Gaussians on Fire: High-Frequency Reconstruction of Flames

Jakob Nazarenus[1], Dominik Michels[2], Wojtek Palubicki[3], Simin Kou[4], Fang-Lue Zhang[4], Sören Pirk[1], and Reinhard Koch[1]

[1] Kiel University, Germany {jna, sp, rk}@informatik.uni-kiel.de
[2] KAUST, CEMSE, Saudi Arabia dominik.michels@kaust.edu.sa
[3] Adam Mickiewicz University, Poland Wojciech.Palubicki@amu.edu.pl
[4] Victoria University of Wellington, New Zealand {simin.kou, fanglue.zhang}@vuw.ac.nz

Abstract. We propose a method to reconstruct dynamic fire in 3D from a limited set of camera views with a Gaussian-based spatiotemporal representation. Capturing and reconstructing fire and its dynamics is highly challenging due to its volatile nature, transparent quality, and multitude of high-frequency features. Despite these challenges, we aim to reconstruct fire from only three views, which consequently requires solving for under-constrained geometry. We solve this by separating the static background from the dynamic fire region by combining dense multi-view stereo images with monocular depth priors. The fire is initialized as a 3D flow field, obtained by fusing per-view dense optical flow projections. To capture the high-frequency features of fire, each 3D Gaussian encodes a lifetime and linear velocity to match the dense optical flow. To ensure sub-frame temporal alignment across cameras, we employ a custom hardware synchronization pattern – allowing us to reconstruct fire with affordable commodity hardware. Our quantitative and qualitative validations across numerous reconstruction experiments demonstrate robust performance for diverse and challenging real and simulated fire scenarios.

Keywords: Gaussian Splatting · 3D Fire Reconstruction · Sparse View Reconstruction

1 Introduction

In recent years, differentiable scene representations have enabled remarkable progress in reconstructing static and dynamic 3D environments. In particular, Gaussian Splatting [43] approaches have gained popularity due to their good performance and efficient reconstruction run-times. However, reconstructing highly dynamic objects remains challenging due to rapid motion, appearance changes, and the lack of consistent geometry across time. Capturing fire, in particular, is uniquely difficult because of its non-rigid structure, high-frequency appearance, and rapid temporal variation. Current Gaussian Splatting-based approaches fundamentally fail when applied to fire, a limitation we verify across different experiments. We compare against state-of-the-art dynamic GS methods, including 4D Gaussian Splatting [92, 105] and FreeTimeGS [90], and observe that

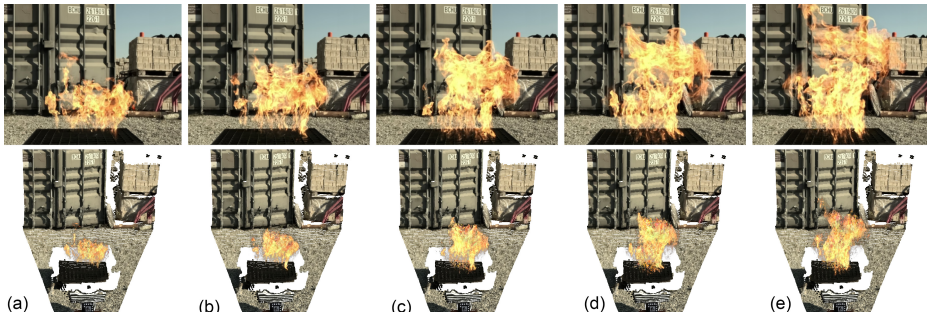

Fig. 1: Temporal progression of a flame encoded as 3D Gaussian splats. The sequence shows the evolution from an initial small flame (a, b) to a fully developed flame (c, d, e) as the rendered 3D Gaussian splats capturing volumetric radiance and motion (top row) and the corresponding point clouds highlighting the spatial distribution and density of the Gaussian primitives over time (bottom row).

all baselines produce degenerate reconstructions for flames: geometry collapses, motion becomes inconsistent, and rendered depth fields are spatially unstable. These methods either blur away the flame structure or incorrectly attribute it to background geometry, resulting in poor reconstruction quality in dynamic regions (Fig. 5 and Tab. 1). Across these baselines, we observe three recurring failure modes: (i) geometric absorption, where flame motion is explained by deforming the background; (ii) depth collapse, where plausible image appearance is obtained without a meaningful 3D structure; and (iii) temporal drift, where the rendered flame no longer matches the observed dynamic evolution. Our method is designed specifically to avoid these three modes. This failure stems from a shared assumption across existing GS methods – that scene content can be represented as persistent primitives undergoing smooth, trackable motion. Fire violates this assumption entirely: it is transient, lacks consistent correspondences, and exhibits strongly view-dependent emission that cannot be explained by surface-based radiance models. Consequently, no existing Gaussian Splatting method is able to reliably reconstruct fire from real-world observations, particularly under sparse-view conditions. This exposes a critical gap in current discrete dynamic Gaussian scene representations and motivates the need for a formulation specifically designed for highly dynamic, emissive, and volumetric phenomena. The key representational change in our approach is that we do not treat flames as persistent scene elements undergoing deformation. Instead, we model fire as a population of transient volumetric carriers of emission whose motion and temporal support must be inferred jointly. This shift in the unit of representation is what makes sparse-view reconstruction of flames feasible within a Gaussian Splatting framework.

We introduce a pipeline that reconstructs high-frequency spatiotemporal flame dynamics with 3D Gaussian Splats. As it is challenging to capture fire in real-world environments, our method targets a highly constrained, sparse-view

setup. Specifically, we present an approach for capturing fire with only three calibrated cameras. Our approach is based on the idea to first decouple the static scene from the flames. We then initialize Gaussian splats for fire based on a 3D flow field fused from per-view dense optical flow. Additionally, we use an explicit temporal parameterization – each Gaussian carries a lifetime and linear velocity – which enables motion priors (e.g., upward bias, smoothness) under sparse views. We refer to these dynamic primitives as transient flame Gaussians: short-lived volumetric carriers of emission whose finite temporal support is essential for modeling fire. This flow-initialized representation couples naturally with efficient Gaussian rasterization, yielding stable optimization and high-fidelity renderings despite the severe view sparsity.

Our method reconstructs temporally coherent 4D representations of fire that capture both its volumetric appearance and dynamic motion from sparse-view video input. Given as few as three calibrated and synchronized cameras, the pipeline jointly estimates a static background scene and a time-varying Gaussian field representing the flames. The resulting model allows photo-realistic novel view synthesis, temporal re-rendering at arbitrary time steps, and explicit analysis of motion trajectories through the per-Gaussian velocity field. In practice, this enables stable reconstruction across diverse flame types – from small localized flames to larger outdoor fires – without requiring dense camera arrays or simulation-based priors. Importantly, our formulation is intentionally minimal: rather than introducing a heavy physics simulator or a fully new rendering backbone, we show that rethinking decomposition, motion initialization, and temporal support is sufficient to make sparse-view fire reconstruction tractable.

A result of our method is shown in Figure 1. In summary, our work makes the following contributions: (1) We identify sparse-view fire reconstruction as a failure case for existing dynamic GS methods and analyze this failure in terms of their persistent-primitive and smooth-motion assumptions; (2) We introduce a transient Gaussian reconstruction formulation that decouples static scene geometry from short-lived emissive flame structures and initializes the latter from fused 3D motion evidence; (3) We propose a practical sparse-view capture and synchronization setup for high-speed real fire acquisition with consumer hardware; (4) We establish an evaluation protocol for dynamic fire reconstruction combining frame-wise appearance, depth consistency, motion plausibility, and video-level temporal quality, and show that prior GS methods fail under this protocol while our method remains effective.

2 Related Work

Building on a long history of physically based fire simulation [67], we address the inverse problem of recovering 3D geometry and appearance from sparse observations. Accurate correspondence and tracking form the basis for such reconstruction, linking observed dynamics to coherent 3D representations with Gaussian splats and connecting to broader efforts in long-range point tracking and 4D dynamic reconstruction [18–20, 26, 32, 42, 45, 64, 66, 81, 87, 95, 96, 98, 109].

The following sections review recent methods in correspondence estimation, view synthesis, and dynamic scene modeling that are related to our approach.

Correspondence and Tracking. While monocular long-range *3D* tracking remains relatively underexplored, there is a rich literature on tracking in 2D image space via dense or semi-dense correspondence estimation. Classical and modern optical-flow style methods estimate dense motion fields between frames [5, 10, 11, 33, 60], paving the way for multi-frame and occlusion-aware variants [3, 36, 40, 41, 73, 79] including high-resolution multi-frame approaches such as MEMFOF [3], which we use in our approach. Sparse keypoint and descriptor methods provide more stable, but sparser, correspondences that can be integrated into SfM pipelines [4, 17, 56, 58, 74, 76, 78], enabling long-term tracks in relatively rigid scenes [2, 30].

Moving from 2D to 3D correspondences, recent work performs geometric matching and reconstruction directly from images: structure-from-motion and MVS pipelines [76, 78], DUSt3R and MASt3R [49, 89], and the fully integrated MASt3R-SfM system [24] provide accurate geometric correspondences in static or slowly changing scenes. Depth priors such as Depth Anything [102,103], Depth Pro [7], and Marigold-DC [83] improve per-frame geometry and therefore implicit tracking, but they do not explicitly maintain persistent long-range 3D trajectories. To incorporate geometry, recent work in monocular settings combines optical flow, depth, and rigidity for self-supervised scene-flow estimation and dynamic reconstruction [36, 44, 53–55, 59, 62, 72, 107, 108]. Feedforward 3D tracking methods estimate trajectories directly in 3D space or focus on rigid reconstruction [26, 66, 95, 96, 98].

View Synthesis and Static 3D Reconstruction. Neural Radiance Fields (NeRF) introduced continuous volumetric radiance fields for view synthesis from sparse posed images and became the de facto standard for static neural scene representations [63]. A broad line of work has extended radiance fields to handle video and dynamic content, including free-viewpoint video [12, 28, 51, 94, 106], articulated humans [14, 39, 50, 91], and factorized space–time–appearance representations [13, 22, 27, 80, 86].

3D Gaussian Splatting (3DGS) replaces volumetric ray marching with a set of view-dependent anisotropic Gaussians rendered by a differentiable splatting pipeline, achieving real-time, high-quality novel-view synthesis [43]. Some approaches adapt the formulation to 2D Gaussian primitives [34], controllable or editable splats [35], and more general camera and light-transport models with distorted cameras and secondary rays [93]. These Gaussian-based approaches complement geometric 3D reconstruction methods such as DUSt3R and MASt3R, and their SfM integration [24, 49, 89], as well as multi-view dynamic reconstruction systems [8, 21, 38, 65, 111].

Dynamic 3D Reconstruction. Dynamic scene modeling extends these representations into space-time for 4D reconstruction and animation. Before neural fields, non-rigid reconstruction often relied on RGB-D sensors [8,21,38,65,111] or strong low-rank and NRSfM priors [9, 16, 47, 68, 75]. NeRF-style dynamic reconstructions deform a canonical radiance field using learned deformation fields or

higher-dimensional embeddings [22,28,51,54,55,69–71,80,84,94,106], while articulated and category-specific methods focus on humans and animals [39,50,53,99–101]. Physics-informed neural fields and tomography-based methods reconstruct phenomena such as smoke and flames from sparse observations [15,29,37], but assume controlled multi-view setups or simplified motion models. A closely related prior work applies a temporal Laplacian pyramid to blend per-view observations into a free-viewpoint representation of dynamic scenes such as flames [82]. While rendering qualitatively compelling results, this approach does not target a full spatiotemporal reconstruction of the underlying geometry and appearance.

More recently, dynamic 3D perception models aim to couple tracking and reconstruction over long temporal horizons [26,30,45,57,59,66,88,95,96,98]. Building on Gaussian Splatting, several works propose explicitly dynamic Gaussian representations: GS4D and 4D Gaussian Splatting learn 4D Gaussian primitives for real-time photorealistic dynamic view synthesis [92,105]; 4D-Rotor Gaussian Splatting introduces rotational dynamics for efficient dynamic scenes [23]; spacetime Gaussian feature splatting embeds Gaussians in a joint space–time feature field [52]; and SC-GS provides sparse controls for editable dynamic scenes [35]. Deformable 3D Gaussians and per-Gaussian embedding methods model continuous canonical-space deformations and per-primitive latents [1,104]. Dynamic 3D Gaussians attach per-frame rigid transforms to canonical Gaussians for tracking and view synthesis [61]. Latent-dynamical and monocular setups are further explored by ODE-GS [85], FreeTimeGS [90], MoSca [48], DynMF [46], and 4DGT [98], which respectively introduce latent ODEs, free-time Gaussian primitives, 4D motion scaffolds, motion factorization, and transformer-based 4D Gaussians for real-world monocular videos. RGB-D based methods [31] produce compelling results on a variety of dynamic real-world scenes but rely on a well-defined surface, which semi-transparent phenomena such as fire and smoke cannot provide.

3 Methodology

As real fire is usually difficult to capture from many viewpoints, our main objective is to use as few cameras as possible for the reconstruction. However, only using a few cameras leads to insufficient view constraints for reconstructing scenes with rendering losses alone. To address this, we separately reconstruct the static scene (background) and the dynamic scene (fire). This enables us to use a view-consistent, rigid representation for the background and a volumetric, dynamic representation for the flames. For the static scene, we use vanilla 3DGS, circumventing the insufficient view constraints by leveraging a dense stereo initialization in combination with monocular depth regularization during optimization. For the dynamic scene, we extend the 3DGS framework with a flow-based initialization that estimates volumetric motion from dense optical flow, enabling the reconstruction of temporally coherent, plausible fire dynamics. All modalities used by our proposed method are depicted in Fig. 3, while Sec. 2 of the Supp. Mat. visualizes the decomposed static/dynamic scene.

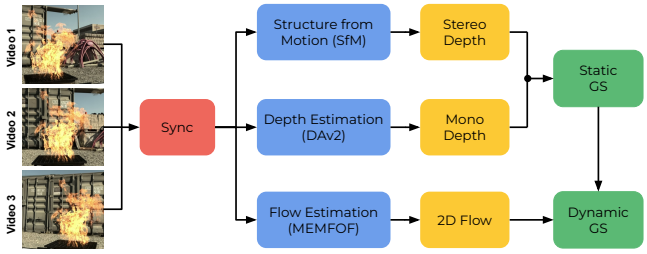

Fig. 2: Pipeline overview: Given the video of a fire from three views, we first calibrate and register cameras using COLMAP and estimate dense stereo depth maps (PatchMatchStereo), complemented by aligned monocular depth predictions for regularization. The static background scene is optimized with vanilla 3DGS using this fused depth initialization. Dynamic regions such as fire are reconstructed by projecting dense optical flow into a 3D voxel grid to estimate a volumetric motion field, which initializes our transient flame Gaussians with position, motion, and temporal support.

3.1 Static Scene

To reconstruct the static scene, we first obtain frames for each view that do not contain any motion. When such frames are not available, we fall back on deriving a background estimate directly from the video sequence. As the flames are actively emitting light, they are usually brighter than their background, allowing us to remove them with a minimum intensity projection along the time dimension (see Sec. 1 of the Supp. Mat.). Based on these static frames, we estimate accurate camera poses using COLMAP [77]. To reduce overfitting caused by limited view coverage, we employ two key strategies: (1) a strong point-cloud initialization, and (2) a continuous monocular regularization term. In order to obtain a dense initial geometry, we apply COLMAP's implementation of PatchMatchStereo [6, 77] to the calibrated and registered static frames and obtain a stereo depth map D_{stereo} for each input image. To initialize regions of the scene observed by at most a single camera, we predict monocular depth maps D_{mono} using DepthAnythingV2 [103]. For each camera, we then perform a linear alignment $D'_{\text{mono}} = aD_{\text{mono}} + b$, where the parameters $a, b \in \mathbb{R}$ are determined by minimizing the least-squares difference

$$(a^*, b^*) = \underset{a,b\in\mathbb{R}}{\arg\min} \|aD_{\text{mono}} + b - D_{\text{stereo}}\|_2^2 \tag{1}$$

over overlapping regions. We create an initial point cloud by back-projecting the three stereo depth maps. In regions with no stereo information available, we utilize the aligned monocular depth maps to fill the scene (Fig. 3, b). This point cloud then serves as the initialization for a 3DGS reconstruction of the static scene. During training, the aligned monocular depths D'_{mono} are incorporated as a regularization term

$$\mathcal{L}_{\text{depth}} = \frac{1}{wh} \sum_{i,j} \|D'_{\text{mono}}(i,j) - D_{\text{render}}(i,j)\|_1 \ , \tag{2}$$

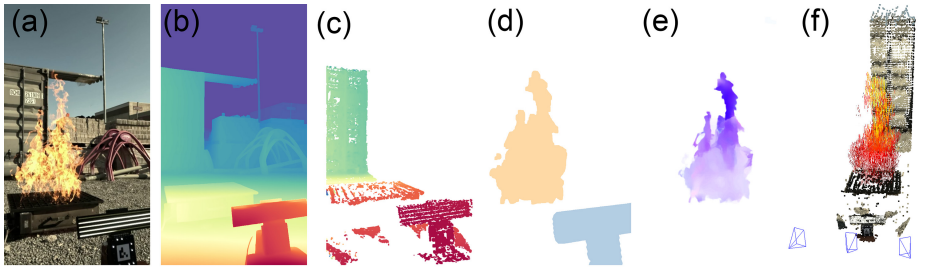

Fig. 3: Intermediate data used for scene reconstruction: (a) Input RGB frame, (b) monocular depth prediction, (c) depth from structure-from-motion (SfM), (d) segmentation mask of the dynamic region, (e) estimated 2D optical flow, and (f) fused voxelized 3D flow field used to initialize dynamic Gaussians.

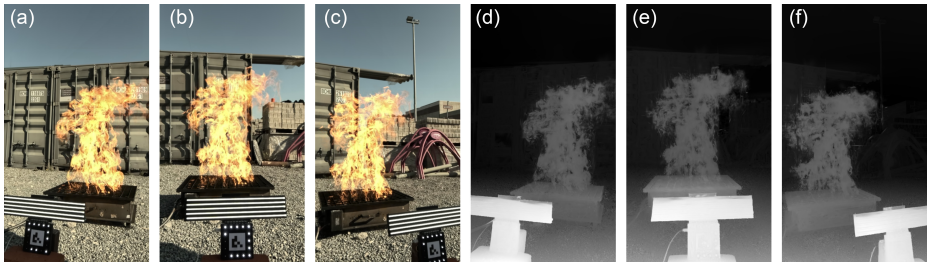

Fig. 4: Multi-view capture of a fire: we capture the RGB images from the three camera viewpoints (a)-(c) and show the corresponding rendered depth maps (d)-(f). The depth reconstructions demonstrate consistent geometry across views despite strong appearance variations caused by the dynamic flames.

to prevent degenerate depth solutions and preserve global consistency. The overall loss is given as the weighted sum of the absolute rendering loss $\mathcal{L}_1$, the SSIM loss $\mathcal{L}_{\text{SSIM}}$, as well as the depth regularization term $\mathcal{L} = \lambda_1\mathcal{L}_1 + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}}$. We utilize standard parameter values $\lambda_1 = 0.8$ and $\lambda_{\text{SSIM}} = 0.2$. To strengthen the depth regularization (λ_{depth}), we increase the default regularization parameter value by a factor of 100, yielding an exponentially decreasing contribution with an initial value of 100 and a final value of 1 [43].

3.2 Dynamic Scene

For the dynamic scene, we represent the flame as a set of transient flame Gaussians, i.e., short-lived volumetric primitives whose appearance and motion are optimized only over a finite temporal support. Reconstructing the dynamic component with just dense stereo is not possible for two main reasons: (1) our commodity camera frames are not synchronized, and (2) it does not support obtaining a continuous motion field. Instead, to address this, we first estimate an approximate flow field, which is subsequently refined through joint optimization guided by the multi-view camera observations.

We estimate the initial flow field by computing the dense optical flow $\mathbf{f}_i(u,v) = (f_{i,u}(u,v), f_{i,v}(u,v))$ for each pixel (u,v) in the image plane of camera i, where $\mathbf{f}_i(u,v)$ represents the 2D motion vector at each pixel. The flow vector, which estimates the displacement of each pixel between consecutive frames, is estimated using MEMFOF [3]. Once the flow is computed, we project all 2D flows onto a voxel grid. For each point $\mathbf{x} \in \mathbb{R}^3$, the corresponding 2D flows $\mathbf{f}_i(\pi_i(\mathbf{x}))$ are back-projected to 3D space

$$\mathbf{u}_i(\mathbf{x}) = \pi_i^{-1}\left(\mathbf{f}_i(\pi_i(\mathbf{x})) + \pi_i(\mathbf{x}), \mathbf{x}\right) - \mathbf{x}\,, \tag{3}$$

where $\pi_i^{-1}(\ldots, \mathbf{x})$ is the back-projection function to the depth of $\mathbf{x}$. For each camera, the 3D flow $\mathbf{F}(\mathbf{x}) \in \mathbb{R}^3$ must adhere to the projection constraint

$$\left(\mathbf{F}(\mathbf{x}) - \mathbf{u}_i(\mathbf{x})\right)^{\mathsf{T}} \mathbf{u}_i(\mathbf{x}) = 0\,. \tag{4}$$

To avoid unstable solutions for under-constrained systems, we apply a Tikhonov regularization with the regularization parameter

$$\tau(\mathbf{x}) = \frac{\tau_0}{m} \sum_i^m \|\mathbf{u}_i(\mathbf{x})\|_2^2\,, \tag{5}$$

where m denotes the number of cameras and τ_0 controls the strength of the regularization. The 3D flow is finally determined by solving the least-squares system[5]

$$\arg\min_{\mathbf{F}} \left\| \begin{bmatrix} \mathbf{u}_1^\top \\ \vdots \\ \mathbf{u}_m^\top \end{bmatrix} \mathbf{F} - \begin{bmatrix} \mathbf{u}_1^\top \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_m^\top \mathbf{u}_m \end{bmatrix} \right\|_2^2 + \tau^2 \|\mathbf{F}\|_2^2\,, \tag{6}$$

which is solved via singular value decomposition and completes the preprocessing stage for the dynamic scene.

For modeling the dynamic flames, we chose a constant-velocity representation inspired by FreeTimeGS [90], which assigns three additional parameters to each Gaussian: a time t_μ, a lifespan t_σ, and a linear velocity $\mathbf{v} \in \mathbb{R}^3$. At any given time, the position of a dynamic Gaussian is determined by $\mathbf{x}(t) = \mathbf{x}_0 + (t - t_\mu)\mathbf{v}$. Additionally, its opacity is modulated by the temporal modifier $\sigma(t) = \exp(-\frac{1}{2}(\frac{t-t_\mu}{t_\sigma})^2)$. This representation has two key advantages: (1) it can be directly initialized from a 3D flow field, and (2) it exposes its velocity field explicitly, which enables a variety of downstream tasks for interpreting fire (e.g., robot navigation).

Given the dynamic representation, for each position $\mathbf{x}$ in the voxel grid at time t, we assign a Gaussian with velocity $\mathbf{F}(\mathbf{x})$ and time $t_\mu = t$. The Gaussian's lifetime is initialized as four times the frame time for smooth motion and to prevent overfitting to single frames. To account for voxel discretization, we introduce a random offset in the position of the Gaussian, sampled uniformly

[5] $\mathbf{F}(\mathbf{x})$, $\mathbf{u}_i(\mathbf{x})$, and $\tau(\mathbf{x})$ are abbreviated as $\mathbf{F}$, $\mathbf{u}_i$, and τ.

from the voxel volume, with $\Delta\mathbf{x} \sim \text{Uniform}([-s/2, s/2]^3)$, where s is the voxel size, to avoid moiré patterns. A multi-view capture of fire is shown in Fig. 4.

To further mitigate overfitting of the dynamic Gaussians to the static background, during training we compare the opacity of the dynamic Gaussians A_{dyn} to a precomputed motion mask M and compute an additional *alpha loss* as $\frac{1}{wh}\sum A_{\text{dyn}}(1 - M)$ with a weight of 0.1, which discourages placing dynamic Gaussians in regions where no flames have been detected in the input video.

3.3 Visual Synchronization

To capture high-frequency phenomena such as fire from multiple viewpoints, precise temporal synchronization is essential. However, most consumer-grade hardware lacks native support for high-frame-rate synchronization. To address this, we design a low-cost synchronization system based on a microcontroller. Specifically, we use an ESP32 microcontroller to control the logic-level outputs of 21 pins. These pins are divided into two groups: five pins generate a sub-frame synchronization pattern, while the remaining 16 pins manage an incrementing frame counter. The frame counter is displayed using 15 white LEDs, which represent the frame number in Gray code. An additional LED is employed to resolve the inherent ambiguity of the Gray code, preventing frame errors that could otherwise reach up to one frame. The synchronization pattern is shown in Fig. 3a.

To achieve sub-frame timing with microsecond-level precision, we exploit the pronounced rolling shutter effect present in most CMOS consumer cameras by sequentially toggling a series of five COB LED strips. The rolling shutter becomes clearly visible by observing the partially illuminated LED strips. For a fully characterized sensor, detecting the image points where the LED strip brightness changes allows us to directly determine the exposure time. The temporal precision is directly dependent on the row time t_{row} and the spatial accuracy Δb in detecting the change in LED strip brightness. For typical values of $t_{\text{row}} \approx 3\,\mu\text{s}$ and $\Delta b \approx 5\,\text{px}$, we achieve a temporal precision of approximately $15\,\mu\text{s}$. For an explicit representation of the rolling shutter within our proposed method, we refer to Sec. 3 of the Supp. Mat.

3.4 Velocity Metrics

In reconstruction tasks, the most commonly used metrics such as PSNR, LPIPS, and SSIM are used to quantitatively assess similarity of 2D images. However, these metrics cannot directly be applied to assess similarity of dynamic 3D scenes. For dynamic similarity evaluation, some methods render 2D optical flow from the scene and evaluate this against estimated optical flow [110], while others observe the motion of the Gaussian centers and compare these against known trajectories [61]. Neither approach is feasible for our case. Comparing against 2D optical flow would introduce a circular dependency into our pipeline, and would rely on estimations of the real flow. For the second approach, we determined that tracking the center position of the Gaussians is insufficient to capture their full dynamics, as each Gaussian represents a spatial distribution and varying

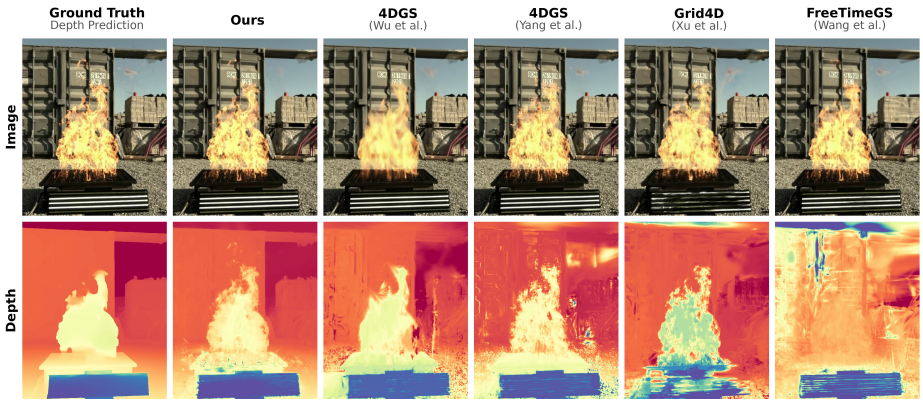

Fig. 5: Qualitative comparison of our method against 4DGS [92], 4DGS [105], Grid4D [97], and FreeTimeGS [90]. We provide the ground truth image as well as the rendered results from all methods. To evaluate the plausibility of the rendered depth, we provide a monocular depth estimation for comparison [103]. While other methods struggle with degradation of visuals or depth, our method provides rendered images of high visual quality while avoiding depth inaccuracies.

opacity and shape. Therefore, we propose a novel metric for assessing similarity of 3D GS reconstructed scenes of transient phenomena – in our case flames. Specifically, we employ synthetic scenes to evaluate the motion reconstruction capability of our model, as these provide, in addition to the dynamics, the underlying mechanism – the velocity field – to be used for evaluation, which we refer to as the density-weighted volumetric evaluation protocol (DVE). For every scene, we obtain the N optimized Gaussians with their center position $\mathbf{x}$, opacity α, scale $\mathbf{s}$, velocity $\mathbf{v}$, and rotation R. Based on this, we compute the Gaussian's weight $w = s_x s_y s_z \alpha$. Given the normalized Gaussian probability distribution $P(\mathbf{x}, \mathbf{s}, R)$ and the underlying ground truth velocity field V, we approximate the expected value of the ground truth velocity function $\bar{\mathbf{v}} = E_P(V)$ by employing a third-order Gauss-Hermite quadrature. We capture the overall similarity of the two motion fields using a relative L2 metric

$$\mathrm{L2} = \sqrt{\frac{\sum_{n=1}^{N} w_n \|\mathbf{v}_n - \bar{\mathbf{v}}_n\|^2}{\sum_{n=1}^{N} w_n \|\bar{\mathbf{v}}_n\|^2}} \,, \tag{7}$$

and measure directional alignment via weighted volumetric cosine similarity

$$\mathrm{CosSim} = \frac{\sum_{n=1}^{N} w_n \frac{\mathbf{v}_n \cdot \bar{\mathbf{v}}_n}{\|\mathbf{v}_n\| \|\bar{\mathbf{v}}_n\|}}{\sum_{n=1}^{N} w_n} \,. \tag{8}$$

Together, these metrics yield a density-aware evaluation protocol capturing the quantitative accuracy and directional fidelity of the reconstructed motion field.

4 Dataset

We record a real-world dataset of fire sequences using three *GoPro Hero 13 Black* cameras arranged in a fixed multi-view setup. Each captures at 400 Hz, which provides a high-frequency temporal resolution necessary for modeling the rapid motion and radiative changes of fire. We use three cameras as this enables multi-view cross-validation for robust COLMAP points, which provide a fair initialization for the baseline methods. Each sequence spans approximately 15 s. Furthermore, we captured fire against different backgrounds to evaluate reconstruction robustness under varying lighting and texture conditions.

To capture diverse flame dynamics and appearance, we include several fire materials such as propane jets, burning wood, gasoline, cardboard, and paper. In total, we captured 17 videos of fire for each camera. For each of these scenes, we created a subsequence containing 100 frames per camera. This setup provides a controlled yet challenging dataset for testing our monocular and multi-view fire reconstruction approach. Code and data will be released with the paper.

To compute the DVE similarity between reconstruction and ground truth and compare our method against existing baselines in the GS domain, we simulated 3 fire scenes in *Mantaflow*. Specifically, we created a setup of three cameras observing fires, placed within three *Blender* scenes. Aside from 50 ray-traced frames per camera over 2 seconds of simulation, this simulation provides us with a ground truth velocity field with a resolution of 640^3, which enables our DVE similarity assessment. In total, we generated 150 synthetic velocity fields. For further details on both datasets, refer to Sec. 4 of the Supp. Mat.

5 Results

To evaluate our proposed method on the real-world fire dataset, we held out every 8th frame for validation. As shown in Fig. 5, our method reconstructs the captured scenes with high visual fidelity while preserving the spatiotemporal evolution of the flames. Qualitative results for the additional scenes across a wide range of materials, including wooden logs, cardboard, and paper, are included in Sec. 5 of the Supp. Mat., while additional videos clarify the recovered spatiotemporal behavior, in particular the temporal coherence of the flames and the failure modes of the baseline methods (see Sec. 11 of the Supp. Mat.).

For a quantitative evaluation, we apply 4DGS (Wu et al.) [92], 4DGS (Yang et al.) [105], Grid4D [97], and FreeTimeGS [90] as established baseline methods. Aside from FreeTimeGS, which applies its own RoMa initialization [25], all other baseline methods were initialized from COLMAP points. We used the author-provided hyperparameter configurations for the DyNeRF *flame salmon* scene. To compare the overall visual quality, we report PSNR, SSIM, and LPIPS. For the more local metrics PSNR and SSIM, we mask out the synchronization pattern. We limit the quantitative comparison to GS-based baselines because they are the only class of methods that operate in the same efficiency regime targeted in this work. Furthermore, we provide additional masked variants $\mathrm{PSNR}_{\mathrm{flame}}$

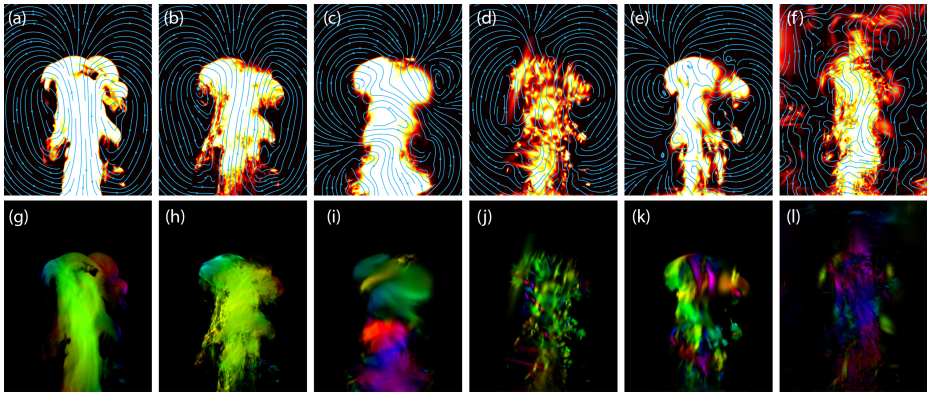

Fig. 6: Stream plot comparisons between maximum projected Gaussian velocities of simulated ground truth (a), our reconstruction (b), Grid4D [97] (c), 4DGS [105] (d), 4DGS [92] (e), FreeTimeGS [90] (f). The streamlines indicate the flow of the Gaussians and the colors indicate their density. The second row (g-l) depicts the color-coded Gaussian velocities.

and $\text{SSIM}_{\text{flame}}$ that only evaluate reconstruction quality in those regions of the image that contain motion. To evaluate the plausibility of the rendered depth of the flame, we compare it against a predicted monocular depth after linear alignment and report the score as $\text{RMSE}_{\text{depth}}$. A central methodological point of our evaluation is that frame-wise image metrics alone are insufficient for dynamic fire reconstruction: methods with partially competitive PSNR can still fail catastrophically in 3D structure and temporal behavior. Taken together, the image, depth, video, and velocity metrics allow us to distinguish between methods that merely match frame appearance and methods that recover a temporally and geometrically meaningful 4D reconstruction. For a novel-view evaluation of our method on synthetic scenes with spatially held-out views, we refer to Sec. 7 of the Supp. Mat. As shown in Tab. 1, our proposed method outperforms all baselines in all visual metrics for both the flame and the full image, as well as in the plausibility of the rendered depth. FreeTimeGS achieves second-best flame reconstruction results, but only at the cost of a high degree of depth degradation. 4DGS [105] achieves second-best full-frame visual metrics. Further video metrics yield similar results and are provided in Sec. 6 of the Supp. Mat.

To evaluate the reconstructed motion, we apply our proposed DVE protocol to the synthetic scenes. We report the weighted volumetric relative L2 metric and cosine similarity. The results are reported in Tab. 2. While the comparison methods mostly fail in reconstructing physically plausible motion, our method shows a comparatively high similarity between the reconstruction and the ground truth flow field. Especially for the directional similarity, the baselines show largely weak alignment with the true flow field, while our velocities are mostly correctly aligned. This is further visualized in Fig. 6, which depicts an X-slice of the reconstructed fire simulation.

Table 1: Quantitative comparison and ablations for the real-world dataset. Our method outperforms all baselines in the visual metrics for the reconstruction of the fire as well as in the accuracy of the predicted flame depth.

Method	$\text{PSNR}_{\text{flame}}$ (↑)	$\text{SSIM}_{\text{flame}}$ (↑)	$\text{RMSE}_{\text{depth}}$ (↓)	LPIPS (↓)	PSNR (↑)	SSIM (↑)
Ours	26.93 ± 1.93	0.843 ± 0.048	0.094 ± 0.038	0.035 ± 0.011	35.89 ± 2.62	0.967 ± 0.011
4DGS [105]	23.28 ± 1.76	0.728 ± 0.069	0.113 ± 0.045	0.053 ± 0.020	32.41 ± 2.86	0.945 ± 0.021
4DGS [92]	20.99 ± 1.14	0.601 ± 0.054	0.102 ± 0.036	0.071 ± 0.025	31.88 ± 3.12	0.933 ± 0.025
FreeTimeGS	24.48 ± 3.48	0.756 ± 0.132	0.131 ± 0.042	0.241 ± 0.131	25.62 ± 1.83	0.794 ± 0.109
Grid4D	18.52 ± 1.47	0.497 ± 0.068	0.225 ± 0.060	0.086 ± 0.025	29.16 ± 3.33	0.909 ± 0.032
Full Method	26.93 ± 1.93	0.843 ± 0.048	0.094 ± 0.038	0.035 ± 0.011	35.89 ± 2.62	0.967 ± 0.011
NoSync	26.89 ± 2.00	0.842 ± 0.049	0.094 ± 0.034	0.035 ± 0.011	35.91 ± 2.67	0.967 ± 0.011
RandFlow	26.68 ± 2.43	0.812 ± 0.076	0.089 ± 0.039	0.036 ± 0.013	35.67 ± 3.21	0.963 ± 0.017
JointOpt	26.52 ± 1.89	0.836 ± 0.048	0.106 ± 0.038	0.060 ± 0.014	33.59 ± 1.89	0.948 ± 0.011
NoMask	27.34 ± 1.98	0.849 ± 0.049	0.098 ± 0.038	0.034 ± 0.011	36.28 ± 2.72	0.968 ± 0.011

Table 2: Quantitative comparison for the synthetic fire scene. Our method yields the most physically plausible reconstruction with well aligned velocities, while the baselines only show a weak directional correlation with the ground truth.

Method	L_2 (↓)	CosSim (↑)	Ablation	L_2 (↓)	CosSim (↑)
Ours	0.825 ± 0.058	0.744 ± 0.032	**Full Method**	0.825 ± 0.058	0.744 ± 0.032
4DGS (Yang et al.)	1.045 ± 0.194	0.225 ± 0.123	**NoSync**	0.838 ± 0.058	0.736 ± 0.034
4DGS (Wu et al.)	2.218 ± 0.276	0.036 ± 0.010	**RandFlow**	1.519 ± 0.026	0.338 ± 0.050
FreeTimeGS	1.164 ± 0.042	0.194 ± 0.114	**JointOpt**	0.820 ± 0.063	0.749 ± 0.033
Grid4D	2.262 ± 1.470	0.154 ± 0.014	**NoMask**	0.850 ± 0.059	0.697 ± 0.035

To demonstrate the impact of the individual components of our approach, we include several ablations over the timing, the flow field initialization, the alpha loss, and the separation of static and dynamic optimization. The evaluation is reported in Tab. 1 for the real-world scenes and in Tab. 2 for the synthetic scenes. Introducing timing offsets of half a frame-time (*NoSync*) results in reduced visual quality in the flames for the real-world experiments and a reduction in the accuracy of the reconstructed motion field for the simulated scenes. The same holds for the omission of our flow-based initialization, which is evaluated by initializing the dynamic Gaussians randomly (*RandFlow*). This indicates that without proper initialization, the in-training optimization of Gaussian velocities is insufficient to find a faithful approximation of the underlying volumetric flow just from photometric supervision. Similarly, omitting the alpha loss causes significant degradation of the reconstructed motion fields while improving the overall visual results (*NoMask*). The last ablation thus demonstrates a delicate trade-off between visual fidelity and physical accuracy: applying the alpha loss removes the method's ability to place dynamic Gaussians in static parts of the scene, thus improving the physical accuracy while simultaneously reducing the visual capacity. In contrast, *JointOpt*'s degradation is concentrated in full-frame fidelity, indicating that this ablation primarily shows the static-pretraining's contribution to the full-frame visuals rather than being a physical-visual trade-off.

6 Discussion and Limitations

Our results show that a dynamic Gaussian representation can be adapted to a regime where existing methods fail: sparse-view reconstruction of transient, emissive, semi-transparent flames. By combining static stereo-monocular depth fusion with dynamic flow-based initialization and masked optimization, our method recovers reconstructions that are not only visually faithful but also substantially more consistent in depth and motion. The resulting pipeline enables efficient reconstruction of high-speed fire events using consumer-grade hardware and lightweight synchronization. Despite these strengths, several limitations remain. For one, the reliance on dense optical flow introduces inaccuracies in regions of strong emissive variation or low texture, which can propagate into the 3D motion field. The failure mode of flames missed by the flow estimator during the flow-based initialization cannot be recovered during the dynamic optimization. Second, the current linear velocity model may not fully capture the complex, turbulent motion of real flames, leading to temporal blurring or oversmoothing. The specific failure modes are discussed further in Sec. 10 of the Supp. Mat.

Another theoretical limitation of our method is the linear transmittance rasterizer. In our method, we rely on the vanilla 3DGS rasterizer that approximates transmittance linearly over a more physically accurate exponential attenuation. Initial experiments showed that due to our specific initialization our dynamic scene consists of a large quantity of small Gaussians (≈ 300 per ray), so that the linear approximation is especially close to the exponential model. Thus, the added complexity of an exponential rasterizer is not justified for these fire scenes. For further details we refer to Sec. 9 of the Supp. Mat.

Additionally, separating the static and dynamic components assumes minimal background-foreground interaction, which may not hold for large-scale or smoke-rich scenes. For reconstructing other volumetric phenomena with our flow-based pipeline, we refer to Sec. 8 of the Supp. Mat. Finally, while our LED-based synchronization mitigates temporal offsets, sub-millisecond inaccuracies can still introduce artifacts at very high motion speeds.

7 Conclusion

We presented a method for reconstructing the dynamics of real fire from high-speed multi-view video using 3D Gaussians. Our approach enables temporally coherent reconstructions of flames with substantially improved motion alignment and depth consistency, even under sparse-view capture and consumer-camera timing offsets. As future work, we plan to integrate physics-based priors for combustion dynamics to improve the photometric modeling of emissive materials. This will enable more adaptive temporal parameterizations of Gaussians and incorporate learning-based flow estimation directly in 3D space. Furthermore, this would allow us to extend the system toward fully monocular dynamic reconstruction for greater fidelity and robustness. Another promising direction is to explore the joint optimization of per-camera timing offsets, thus eliminating the need for our custom LED synchronization setup.

Acknowledgments

We thank Marvin Voigt, Anton Wagner, and Helge Wrede for their assistance in executing the experiments and Michael Lütten for providing the hardware used for the propane fire setup. Furthermore, the authors acknowledge the financial support of Catalyst: Leaders Julius von Haast Fellowship (23-VUW-019-JVH). Sören Pirk wishes to acknowledge the European Research Council (ERC) who partially funded this research through the ERC Consolidator Grant WildfireTwins (Grant agreement ID: 101170158).

References

1. Bae, J., Kim, S., Yun, Y., Lee, H., Bang, G., Uh, Y.: Per-gaussian embedding-based deformation for deformable 3d gaussian splatting. In: Eur. Conf. Comput. Vis. vol. 15073, pp. 321–335 (2024). https://doi.org/10.1007/978-3-031-72633-0_18
2. Bansal, A., Vo, M., Sheikh, Y., Ramanan, D., Narasimhan, S.G.: 4d visualization of dynamic events from unconstrained multi-view videos. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5365–5374 (2020). https://doi.org/10.1109/CVPR42600.2020.00541
3. Bargatin, V., Chistov, E., Yakovenko, A., Vatolin, D.: MEMFOF: High-resolution training for memory-efficient multi-frame optical flow estimation. In: Int. Conf. Comput. Vis. pp. 8187–8196 (2025). https://doi.org/10.1109/ICCV51701.2025.00767
4. Bay, H., Tuytelaars, T., Van Gool, L.: SURF: Speeded up robust features. In: Eur. Conf. Comput. Vis. vol. 3951, pp. 404–417 (2006). https://doi.org/10.1007/11744023_32
5. Black, M.J., Anandan, P.: A framework for the robust estimation of optical flow. In: Int. Conf. Comput. Vis. pp. 231–236 (1993). https://doi.org/10.1109/ICCV.1993.378214
6. Bleyer, M., Rhemann, C., Rother, C.: PatchMatch Stereo - stereo matching with slanted support windows. In: Brit. Mach. Vis. Conf. vol. 11, pp. 1–11 (2011). https://doi.org/10.5244/C.25.14
7. Bochkovskiy, A., Delaunoy, A., Germain, H., Santos, M., Zhou, Y., Richter, S., Koltun, V.: Depth Pro: Sharp monocular metric depth in less than a second. In: Yue, Y., Garg, A., Peng, N., Sha, F., Yu, R. (eds.) Int. Conf. Learn. Represent. pp. 75602–75637 (2025)
8. Bozic, A., Zollhöfer, M., Theobalt, C., Nießner, M.: DeepDeform: Learning non-rigid RGB-D reconstruction with semi-supervised data. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7000–7010 (2020). https://doi.org/10.1109/CVPR42600.2020.00703
9. Bregler, C., Hertzmann, A., Biermann, H.: Recovering non-rigid 3d shape from image streams. In: IEEE Conf. Comput. Vis. Pattern Recog. vol. 2, pp. 690–696 (2000). https://doi.org/10.1109/CVPR.2000.854941
10. Brox, T., Bregler, C., Malik, J.: Large displacement optical flow. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 41–48 (2009). https://doi.org/10.1109/CVPR.2009.5206697

11. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.: High accuracy optical flow estimation based on a theory for warping. In: Eur. Conf. Comput. Vis. vol. 3024, pp. 25–36 (2004). https://doi.org/10.1007/978-3-540-24673-2_3
12. Broxton, M., Flynn, J., Overbeck, R., Erickson, D., Hedman, P., Duvall, M., Dourgarian, J., Busch, J., Whalen, M., Debevec, P.: Immersive light field video with a layered mesh representation. ACM Trans. Graph. **39**(4), 86:1–86:15 (2020). https://doi.org/10.1145/3386569.3392485
13. Cao, A., Johnson, J.: HexPlane: A fast representation for dynamic scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 130–141 (2023). https://doi.org/10.1109/cvpr52729.2023.00021
14. Chen, X., Zheng, Y., Black, M.J., Hilliges, O., Geiger, A.: SNARF: Differentiable forward skinning for animating non-rigid neural implicit shapes. In: Int. Conf. Comput. Vis. pp. 11574–11584 (2021). https://doi.org/10.1109/ICCV48922.2021.01139
15. Chu, M., Liu, L., Zheng, Q., Franz, A., Seidel, H.P., Theobalt, C., Zayer, R.: Physics informed neural fields for smoke reconstruction with sparse data. ACM Trans. Graph. **41**(4), 119:1–119:14 (Aug 2022). https://doi.org/10.1145/3528223.3530169
16. Dai, Y., Li, H., He, M.: A simple prior-free method for non-rigid structure-from-motion factorization. Int. J. Comput. Vis. **107**(2), 101–122 (2014). https://doi.org/10.1007/S11263-013-0684-2
17. DeTone, D., Malisiewicz, T., Rabinovich, A.: SuperPoint: Self-supervised interest point detection and description. In: IEEE Conf. Comput. Vis. Pattern Recog. Worksh. pp. 224–236 (2018). https://doi.org/10.1109/CVPRW.2018.00060
18. Doersch, C., Gupta, A., Markeeva, L., Recasens, A., Smaira, L., Aytar, Y., Carreira, J., Zisserman, A., Yang, Y.: TAP-Vid: A benchmark for tracking any point in a video. In: Adv. Neural Inform. Process. Syst. pp. 13610–13626 (2022). https://doi.org/10.52202/068431-0989
19. Doersch, C., Luc, P., Yang, Y., Gokay, D., Koppula, S., Gupta, A., Heyward, J., Rocco, I., Goroshin, R., Carreira, J., et al.: Bootstap: Bootstrapped training for tracking-any-point. In: Asian Conf. Comput. Vis. vol. 15473, pp. 483–500 (2024). https://doi.org/10.1007/978-981-96-0901-7_28
20. Doersch, C., Yang, Y., Vecerik, M., Gokay, D., Gupta, A., Aytar, Y., Carreira, J., Zisserman, A.: TAPIR: Tracking any point with per-frame initialization and temporal refinement. In: Int. Conf. Comput. Vis. pp. 10027–10038 (2023). https://doi.org/10.1109/ICCV51070.2023.00923
21. Dou, M., Khamis, S., Degtyarev, Y., Davidson, P., Fanello, S.R., Kowdle, A., Orts-Escolano, S., Rhemann, C., Kim, D., Taylor, J., Kohli, P., Tankovich, V., Izadi, S.: Fusion4D: real-time performance capture of challenging scenes. ACM Trans. Graph. **35**(4), 114:1–114:13 (2016). https://doi.org/10.1145/2897824.2925969
22. Du, Y., Zhang, Y., Yu, H.X., Tenenbaum, J.B., Wu, J.: Neural radiance flow for 4d view synthesis and video processing. In: Int. Conf. Comput. Vis. pp. 14304–14314 (2021). https://doi.org/10.1109/ICCV48922.2021.01406
23. Duan, Y., Wei, F., Dai, Q., He, Y., Chen, W., Chen, B.: 4D-Rotor gaussian splatting: Towards efficient novel view synthesis for dynamic scenes. In: ACM SIGGRAPH Conf. Papers. pp. 87:1–87:11 (2024). https://doi.org/10.1145/3641519.3657463
24. Duisterhof, B.P., Zust, L., Weinzaepfel, P., Leroy, V., Cabon, Y., Revaud, J.: MASt3R-SfM: a fully-integrated solution for unconstrained structure-from-motion. In: Int. Conf. 3D Vis. pp. 1–10 (2025). https://doi.org/10.1109/3DV66043.2025.00008

25. Edstedt, J., Sun, Q., Bökman, G., Wadenbäck, M., Felsberg, M.: RoMa: Robust dense feature matching. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 19790–19800 (2024). https://doi.org/10.1109/CVPR52733.2024.01871
26. Feng, H., Zhang, J., Wang, Q., Ye, Y., Yu, P., Black, M.J., Darrell, T., Kanazawa, A.: St4RTrack: Simultaneous 4d reconstruction and tracking in the world. In: Int. Conf. Comput. Vis. pp. 8503–8513 (2025). https://doi.org/10.1109/ICCV51701.2025.00796
27. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-Planes: Explicit radiance fields in space, time, and appearance. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 12479–12488 (2023). https://doi.org/10.1109/CVPR52729.2023.01201
28. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: Int. Conf. Comput. Vis. pp. 5692–5701 (2021). https://doi.org/10.1109/ICCV48922.2021.00566
29. Gao, Y., Yu, H.X., Zhu, B., Wu, J.: FluidNexus: 3d fluid reconstruction and prediction from a single video. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 26091–26101 (June 2025). https://doi.org/10.1109/CVPR52734.2025.02430
30. Goli, L., Sabour, S., Matthews, M., Brubaker, M., Lagun, D., Jacobson, A., Fleet, D.J., Saxena, S., Tagliasacchi, A.: RoMo: Robust motion segmentation improves structure from motion. pp. 6155–6164 (2025). https://doi.org/10.1109/ICCV51701.2025.00581
31. Ha, H., Xiao, L., Richardt, C., Nguyen-Phuoc, T., Kim, C., Kim, M.H., Lanman, D., Khan, N.: Geometry-guided online 3d video synthesis with multi-view temporal consistency. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 11275–11285 (2025). https://doi.org/10.1109/cvpr52734.2025.01053
32. Harley, A.W., Fang, Z., Fragkiadaki, K.: Particle video revisited: Tracking through occlusions using point trajectories. In: Eur. Conf. Comput. Vis. vol. 13682, pp. 59–75 (2022). https://doi.org/10.1007/978-3-031-20047-2_4
33. Horn, B.K.P., Schunck, B.G.: Determining optical flow. Artif. Intell. **17**, 185–203 (1981). https://doi.org/10.1016/0004-3702(81)90024-2
34. Huang, B., Yu, Z., Chen, A., Geiger, A., Gao, S.: 2d gaussian splatting for geometrically accurate radiance fields. In: ACM SIGGRAPH Conf. Papers. pp. 32:1–32:11 (2024). https://doi.org/10.1145/3641519.3657428
35. Huang, Y.H., Sun, Y.T., Yang, Z., Lyu, X., Cao, Y.P., Qi, X.: SC-GS: Sparse-controlled gaussian splatting for editable dynamic scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4220–4230 (June 2024)
36. Hur, J., Roth, S.: Self-supervised monocular scene flow estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 7394–7403 (2020). https://doi.org/10.1109/CVPR42600.2020.00742
37. Ihrke, I., Magnor, M.A.: Image-based tomographic reconstruction of flames. In: Symp. Comput. Animation. pp. 365–373 (2004). https://doi.org/10.2312/SCA/SCA04/367-375
38. Innmann, M., Zollhöfer, M., Nießner, M., Theobalt, C., Stamminger, M.: VolumeDeform: Real-time volumetric non-rigid reconstruction. In: Eur. Conf. Comput. Vis. pp. 362–379 (2016). https://doi.org/10.1007/978-3-319-46484-8_22
39. Isik, M., Rünz, M., Georgopoulos, M., Khakhulin, T., Starck, J., Agapito, L., Nießner, M.: HumanRF: High-fidelity neural radiance fields for humans in motion. ACM Trans. Graph. pp. 160:1–160:12 (2023). https://doi.org/10.1145/3592415

40. Jiang, S., Campbell, D., Lu, Y., Li, H., Hartley, R.: Learning to estimate hidden motions with global motion aggregation. In: Int. Conf. Comput. Vis. pp. 9752–9761 (2021). https://doi.org/10.1109/ICCV48922.2021.00963
41. Jiang, S., Lu, Y., Li, H., Hartley, R.: Learning optical flow from a few matches. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16592–16600 (2021). https://doi.org/10.1109/CVPR46437.2021.01632
42. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: CoTracker: It is better to track together. In: Eur. Conf. Comput. Vis. vol. 15120, pp. 18–35 (2024). https://doi.org/10.1007/978-3-031-73033-7_2
43. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139:1–139:14 (2023). https://doi.org/10.1145/3592433
44. Kopf, J., Rong, X., Huang, J.B.: Robust consistent video depth estimation. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 1611–1621 (2021). https://doi.org/10.1109/CVPR46437.2021.00166
45. Koppula, S., Rocco, I., Yang, Y., Heyward, J., Carreira, J., Zisserman, A., Brostow, G., Doersch, C.: TAPVid-3D: A benchmark for tracking any point in 3d. In: Adv. Neural Inform. Process. Syst. (2024). https://doi.org/10.52202/079017-2611
46. Kratimenos, A., Lei, J., Daniilidis, K.: DynMF: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. In: Eur. Conf. Comput. Vis. vol. 15137, pp. 252–269 (2024). https://doi.org/10.1007/978-3-031-72986-7_15
47. Kumar, S., Dai, Y., Li, H.: Monocular dense 3d reconstruction of a complex dynamic scene from two perspective frames. In: Int. Conf. Comput. Vis. pp. 4659–4667 (2017). https://doi.org/10.1109/ICCV.2017.498
48. Lei, J., Weng, Y., Harley, A.W., Guibas, L., Daniilidis, K.: MoSca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6165–6177 (2025). https://doi.org/10.1109/CVPR52734.2025.00578
49. Leroy, V., Cabon, Y., Revaud, J.: Grounding image matching in 3d with MASt3R. In: Eur. Conf. Comput. Vis. pp. 71–91 (2024). https://doi.org/10.1007/978-3-031-73220-1_5
50. Li, R., Tanke, J., Vo, M., Zollhöfer, M., Gall, J., Kanazawa, A., Lassner, C.: TAVA: Template-free animatable volumetric actors. In: Eur. Conf. Comput. Vis. vol. 13692, pp. 419–436 (2022). https://doi.org/10.1007/978-3-031-19824-3_25
51. Li, T., Slavcheva, M., Zollhöfer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Newcombe, R.A., Lv, Z.: Neural 3d video synthesis from multi-view video. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5511–5521 (2022). https://doi.org/10.1109/CVPR52688.2022.00544
52. Li, Z., Chen, Z., Li, Z., Xu, Y.: Spacetime gaussian feature splatting for real-time dynamic view synthesis. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 8508–8520 (June 2024). https://doi.org/10.1109/CVPR52733.2024.00813
53. Li, Z., Dekel, T., Cole, F., Tucker, R., Snavely, N., Liu, C., Freeman, W.T.: Learning the depths of moving people by watching frozen people. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4521–4530 (2019). https://doi.org/10.1109/CVPR.2019.00465
54. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 6498–6508 (2021). https://doi.org/10.1109/CVPR46437.2021.00643

55. Li, Z., Wang, Q., Cole, F., Tucker, R., Snavely, N.: DynIBaR: Neural dynamic image-based rendering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4273–4284 (2023). https://doi.org/10.1109/CVPR52729.2023.00416
56. Liu, C., Yuen, J., Torralba, A.: SIFT flow: Dense correspondence across scenes and its applications. IEEE Trans. Pattern Anal. Mach. Intell. **33**(5), 978–994 (2011). https://doi.org/10.1109/TPAMI.2010.147
57. Liu, Q., Liu, Y., Wang, J., Lyu, X., Wang, P., Wang, W., Hou, J.: MoDGS: Dynamic gaussian splatting from casually-captured monocular videos with depth priors. In: Int. Conf. Learn. Represent. (2025)
58. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vis. **60**(2), 91–110 (2004). https://doi.org/10.1023/B:VISI.0000029664.99615.94
59. Lu, J., Huang, T., Li, P., Dou, Z., Lin, C., Cui, Z., Dong, Z., Yeung, S.K., Wang, W., Liu, Y.: Align3R: Aligned monocular depth estimation for dynamic videos. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 22820–22830 (2025). https://doi.org/10.1109/CVPR52734.2025.02125
60. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: IJCAI. vol. 2, pp. 674–679 (1981)
61. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In: Int. Conf. 3D Vis. pp. 800–809 (2024). https://doi.org/10.1109/3DV62453.2024.00044
62. Luo, X., Huang, J.B., Szeliski, R., Matzen, K., Kopf, J.: Consistent video depth estimation. ACM Trans. Graph. **39**(4), 71:1–71:13 (2020). https://doi.org/10.1145/3386569.3392377
63. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: Eur. Conf. Comput. Vis. pp. 405–421. Springer (2020). https://doi.org/10.1007/978-3-030-58452-8_24
64. Neoral, M., Serych, J., Matas, J.: MFT: Long-term tracking of every pixel. In: IEEE Winter Conf. Appl. Comput. Vis. pp. 6823–6833 (2024). https://doi.org/10.1109/WACV57701.2024.00669
65. Newcombe, R.A., Fox, D., Seitz, S.M.: DynamicFusion: Reconstruction and tracking of non-rigid scenes in real-time. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 343–352 (2015). https://doi.org/10.1109/CVPR.2015.7298631
66. Ngo, T.D., Zhuang, P., Gan, C., Kalogerakis, E., Tulyakov, S., Lee, H.Y., Wang, C.: DELTA: Dense efficient long-range 3d tracking for any video. In: Int. Conf. Learn. Represent. (2025)
67. Nguyen, D.Q., Fedkiw, R., Jensen, H.W.: Physically based modeling and animation of fire. ACM Trans. Graph. **21**(3), 721–728 (2002). https://doi.org/10.1145/566654.566643
68. Novotny, D., Ravi, N., Graham, B., Neverova, N., Vedaldi, A.: C3DPO: Canonical 3d pose networks for non-rigid structure from motion. In: Int. Conf. Comput. Vis. pp. 7687–7696 (2019). https://doi.org/10.1109/ICCV.2019.00778
69. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: NeRFies: Deformable neural radiance fields. In: Int. Conf. Comput. Vis. pp. 5845–5854 (2021). https://doi.org/10.1109/ICCV48922.2021.00581
70. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: HyperNeRF: A higher-dimensional representation for topologically varying neural radiance fields. ACM Trans. Graph. **40**(6), 238:1–238:12 (2021). https://doi.org/10.1145/3478513.3480487

71. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: Neural radiance fields for dynamic scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10318–10327 (2021). https://doi.org/10.1109/CVPR46437.2021.01018
72. Ranftl, R., Vineet, V., Chen, Q., Koltun, V.: Dense monocular depth estimation in complex dynamic scenes. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4058–4066 (2016). https://doi.org/10.1109/CVPR.2016.440
73. Ren, Z., Gallo, O., Sun, D., Yang, M.H., Sudderth, E.B., Kautz, J.: A fusion approach for multi-frame optical flow estimation. In: IEEE Winter Conf. Appl. Comput. Vis. pp. 2077–2086 (2019). https://doi.org/10.1109/WACV.2019.00225
74. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.R.: ORB: An efficient alternative to SIFT or SURF. In: Int. Conf. Comput. Vis. pp. 2564–2571 (2011). https://doi.org/10.1109/ICCV.2011.6126544
75. Russell, C., Yu, R., Agapito, L.: Video pop-up: Monocular 3d reconstruction of dynamic scenes. In: Eur. Conf. Comput. Vis. vol. 8695, pp. 583–598 (2014). https://doi.org/10.1007/978-3-319-10584-0_38
76. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 4104–4113 (2016). https://doi.org/10.1109/CVPR.2016.445
77. Schönberger, J.L., Price, T., Sattler, T., Frahm, J.M., Pollefeys, M.: A vote-and-verify strategy for fast spatial verification in image retrieval. In: Asian Conf. Comput. Vis. vol. 10111, pp. 321–337 (2016). https://doi.org/10.1007/978-3-319-54181-5_21
78. Schönberger, J.L., Zheng, E., Pollefeys, M., Frahm, J.M.: Pixelwise view selection for unstructured multi-view stereo. In: Eur. Conf. Comput. Vis. vol. 9907, pp. 501–518 (2016). https://doi.org/10.1007/978-3-319-46487-9_31
79. Shi, X., Huang, Z., Bian, W., Li, D., Zhang, M., Cheung, K.C., See, S., Qin, H., Dai, J., Li, H.: VideoFlow: Exploiting temporal cues for multi-frame optical flow estimation. In: Int. Conf. Comput. Vis. pp. 12435–12446 (2023). https://doi.org/10.1109/ICCV51070.2023.01146
80. Song, L., Chen, A., Li, Z., Chen, Z., Chen, L., Yuan, J., Xu, Y., Geiger, A.: NeRF-Player: A streamable dynamic scene representation with decomposed neural radiance fields. IEEE Trans. Vis. Comput. Graph. **29**(5), 2732–2742 (2023). https://doi.org/10.1109/TVCG.2023.3247082
81. Song, Y., Lei, J., Wang, Z., Liu, L., Daniilidis, K.: Track everything everywhere fast and robustly. In: Eur. Conf. Comput. Vis. vol. 15061, pp. 343–359 (2024). https://doi.org/10.1007/978-3-031-72646-0_20
82. Thonat, T., Aksoy, Y., Aittala, M., Paris, S., Durand, F., Drettakis, G.: Video-based rendering of dynamic stationary environments from unsynchronized inputs. In: Comput. Graph. Forum. vol. 40, pp. 73–86 (2021). https://doi.org/10.1111/CGF.14342
83. Viola, M., Qu, K., Metzger, N., Ke, B., Becker, A., Schindler, K., Obukhov, A.: Marigold-DC: Zero-shot monocular depth completion with guided diffusion. In: Int. Conf. Comput. Vis. pp. 5359–5370 (2025). https://doi.org/10.1109/ICCV51701.2025.00509
84. Wang, C., Eckart, B., Lucey, S., Gallo, O.: Neural trajectory fields for dynamic novel view synthesis. arXiv preprint arXiv:2105.05994 (2021)
85. Wang, D., Rim, P., Tian, T., Lao, D., Wong, A., Sundaramoorthi, G.: ODE-GS: Latent odes for dynamic scene extrapolation with 3d gaussian splatting. arXiv preprint arXiv:2506.05480 (2025)

86. Wang, L., Zhang, J., Liu, X., Zhao, F., Zhang, Y., Zhang, Y., Wu, M., Yu, J., Xu, L.: Fourier plenoctrees for dynamic radiance field rendering in real-time. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 13514–13524 (2022). https://doi.org/10.1109/CVPR52688.2022.01316
87. Wang, Q., Chang, Y.Y., Cai, R., Li, Z., Hariharan, B., Holynski, A., Snavely, N.: Tracking everything everywhere all at once. In: Int. Conf. Comput. Vis. pp. 19738–19749 (2023). https://doi.org/10.1109/ICCV51070.2023.01813
88. Wang, Q., Zhang, Y., Holynski, A., Efros, A.A., Kanazawa, A.: Continuous 3d perception model with persistent state. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10510–10522 (2025). https://doi.org/10.1109/CVPR52734.2025.00983
89. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: DUSt3R: Geometric 3d vision made easy. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 20697–20709 (2024). https://doi.org/10.1109/CVPR52733.2024.01956
90. Wang, Y., Yang, P., Xu, Z., Sun, J., Zhang, Z., Chen, Y., Bao, H., Peng, S., Zhou, X.: FreeTimeGS: Free gaussian primitives at anytime anywhere for dynamic scene reconstruction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 21750–21760 (2025). https://doi.org/10.1109/CVPR52734.2025.02026
91. Weng, C.Y., Curless, B., Srinivasan, P.P., Barron, J.T., Kemelmacher-Shlizerman, I.: HumanNeRF: Free-viewpoint rendering of moving people from monocular video. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 16189–16199 (2022). https://doi.org/10.1109/CVPR52688.2022.01573
92. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 20310–20320 (2024). https://doi.org/10.1109/CVPR52733.2024.01920
93. Wu, Q., Esturo, J.M., Mirzaei, A., Moënne-Loccoz, N., Gojcic, Z.: 3DGUT: Enabling distorted cameras and secondary rays in gaussian splatting. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 26036–26046 (2025). https://doi.org/10.1109/CVPR52734.2025.02425
94. Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 9421–9431 (2021). https://doi.org/10.1109/CVPR46437.2021.00930
95. Xiao, Y., Wang, J., Xue, N., Karaev, N., Makarov, Y., Kang, B., Zhu, X., Bao, H., Shen, Y., Zhou, X.: SpatialTrackerV2: 3d point tracking made easy. In: Int. Conf. Comput. Vis. pp. 6726–6737 (2025). https://doi.org/10.1109/ICCV51701.2025.00633
96. Xiao, Y., Wang, Q., Zhang, S., Xue, N., Peng, S., Shen, Y., Zhou, X.: SpatialTracker: Tracking any 2d pixels in 3d space. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 20406–20417 (2024). https://doi.org/10.1109/CVPR52733.2024.01929
97. Xu, J., Fan, Z., Yang, J., Xie, J.: Grid4D: 4d decomposed hash encoding for high-fidelity dynamic gaussian splatting. In: Adv. Neural Inform. Process. Syst. pp. 123787–123811 (2024). https://doi.org/10.52202/079017-3934
98. Xu, Z., Li, Z., Dong, Z., Zhou, X., Newcombe, R., Lv, Z.: 4DGT: Learning a 4d gaussian transformer using real-world monocular videos. In: Adv. Neural Inform. Process. Syst. (2025)
99. Yang, G., Sun, D., Jampani, V., Vlasic, D., Cole, F., Chang, H., Ramanan, D., Freeman, W.T., Liu, C.: LASR: Learning articulated shape reconstruction from a monocular video. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 15980–15989 (2021). https://doi.org/10.1109/CVPR46437.2021.01572

100. Yang, G., Sun, D., Jampani, V., Vlasic, D., Cole, F., Liu, C., Ramanan, D.: ViSER: Video-specific surface embeddings for articulated 3d shape reconstruction. In: Adv. Neural Inform. Process. Syst. pp. 19326–19338 (2021)
101. Yang, G., Vo, M., Neverova, N., Ramanan, D., Vedaldi, A., Joo, H.: BANMo: Building animatable 3d neural models from many casual videos. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 2853–2863 (2022). https://doi.org/10.1109/CVPR52688.2022.00288
102. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything: Unleashing the power of large-scale unlabeled data. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 10371–10381 (2024). https://doi.org/10.1109/CVPR52733.2024.00987
103. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth Anything V2. In: Adv. Neural Inform. Process. Syst. pp. 21875–21911 (2024). https://doi.org/10.52202/079017-0688
104. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 20331–20341 (2024). https://doi.org/10.1109/CVPR52733.2024.01922
105. Yang, Z., Yang, H., Pan, Z., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In: Int. Conf. Learn. Represent. (2024)
106. Yoon, J.S., Kim, K., Gallo, O., Park, H.S., Kautz, J.: Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In: IEEE Conf. Comput. Vis. Pattern Recog. pp. 5335–5344 (2020). https://doi.org/10.1109/CVPR42600.2020.00538
107. Zhang, Z., Cole, F., Li, Z., Rubinstein, M., Snavely, N., Freeman, W.T.: Structure and motion from casual videos. In: Eur. Conf. Comput. Vis. vol. 13693, pp. 20–37 (2022). https://doi.org/10.1007/978-3-031-19827-4_2
108. Zhang, Z., Cole, F., Tucker, R., Freeman, W.T., Dekel, T.: Consistent depth of moving objects in video. ACM Trans. Graph. **40**(4), 148:1–148:12 (2021). https://doi.org/10.1145/3450626.3459871
109. Zheng, Y., Harley, A.W., Shen, B., Wetzstein, G., Guibas, L.J.: PointOdyssey: A large-scale synthetic dataset for long-term point tracking. In: Int. Conf. Comput. Vis. pp. 19798–19808 (2023). https://doi.org/10.1109/ICCV51070.2023.01818
110. Zhu, R., Liang, Y., Chang, H., Deng, J., Lu, J., Yang, W., Zhang, T., Zhang, Y.: MotionGS: Exploring explicit motion guidance for deformable 3d gaussian splatting. Adv. Neural Inform. Process. Syst. **37**, 101790–101817 (2024). https://doi.org/10.52202/079017-3229
111. Zollhöfer, M., Nießner, M., Izadi, S., Rhemann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A.W., Loop, C.T., Theobalt, C., Stamminger, M.: Real-time non-rigid reconstruction using an RGB-D camera. ACM Trans. Graph. **33**(4), 156:1–156:12 (2014). https://doi.org/10.1145/2601097.2601165