend these ideas to dynamic scenes [6,25,33,38,39,54], but the problem remains largely unsolved. Existing approaches often require extensive training time, substantial memory and storage resources, and suffer from slow rendering, while still exhibiting noticeable artifacts in regions with complex motion. Many of these methods also prioritize visual realism over physical consistency, leading to distortions when observed from unseen viewpoints. This issue is particularly pronounced in settings with restricted camera viewpoints or monocular video acquisition, where the limited field of view further complicates physically consistent reconstruction. Such scenarios are common in robotics and embedded devices, where cameras typically provide only narrow or stereo views rather than wide multi-view coverage. Moreover, the limited computational and memory resources of these platforms make compact and efficient scene representations essential.

To address these challenges, we present MOSAIC-GS, an explicit and computationally efficient framework for dynamic scene reconstruction from monocular videos. MOSAIC-GS leverages a 3D Gaussian Splatting representation [20] constructed from different viewpoints of a dynamic scene acquired over an interval of time. Each view is paired with a camera pose and a depth estimate, obtained either by dedicated sensors or learned models. A key insight of our work is that inferring scene dynamics from purely visual data during photometric optimization is both inefficient and unreliable. Instead, we emphasize the importance of high-quality initialization, which has been shown to strongly influence reconstruction fidelity [20] and convergence speed [34].

Our pre-processing pipeline first focuses on robust identification and tracking of dynamic regions, enabling decomposition of the scene into static and dynamic components. This separation improves parameter efficiency and simplifies the subsequent optimization. We then estimate preliminary Gaussian motions from all available geometric and physical cues in the video, leveraging recent advances in optical flow [48], object segmentation [40], and per-point tracking [3,4] to extract per-point trajectories and object associations. We further incorporate rigidity constraints during this stage to enable refinement of these motions and efficient inference even in unobserved regions of the scene.

To compactly encode motion, we represent these trajectories using time-dependent Poly-Fourier curves, following Lin *et al*. [28], instead of the per-frame deformation models commonly used in prior monocular methods [24, 45, 52]. These trajectories are then used to initialize the dynamic Gaussians. Combined with disentanglement into static and dynamic components, this representation enables efficient modeling, fast training, real-time rendering, and robust handling of non-rigid deformations. The proposed MOSAIC-GS framework is particularly effective in monocular settings, where limited viewpoints make motion inference from visual cues alone unreliable.

The main contributions of our work can be summarized as follows:

- An advanced pre-processing framework for dynamic Gaussian Splatting that improves reconstruction quality and significantly accelerates training.
- A robust pipeline for dynamic object detection and tracking across monocular video sequences.
- A novel trajectory refinement method that extracts and optimizes 3D motion using per-instance rigidity constraints.
- State-of-the-art training efficiency and inference speed, achieved through scene disentanglement, compact deformation representation, and accurate motion initialization.

2. Related Works

2.1. Moving object detection

Detection of moving objects in video remains a challenging problem, particularly under changing illumination and camera motion. However, advances in deep learning, particularly convolutional neural networks (CNNs) [14,21,22], recurrent neural networks (RNNs) [47], and more recently transformer-based architectures [5, 51], have significantly improved robustness and accuracy. Building on these developments, recent motion detection frameworks integrate instance segmentation [13, 41] of common dynamic objects, dense optical flow [17, 46, 48], and epipolar geometry constraints to reliably detect motion even in challenging scenarios such as moving cameras or complex illumination [15,24,27,32,58].

2.2. Novel-View Synthesis of Dynamic Scenes

Novel-view synthesis aims to generate realistic images of a scene from previously unseen viewpoints. A major breakthrough came with the introduction of Neural Radiance Fields (NeRFs) [36], which enabled high-fidelity 3D scene representations and have since been extended to capture dynamic content [27, 37–39]. Despite their impressive quality, NeRF-based methods are computationally expensive and slow to render. To address this, hybrid approaches combine implicit neural fields with explicit 3D structures

such as voxels or planes [1, 2, 9, 42, 44], which improve efficiency but still struggle with real-time photorealistic rendering.

This limitation has motivated the development of explicit point-based models such as 3D Gaussian Splatting [20], which enable fast and high-quality rendering. Building on this foundation, recent works have extended Gaussian Splatting to dynamic scenes by modeling temporal motion in two main ways. The first line of work applies per-frame deformations, storing Gaussian transformations for each frame, either for all Gaussians or a smaller skeletal subset [16, 24, 33, 45, 52]. These methods estimate per-point trajectories either during optimization or via point tracking, but they are memory-intensive and do not scale well to long sequences. The second direction adopts continuous motion modeling, where Gaussian trajectories are represented with learnable functions [26, 28, 55]. While more compact and efficient, these methods often struggle to capture complex or rapid motions, especially when relying solely on photometric cues.

2.3. Monocular Dynamic 3D Scene Reconstruction

Monocular dynamic scene reconstruction is highly practical in real-world applications, but inherently ambiguous due to partial visibility of scene geometry and motion. Several NeRF-based methods proposed to address this by incorporating object-specific priors such as human shape models [30, 53], shadow modeling [56], and facial expressions [10]. While effective, these methods generalize poorly beyond specific object categories. To improve generalization, geometric and physical priors such as depth, optical flow, and rigidity-based deformation constraints have been explored [7, 11, 50], but monocular NeRFs still suffer from long training times and slow rendering.

Recent research has shifted toward explicit representations using 3D Gaussian Splatting, which offers faster and more efficient reconstruction. The progress in this area began with the work of Stearns *et al*. [45], who introduced isotropic Gaussian representations combined with motion trajectory blending from short sequences, guided by segmentation and depth priors. With advancements in point tracking, newer approaches propose to extract per-point trajectories from video and use them as motion bases shared across dynamic Gaussians [52]. The current state-of-the-art by Lei *et al*. [24] extends this idea with a Motion Scaffold Graph that connects trajectories into a spatio-temporal structure and refines them using rigidity and smoothness constraints.

Although these approaches achieve high-quality results, they remain limited by per-frame deformation modeling, leading to long training times and high memory usage. In contrast, we propose a compact and efficient representation for monocular dynamic scene reconstruction that achieves comparable quality to state-of-the-art methods while being significantly faster in both training and rendering.

3. Method

Overview. Given a monocular video of a dynamic scene, along with camera intrinsics $\mathbf{K}$, extrinsics $[\mathbf{R} \mid \mathbf{t}]_t$, and per-frame depth maps $\mathbf{D_t}$ obtained from sensors or monocular depth estimation models [57], our pipeline accurately reconstructs dynamic 3D scenes while reducing training and rendering time.

Preliminary experiments show that photometric optimization is highly sensitive to scene initialization: without accurate motion estimates, most dynamic-region points are initially pruned. This leads to longer training times and reduced reconstruction quality, as the model must recover lost data over many iterations. Meuleman *et al*. [34] similarly highlight that proper initialization accelerates convergence and improves visual quality. To prevent this data loss, we argue that the initialization should include richer information, rather than relying solely on Gaussian position and color as in most prior works [20, 28, 54].

To achieve this, we propose an advanced pre-processing pipeline with four main stages: (1) dynamic region detection, (2) segmentation and tracking of dynamic instances, (3) scene flow estimation and per-instance refinement using rigidity constraints, and finally (4) initialization of static and dynamic regions. In the last stage, the extracted scene flow is encoded as time-dependent Poly-Fourier curves $\mathbf{F}_t^i$ for compact trajectory representation, and initialization points are sampled adaptively to match the density of detail in each region.

The initialized Gaussians are then optimized and densified during the photometric optimization phase, regularized by a depth-based supervision loss. A full overview of the framework is shown in Fig. 2, and each stage is described in detail below.

3.1. Dynamic Regions Detection

To identify dynamic objects in the scene, we leverage dense optical flow in combination with epipolar geometry constraints. Instead of relying on full 3D reconstruction errors, which can be unreliable due to noisy depth maps, we employ the Sampson epipolar error.

Given two consecutive frames $\mathbf{I_t}$ and $\mathbf{I_{t+1}}$, we first compute the dense optical flow $\mathbf{u}_t : \Omega \rightarrow \mathbb{R}^2$ using RAFT [48]. Following Liu *et al*. [32], we compute the Sampson epipolar error for each pixel correspondence $\mathbf{x} \in \Omega$ using known camera extrinsics $[\mathbf{R} \mid \mathbf{t}]_t$, $[\mathbf{R} \mid \mathbf{t}]_{t+1}$ and intrinsics $\mathbf{K}$,:

$$e_{\text{epi}}(\mathbf{x}) = \frac{(\mathbf{x}'^\top \mathbf{F} \mathbf{x})^2}{(\mathbf{F}\mathbf{x})_1^2 + (\mathbf{F}\mathbf{x})_2^2 + (\mathbf{F}^\top \mathbf{x}')_1^2 + (\mathbf{F}^\top \mathbf{x}')_2^2}, \quad (1)$$

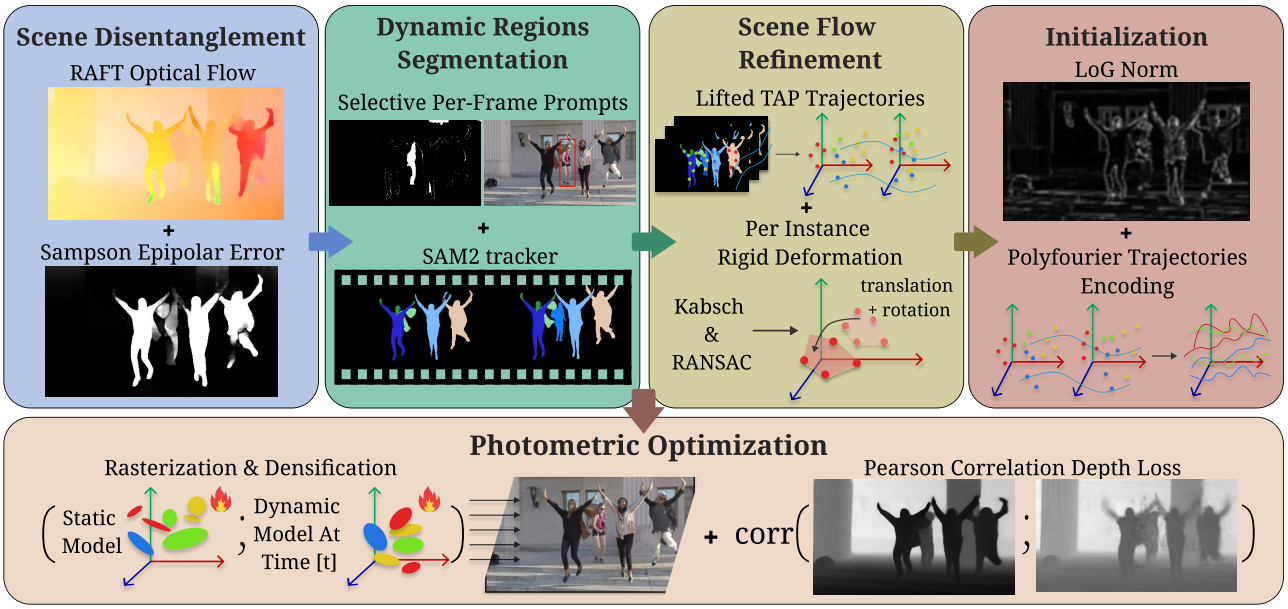

Figure 2. **Overview of the MOSAIC-GS framework.** The pipeline begins with four pre-processing steps: (1) dynamic region detection using optical flow and epipolar error, (2) segmentation and tracking of dynamic instances, (3) scene flow estimation with per-instance rigidity refinement, and (4) initialization of static and dynamic Gaussians with Poly-Fourier encoded trajectories. The extracted parameters serve as initialization for the main photometric optimization phase, where separate Gaussian models are created for static and dynamic regions to improve parameter efficiency, while being jointly rasterized for rendering. A depth-based Pearson correlation loss further enhances geometric consistency and reconstruction fidelity.

where $\mathbf{F}$ is the fundamental matrix derived from $[\mathbf{R} \mid \mathbf{t}]_t$, $[\mathbf{R} \mid \mathbf{t}]_{t+1}$ and $\mathbf{K}$, and $\mathbf{x}'$ is the corresponding pixel in the next frame. Pixels with high epipolar error correspond to regions that cannot be explained by camera motion alone, which may arise from true scene motion, inaccuracies in optical flow estimation, or errors in camera parameters. To filter out false positives, we consider dynamic only those regions where the epipolar error satisfies $e_{\text{epi}}(\mathbf{x}) > \tau_{\text{epi}}$.

3.2. Segmentation and Tracking of Dynamic Instances

Relying solely on epipolar error thresholding can be insufficient in certain scenarios. To improve the detection, tracking, and segmentation of dynamic regions across all frames, we incorporate the prompt-based semantic segmentation and tracking model SAM2 [40].

For each frame, we first extract bounding boxes around dynamic regions detected in the previous step. If no dynamic objects have been detected in earlier frames, these boxes are used directly as prompts for the segmentation model [40]. Otherwise, masks of previously detected and tracked instances are first subtracted from the dynamic regions to prevent duplicate instance tracking. Bounding boxes around the remaining regions are then used as prompts to generate accurate instance masks $\mathbf{M_t^j}$ for each new dynamic object j. Masks with low confidence scores $p_{\text{conf}}(\mathbf{M_t^j}) < \tau_{\text{mask}}$, are filtered out to remove unreliable detections.

High-confidence masks are added to the tracker and propagated forward at fixed intervals, reducing the computational overhead of re-tracking frames when new dynamic objects appear. After processing all frames, we obtain a set of tracked dynamic object masks $\{\mathbf{M_t^j}\}_{t=1}^{T}$ for the entire video. To account for objects that were static in earlier frames, we perform a reverse propagation pass, extending object masks to previous frames. A visualization of this process is shown in Fig. 3.

3.3. Scene Flow Estimation and Refinement

To extract scene flow for dynamic regions, we leverage recent advances in point tracking [3, 19], following prior works [24, 52]. We randomly sample $N_p = 10,000$ query points $\mathbf{q_t^i}$ within dynamic regions across all T frames and track them through the video, where i denotes the point index and t the index of a frame where the point has been queried.

Each trajectory $\{\mathbf{p_t^i}\}_{t=1}^{T}$ is assigned to a dynamic object j based on majority overlap with segmentation masks $\mathbf{M_t^j}$, producing per-object 3D point sets $\{\mathcal{P}_j\}_{j=1}^{J}$. Trajectories primarily associated with the static background are discarded. Combining segmentation and tracking resolves ambiguities from occlusions and imperfect masks, producing more temporally consistent instance identifiers.

The 2D trajectories are then lifted to 3D using depth

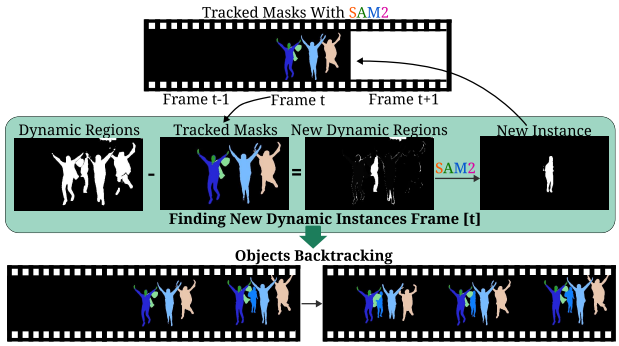

Figure 3. **Dynamic object detection and tracking.** Masks of previously detected objects are subtracted from dynamic regions to identify new instances. The remaining regions are used as prompts for SAM2 [40] to generate segmentation masks, which are tracked forward through all frames. A reverse propagation pass extends object masks to earlier frames where the object was initially static.

maps $\mathbf{D_t}$, camera intrinsics $\mathbf{K}$, and extrinsics $[\mathbf{R} \mid \mathbf{t}]_t$ to produce scene flow $\{\mathbf{P_t^i} \mid i = 1, \ldots, N_p;\ t = 1, \ldots, T\}$. To reduce noise and fill unobserved regions, we refine the flow of each dynamic object using rigid transformations. For object j, we consider points $\{(\mathbf{P_t^i}, \mathbf{P_{t+1}^i}) \mid i \in \mathcal{P}_j,\ v_t^i = 1,\ v_{t+1}^i = 1\}$, where v_t^i indicates visibility at frame t, and estimate the rigid transformation $(\mathbf{R_t^j}, \mathbf{t_t^j})$ that best aligns them using the Kabsch algorithm [18] with RANSAC [8]. Refined positions at $t+1$ are then computed as:

$$\mathbf{P}_{t+1}^i := \mathbf{R}_t^j \mathbf{P}_t^i + \mathbf{t}_t^j, \quad v_{t+1}^i := 1 \mid v_t^i = 1 \tag{2}$$

After processing all frames forward, we apply the same refinement backward to update points that were not observed in earlier frames but whose corresponding objects were visible at those frames. Remaining points with $v_t^i = 0$ are interpolated using the nearest frames $t' < t$ and $t'' > t$, where $v_{t'}^i = 1$ and $v_{t''}^i = 1$.

3.4. Initialization of Static and Dynamic Regions

To provide a compact representation of the scene we propose representing scene dynamics as time-dependent Poly-Fourier curves, following Lin *et al.* [28]. However, instead of learning these curves' coefficients during the photometric optimization phase, we propose initializing them directly from the refined scene flow obtained in the previous step. To achieve this, we are solving this system of linear equations for each trajectory $\{\mathbf{P_t^i}\}_{t=1}^T$:

$$Ax = y \tag{3}$$

where:

$$\begin{aligned} A &= \begin{bmatrix} \phi(t_1) & \phi(t_2) & \cdots & \phi(t_T) \end{bmatrix}^\top, \\ \phi(t) &= [\,1,\, t,\, t^2,\, \sin(\omega t),\, \cos(\omega t)\,]^\top, \\ y &= [\,P_i(t_1),\, P_i(t_2),\, \ldots,\, P_i(t_T)\,]^\top, \\ x &= [\,a_0,\, a_1,\, a_2,\, \ldots,\, b_1,\, c_1,\, \ldots\,]^\top. \end{aligned}$$

The resulting coefficients $\mathbf{x}$ then compactly encode each trajectory over time and are used to initialize the deformation parameters of dynamic Gaussians. In addition to motion, we initialize each Gaussian's color and scale to match the local appearance and level of details. Specifically, for each trajectory query point $\mathbf{q}_t^i$, we extract its color from the corresponding pixel in the RGB frame $\mathbf{I}_t$. To estimate the Gaussian scale, we evaluate the Laplacian of Gaussian (LoG) norm on $\mathbf{I}_t$ and incorporate its depth map $\mathbf{D}_t$ as proposed by Meuleman *et al.* [34]. The extracted Poly-Fourier coefficients, colors, and scales together initialize the mean position, deformation parameters, color, and scale of each dynamic Gaussian.

For static regions, we adopt the sampling strategy of Meuleman *et al.* [34]. To ensure broad scene coverage while avoiding oversampling of flat areas, sampling is performed only on a subset of temporally spaced training frames.

3.5. Photometric Optimization

The proposed MOSAIC-GS framework represents the scene using two distinct sets of 3D Gaussians: a static component $\mathcal{G}_s$ and a dynamic component $\mathcal{G}_d$. Each component is initialized separately and follows a distinct densification strategy to account for their differing parameter complexity. Formally, the full scene representation is defined as $\mathcal{G} = \mathcal{G}_s \cup \mathcal{G}_d$.

Each Gaussian in the static set $\mathcal{G}_s$ is parameterized by its mean position $\boldsymbol{\mu} \in \mathbb{R}^3$, rotation quaternion $\mathbf{q} \in \mathbb{H}$, scale $\mathbf{s} \in \mathbb{R}^3$, opacity $\alpha \in [0, 1]$, and color coefficients $\mathbf{c} \in \mathbb{R}^{3\times(B+1)^2}$, where B denotes the degree of spherical harmonics.

Dynamic Gaussians in $\mathcal{G}_d$ share the same canonical parameters but additionally store time-dependent deformation parameters encoded as Poly-Fourier coefficients $\{\mathbf{a}_k\}_{k=1}^{d_{\text{pol}}}$, $\{\mathbf{b}_k, \mathbf{c}_k\}_{k=1}^{d_{\text{Fourier}}}$ representing offsets to the Gaussian mean and rotation. Position and rotation deformations are modeled separately, with positional parameters initialized from the Poly-Fourier coefficients estimated during the preprocessing stage. The position offset at time t is given by:

$$\Delta\boldsymbol{\mu}(t) = \sum_{k=1}^{d_p} a_k t^k + \sum_{k=1}^{d_F} \Big(b_k \cos(k\omega t) + c_k \sin(k\omega t) \Big), \tag{4}$$

PSNR↑ / LPIPS↓	Apple	Block	Spin	Paper Windmill	Space-Out	Teddy	Wheel	Mean
Dyn. Gaussians [33]	7.65 / 0.766	7.55 / 0.684	8.08 / 0.651	6.24 / 0.729	6.79 / 0.733	7.41 / 0.690	7.28 / 0.593	7.29 / 0.692
4D Gaussians [54]	15.41 / 0.450	11.28 / 0.633	14.42 / 0.339	15.60 / 0.297	14.60 / 0.372	12.36 / 0.466	11.79 / 0.436	13.64 / 0.428
Marbles [45]	17.70 / 0.492	17.42 / 0.384	18.88 / 0.428	17.04 / 0.394	15.94 / 0.435	13.95 / 0.442	16.14 / 0.351	16.72 / 0.418
Gaussian Flow [29]	18.37 / **0.322**	15.62 / 0.352	16.66 / 0.302	18.14 / 0.217	17.12 / 0.286	13.48 / 0.376	14.18 / 0.323	16.22 / 0.311
Shape of Motion [52]	18.57 / 0.341	17.41 / 0.323	19.35 / 0.247	18.14 / 0.225	16.85 / 0.324	13.69 / 0.38	17.21 / 0.23	17.32 / 0.295
MoSca [24]	**19.4** / 0.34	**18.06** / 0.33	**21.31** / **0.19**	**22.34** / 0.15	**20.48** / 0.26	**15.47** / 0.35	**18.17** / 0.23	**19.32** / 0.264
Ours	18.92 / 0.328	17.5 / **0.29**	20.28 / 0.233	20.31 / 0.15	19.96 / **0.239**	14.42 / **0.343**	17.40 / **0.204**	18.4 / **0.255**

Table 1. Per-scene quantitative results on the DyCheck [12] dataset. Best results are in **bold**, second best are underlined. MOSAIC-GS achieves comparable performance to the state-of-the-art MoSca [24] method, while being significantly faster in both training and rendering and achieves a new state-of-the-art in terms of LPIPS for this dataset.

and the mean position becomes $\boldsymbol{\mu}(t) = \boldsymbol{\mu}_0 + \Delta\boldsymbol{\mu}(t)$.

For rotation, instead of directly adding time-dependent offsets to the base quaternion as in Gaussian Flow [28], we convert the Poly-Fourier output into a valid rotation by adding the identity quaternion and normalizing it to form a unit quaternion $\Delta\mathbf{q}(t)$. The final rotation is then obtained as:

$$\mathbf{q}(t) = \Delta\mathbf{q}(t) \otimes \mathbf{q}_0, \tag{5}$$

where $\mathbf{q}_0$ denotes the canonical rotation and $\otimes$ indicates quaternion multiplication. This design enforces consistent rotation parameterization and avoids artifacts from invalid quaternion updates.

Unlike Gaussian Flow [28], we do not model time-varying color deformations, which improves compactness and prevents the model from compensating motion inaccuracies through artificial color changes.

The overall loss function is defined as:

$$\mathcal{L} = (1 - \lambda_{\text{ssim}})\,\mathcal{L}_{\text{L1}} + \lambda_{\text{ssim}}\,\mathcal{L}_{\text{SSIM}} + \lambda_{\text{depth}}\,\mathcal{L}_{\text{depth}}, \tag{6}$$

where $\mathcal{L}_{\text{L1}}$ and $\mathcal{L}_{\text{SSIM}}$ denote photometric losses, and $\mathcal{L}_{\text{depth}}$ enforces geometric consistency with reference depth maps. To handle temporally inconsistent depth scales, $\mathcal{L}_{\text{depth}}$ is implemented as a Pearson correlation loss [31, 59], which preserves relative geometry while remaining invariant to absolute scale variations. Further implementation details are provided in supplementary materials.

4. Experiments

Datasets. We evaluate MOSAIC-GS on standard monocular dynamic scene reconstruction benchmarks, including the *iPhone DyCheck* [12] and *NVIDIA Dynamic Scene* dataset [43]. The *iPhone DyCheck* dataset contains seven casually captured dynamic scenes recorded with a moving monocular camera, each with up to 500 frames [12]. Each scene provides RGB images, camera intrinsics/extrinsics, noisy LiDAR depth maps, and ground-truth covisibility masks, along with one or two static cameras for novel-view evaluation.

Two distinct versions of the *NVIDIA Dynamic Scene* dataset exist: (1) the *original* release by Yoon et al. [43], and (2) the modified *Gaussian Marbles* variant introduced by Stearns et al. [45]. The *original* dataset [43] includes seven dynamic sequences captured by twelve static cameras. Monocular input is simulated by selecting one frame from each consecutive camera, yielding twelve spatially and temporally sparse frames. Evaluation uses images from the first camera as reference views. The *Gaussian Marbles* version [45] contains seven scenes with 100–200 frames each, captured by four static cameras. One camera provides monocular input, while the others serve as evaluation views.

This distinction has been overlooked in prior works [24], leading to inconsistent comparisons. To ensure fairness, we report results on both versions and compare with methods evaluated under the corresponding settings.

4.1. Novel-View Synthesis

DyCheck [12]. Following MoSca [24], we employ the BootsTAPIR tracker [3] to extract dense point trajectories across frames. MOSAIC-GS focuses exclusively on novel-view synthesis and does not include a camera pose refinement stage. In contrast, prior works often perform additional pose optimization to compensate for noisy ground-truth poses [24, 52]. For fair comparison, we adopt the refined poses extracted by MoSca [24].

Table 1 reports quantitative results comparing MOSAIC-GS with state-of-the-art monocular dynamic scene reconstruction methods, evaluated using PSNR and LPIPS. MOSAIC-GS achieves competitive performance, outperforming all prior methods except MoSca [24] in PSNR, while setting a new state-of-the-art in LPIPS across most scenes and on average. Notably, these results are obtained with substantially lower training and rendering times as shown in Table 2, enabled by our compact scene representation, static–dynamic scene disentanglement, and accurate motion initialization, which reduce the number of optimization steps required for reconstruction. Qualitative comparisons are shown in Fig. 4. Results for Gaussian Marbles are omitted due to the unavailability of their preprocessed dataset. MOSAIC-GS reconstructs finer details and produces sharper renderings in dynamic regions, resulting in noticeably better perceptual quality. While this level of de-

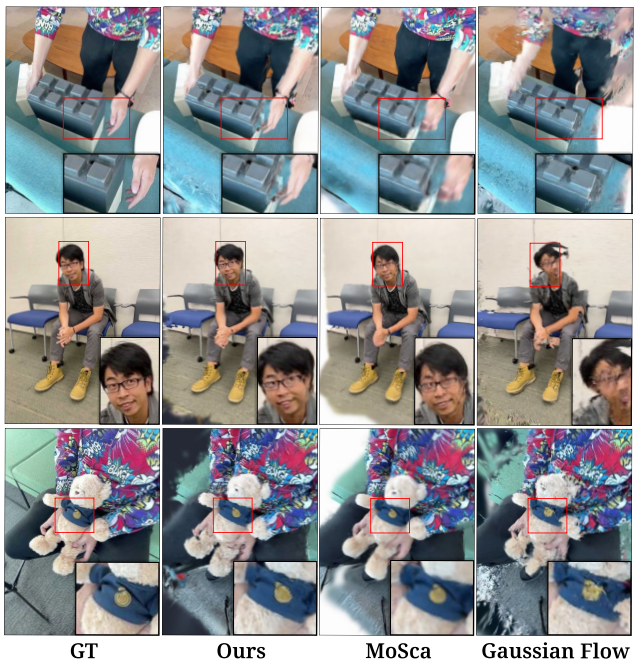

Figure 4. Qualitative comparison of novel view synthesis on the DyCheck [12] dataset. Our method produces sharper details in dynamic regions compared to prior approaches resulting in higher perceptual similarity with the ground truth evaluation view.

tail may slightly reduce PSNR, the improvement is better reflected by perceptual metrics such as LPIPS, which align more closely with visual quality.

NVIDIA Dynamic Scene Dataset [43]. Table 3 presents PSNR and LPIPS scores averaged across all scenes. Similarly to DyCheck, MOSAIC-GS achieves second-best PSNR after MoSca [24] and attains the best LPIPS among all compared methods. As illustrated in Fig. 5, MOSAIC-GS more faithfully captures motion and preserves object geometry when observed from novel viewpoints, resulting in visually more coherent reconstructions.

NVIDIA Dynamic Scene Dataset (modified [45]). Since this dataset does not provide reliable covisibility masks, prior works evaluated reconstruction quality on the entire image, often applying post-processing to fill unobserved regions [45]. We argue that this approach is less accurate than restricting evaluation to covisible regions. Accordingly, we generate covisibility masks analytically from depth maps and camera parameters and limit quantitative evaluation to these areas.

Table 4 presents results using this covisibility-based protocol, re-evaluating all prior approaches using the same procedure. On this dataset MOSAIC-GS achieves state-of-the-art performance in both PSNR and LPIPS. Additional qualitative comparisons are provided in the supplementary materials.

Methods	Training Time↓	Rendering Speed↑
Gaussian Marbles*	5-9h	**200 FPS**
Gaussian Flow	23 min	52 FPS
MoSca	50 min	38 FPS
Ours	**10.5 min**	180 FPS

Table 2. Comparison of training time and rendering speed averaged across the DyCheck [12] dataset. *: The values for Gaussian Marbles are taken from the original paper [45]. All other methods were benchmarked on one NVIDIA RTX 4090 GPU. The training time reports the total time including pre-processing, the photometric optimization takes approximately 5 minutes.

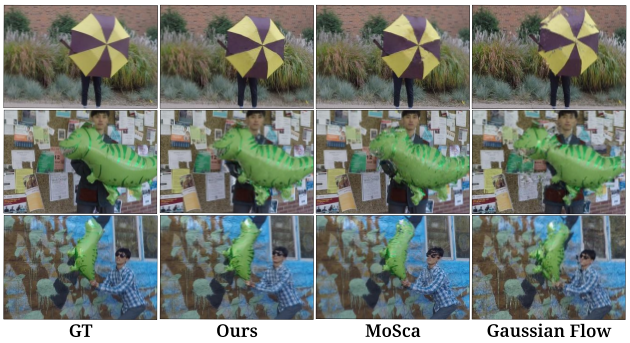

Figure 5. Qualitative comparison of novel view synthesis on the NVIDIA Dynamic Scene (original) dataset. The proposed method better captures motion and preserves objects' geometry when observed from different viewpoints.

Method	PSNR↑	LPIPS↓
Casual-FVS [23]	24.57	0.081
MonoNeRF [49]	25.62	0.106
CTNeRF [35]	26.13	0.082
RoDynRF [32]	25.89	0.067
MoSca [24]	**26.72**	0.070
Ours	26.26	**0.060**

Table 3. Comparison on NVIDIA Dataset (original) [43], averaged over all scenes.

Method	PSNR↑	LPIPS↓	Training Time
Gaussian Flow [29]	21.89	0.095	18 min
Gaussian Marbles [45]	23.68	0.069	3.5-8 h
Ours	**23.79**	0.069	**7 min**

Table 4. Results averaged over all scenes on the Gaussian Marbles variant of the NVIDIA dataset [45]. Our method achieves state-of-the-art quality with a fraction of the training time of Dynamic Gaussian Marbles [45].

4.2. Ablation Study

We conduct an ablation study on the DyCheck dataset [12] to evaluate the impact of key components of

Components	mPSNR↑	mSSIM↑	mLPIPS↓	Time↓
Full model	**18.40**	**0.668**	**0.255**	5.06
No deformations initialization	16.99	0.641	0.298	5.41
No flow refinement	18.11	0.663	0.265	4.83
No scene disentanglement	14.65	0.521	0.456	8.26
Fourier degree 24	18.37	0.667	0.261	4.91
Fourier degree 16	18.29	0.665	0.265	**4.68**
No depth supervision	18.13	0.665	0.264	4.87

Table 5. Ablation study on different components of the system on the DyCheck dataset [12]. The *Time* column reports the training duration of the photometric optimization stage (in minutes).

the MOSAIC-GS framework. Specifically, we investigate the effects of: (1) initialization of mean deformation coefficients for dynamic Gaussians, (2) scene flow refinement using rigid transformations, (3) scene disentanglement into static and dynamic Gaussians, (4) different degrees of Fourier encoding for motion representation, compared to the default degree of 32, and (5) depth supervision during photometric optimization.

As shown in Table 5, each component of our approach improves the reconstruction quality. The largest drop occurs when the scene disentanglement module is removed, leading to poor initialization of static regions and incorrect scene flow refinement. This also significantly increases computational cost, as every Gaussian must store deformation parameters. Removing the initialization of deformation parameters for dynamic Gaussians also substantially reduces visual quality, highlighting the importance of accurate motion priors. Additional scene flow refinement further improves results, especially in scenes with complex motion, where point tracking benefits from extra refinement.

The choice of Fourier degree has a relatively small overall effect, though higher degrees better capture complex motion in detailed, fast-moving scenes like *Wheel* (Fig. 6). Fourier degrees higher than 32 may offer marginal gains in certain cases, but they sacrifice compactness and were therefore not explored further. Lower degrees, on the other hand, provide a more compact motion representation, which can be beneficial in resource-constrained settings.

Finally, adding the depth supervision term further improves reconstruction quality by enhancing both photometric accuracy and geometric consistency.

4.3. Additional Applications

MOSAIC-GS additionally allows temporally consistent scene segmentation and dynamic scene editing without any extra computational cost. During pre-processing, temporal refinement of mask identifiers with point tracking ensures consistent segmentation of dynamic objects over time, avoiding identity swaps that occur with simple 3D reprojection of masks. By assigning segmentation IDs to dynamic Gaussians during initialization, the densification and optimization steps naturally produce temporally consistent segmentation of dynamic regions as a by-product of the reconstruction process.

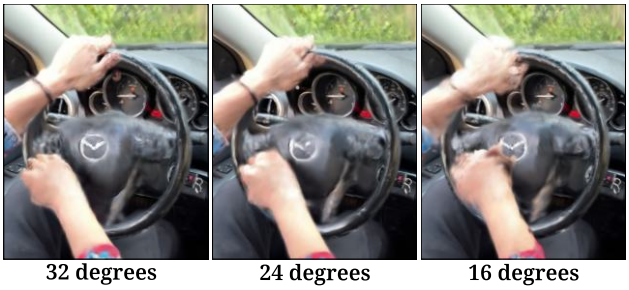

Figure 6. Effect of Fourier degree on the reconstruction of fine details in dynamic regions.

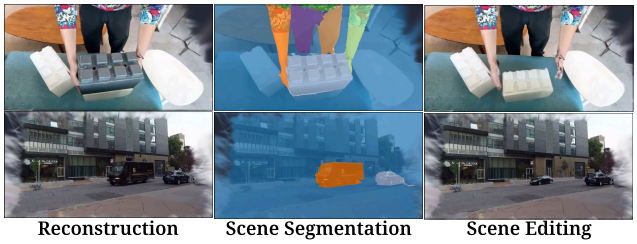

Figure 7. Examples of scene segmentation and editing enabled by the instance IDs stored in dynamic Gaussians. In the top image, the gray cube has been removed from the scene, while in the bottom image, the truck has been removed.

This allows us not only to remove all dynamic objects from the scene, or, conversely, render only the dynamic objects on a blank background, but also to perform several editing operations, such as changing object color, adjusting motion or position, or simply isolating them from the rest of the scene during rendering. Examples of these editing results are shown in Fig. 7

5. Limitations

While MOSAIC-GS achieves efficient and accurate monocular dynamic scene reconstruction, it has several limitations. First, the approach relies heavily on the quality of initial segmentation masks and scene flow, which may inherit errors from the external models. Second, the scene flow refinement may be insufficient, particularly when an entire dynamic object is invisible from the training viewpoint. Further discussion of limitations and future research directions is provided in the supplementary materials.

6. Conclusion

In this work, we presented MOSAIC-GS, a framework for efficient and accurate monocular dynamic scene reconstruction using 3D Gaussian Splatting. Experiments on standard benchmarks show that our method achieves

the state-of-the-art training and rendering efficiency, along with superior perceptual quality (LPIPS) in novel-view synthesis. These improvements stem from an efficient preprocessing stage, including scene disentanglement, scene flow refinement, and compact motion encoding using Poly-Fourier curves. We believe this work takes a meaningful step toward fast dynamic scene reconstruction with high visual fidelity, opening up new possibilities for applications in AR/VR, robotics, and 3D content creation.

Acknowledgment. This work was funded by NCCR Automation, Swiss Federal Railways (SBB) via ETH Mobility Initiative, and ETHAR.

References

[1] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O'Toole, and Changil Kim. HyperReel: High-fidelity 6-DoF video with ray-conditioned sampling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[2] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 130–141, 2023. 3

[3] 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. 2, 4, 6

[4] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10027–10038, 2023. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 2

[6] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14304–14314, 2021. 2

[7] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021. 3

[8] Martin A Fischler and Robert C Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 5

[9] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12479–12488, 2023. 3

[10] Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8645–8654, 2021. 3

[11] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE International Conference on Computer Vision*, 2021. 3

[12] 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. 6, 7, 8

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 2

[15] Nan Huang, Wenzhao Zheng, Chenfeng Xu, Kurt Keutzer, Shanghang Zhang, Angjoo Kanazawa, and Qianqian Wang. Segment any motion in videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 3406–3416, June 2025. 2

[16] Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4220–4230, 2024. 3

[17] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. FlowFormer: A transformer architecture for optical flow. *ECCV*, 2022. 2

[18] Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography*, 32(5):922–923, 1976. 5

[19] 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. 4

[20] 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. 2, 3

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 2

[22] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. 2

[23] Yao-Chih Lee, Zhoutong Zhang, Kevin Blackburn-Matzen, Simon Niklaus, Jianming Zhang, Jia-Bin Huang, and Feng Liu. Fast view synthesis of casual videos with soup-of-planes. In *European Conference on Computer Vision (ECCV)*, 2024. 7

[24] 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. 2, 3, 4, 6, 7

[25] 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. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5511–5521, 2022. 2

[26] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8508–8520, 2024. 3

[27] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[28] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21136–21145, 2024. 2, 3, 5, 6

[29] 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. 6, 7

[30] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Trans. Graph.(ACM SIGGRAPH Asia)*, 2021. 3

[31] Qingming LIU, Yuan Liu, Jiepeng Wang, Xianqiang Lyu, Peng Wang, Wenping Wang, and Junhui Hou. MoDGS: Dynamic gaussian splatting from casually-captured monocular videos with depth priors. In *The Thirteenth International Conference on Learning Representations*, 2025. 6

[32] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3, 7

[33] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 800–809, 2024. 2, 3, 6

[34] 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. 2, 3, 5

[35] Xingyu Miao, Yang Bai, Haoran Duan, Fan Wan, Yawen Huang, Yang Long, and Yefeng Zheng. Ctnerf: Cross-time transformer for dynamic neural radiance field from monocular video. *Pattern Recogn.*, 156(C), Dec. 2024. 7

[36] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *Computer Vision – ECCV 2020*, pages 405–421. Springer International Publishing, 2020. 2

[37] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5845–5854, 2021. 2

[38] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), dec 2021. 2

[39] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10313–10322, 2021. 2

[40] 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. 2, 4, 5

[41] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection, 2015. cite arxiv:1506.02640. 2

[42] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 3

[43] 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. 6, 7

[44] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 3

[45] Colton Stearns, Adam W Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthe-

sis of casual monocular videos. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2, 3, 6, 7

[46] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018. 2

[47] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA, 2014. MIT Press. 2

[48] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 402–419, Cham, 2020. Springer International Publishing. 2, 3

[49] Fengrui Tian, Shaoyi Du, and Yueqi Duan. MonoNeRF: Learning a generalizable dynamic radiance field from monocular videos. In *Proceedings of the International Conference on Computer Vision (ICCV)*, October 2023. 7

[50] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video, 2020. 3

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[52] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. In *International Conference on Computer Vision (ICCV)*, 2025. 2, 3, 4, 6

[53] Chung–Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16189–16199, 2022. 3

[54] 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. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, 2024. 2, 3, 6

[55] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20310–20320, June 2024. 3

[56] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. D2nerf: self-supervised decoupling of dynamic and static objects from a monocular video. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. 3

[57] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3

[58] Tianwei Zhang, Huayan Zhang, Yang Li, Yoshihiko Nakamura, and Lei Zhang. Flowfusion: Dynamic dense rgb-d slam based on optical flow. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7322–7328, 2020. 2

[59] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXXIX*, page 145–163, Berlin, Heidelberg, 2024. Springer-Verlag. 6