MoSca: Dynamic Gaussian Fusion from Casual Videos via 4D Motion Scaffolds

Jiahui Lei[1] Yijia Weng[2] Adam W. Harley[2] Leonidas Guibas[2] Kostas Daniilidis[1,3]
[1] University of Pennsylvania [2] Stanford University [3] Archimedes, Athena RC
{leijh, kostas}@cis.upenn.edu, {yijiaw, aharley, guibas}@cs.stanford.edu

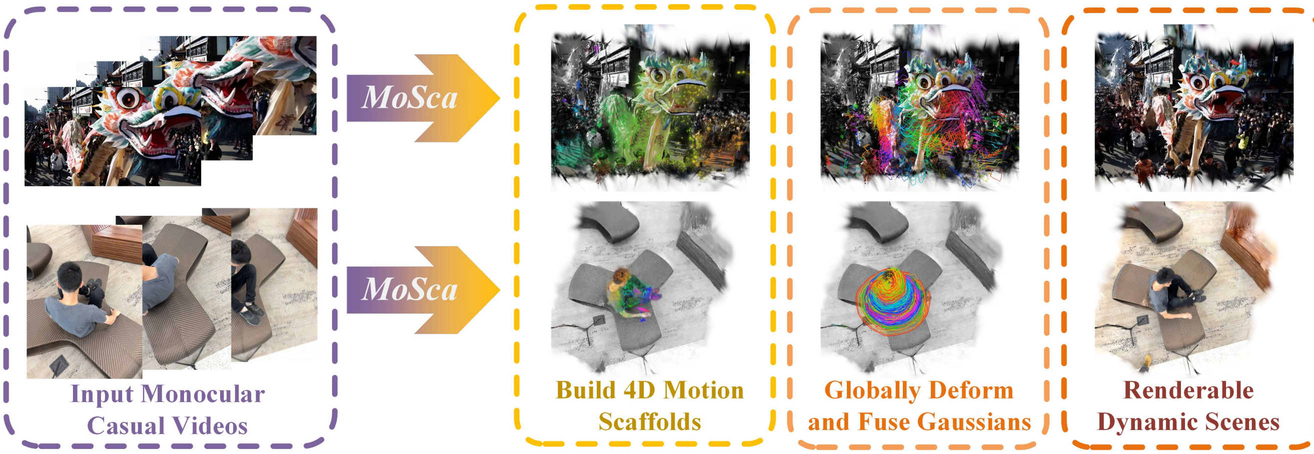

Figure 1. MoSca reconstructs renderable dynamic scenes from monocular casual videos.

Abstract

We introduce 4D Motion Scaffolds (MoSca), a modern 4D reconstruction system designed to reconstruct and synthesize novel views of dynamic scenes from monocular videos captured casually in the wild. To address such a challenging and ill-posed inverse problem, we leverage prior knowledge from foundational vision models and lift the video data to a novel Motion Scaffold (MoSca) representation, which compactly and smoothly encodes the underlying motions/deformations. The scene geometry and appearance are then disentangled from the deformation field and are encoded by globally fusing the Gaussians anchored onto the MoSca and optimized via Gaussian Splatting. Additionally, camera focal length and poses can be solved using bundle adjustment without the need of any other pose estimation tools. Experiments demonstrate state-of-the-art performance on dynamic rendering benchmarks and its effectiveness on real videos. Project page and code: https://www.cis.upenn.edu/~leijh/projects/mosca

1. Introduction

This paper presents 4D Motion Scaffolds (*MoSca*), a fully automated system for reconstructing and rendering dynamic scenes from casual monocular video inputs with unknown camera parameters—the most typical data format for such a system in the wild. Robust 4D scene reconstruction from such input is increasingly vital for constructing datasets for future AGI models, content creation for spatial computing and VR/MR/AR, and building embodied agents to perceive and learn from real video data. However, this task is known to be highly challenging and inherently ill-posed [30, 51, 66] due to the limited availability of multi-view stereo cues in casual video footage.

To tackle this challenging task, our first insight is to leverage the recent advances of pretrained vision models (Sec. 3.2.1), which today are very effective at fundamental computer vision tasks such as tracking and depth estimation. While this knowledge provides a critical boost to understanding the complete dynamic scene, it is inherently insufficient, as it fails to capture occluded parts of the scene and it is usually noisy, local, and partial. Our second insight is to design a deformation representation, *MoSca*, derived from the above foundational priors, exploiting a *physical* deformation prior. Although the real-world geometry and appearance are complex and include high-frequency details, the underlying deformation that drives these geometries is usually compact (low-rank) and smooth. *MoSca* leverages this property by disentangling the 3D geometry and motion, representing the deformation with sparse graph nodes that can be smoothly interpolated (Sec. 3.1). Another physical prior we exploit is the as-rigid-as-possible (ARAP) deformation, which can be efficiently applied via the trajectory topology of *MoSca*. Two important benefits arise from the above two insights: firstly, *MoSca* can be lifted into 3D and optimized from the inferred 2D foundational priors (Sec. 3.2.3), and secondly, the observations from all timesteps can be globally fused and rendered for any query time (Sec. 3.2.4). Gaussian fusion happens when we deform all Gaussians observed at different times to the query time, forming a complete reconstruction, which can be supervised through Gaussian Splatting [44]. Furthermore, our system estimates the camera poses and focal lengths via a bundle adjustment and the photometric optimization (Sec. 3.2.2), obviating the need for other poes estimators such as COLMAP.

In summary, our main contributions can be summarized as: (1) An automatic 4D reconstruction system that works in the real world for pose-free monocular videos. (2) A novel Motion Scaffold deformation representation, which we build using knowledge from 2D foundational models, and optimize via physically-inspired deformation regularization. (3) An efficient and explicit Gaussian-based dynamic scene representation, driven by *MoSca*, which globally fuses observations across an input video to render this data into any new viewpoint and query time of choice. (4) State-of-the-art performance on dynamic scene rendering benchmarks.

2. Related Works

Dynamic Novel-View Synthesis. Novel-view synthesis of dynamic scenes is challenging. Many existing works [2, 3, 5, 13, 28, 55, 60, 67, 78, 87, 120] assume available synchronized multi-view video inputs. Another line of works [11, 29, 56, 59, 66, 68, 85, 94, 96, 97, 104, 105, 110, 112, 113] tackles the more practical setting of monocular inputs, where ambiguities from limited observations further complicate the problem. As [30] pointed out, most methods struggle with realistic single-view videos. To disambiguate, some works [1, 16, 33, 35, 48, 52, 54, 65, 79, 82, 84, 90, 101, 102] target specific scenes and exploit domain knowledge like template models [8, 95]. A few recent works [51, 58, 118, 119] fuse information across frames, but only from a small temporal window.

Neural radiance fields [4, 14, 27, 69, 70, 74, 75] and 3D Gaussian Splatting [44–46, 114] are promising approaches to novel view synthesis. The latter's explicit point-based representation fits particularly well into the dynamic setting [18, 21, 25, 26, 37, 42, 50, 57, 59, 61, 67, 103, 110, 111]. We employ 3D Gaussians for long-term, global aggregation. Compared to concurrent works [64, 83, 86, 99], *MoSca* has a more structured deformation representation exploiting powerful 2D foundation models, and is a full-stack automated system that directly outputs 4D reconstruction from an unposed RGB video.

Non-Rigid Structure-from-Motion. Geometric reconstruction of non-rigidly deforming scenes from a single camera is a long-standing problem. [7, 8, 81, 107, 108, 121] focus on specific object categories or articulated shapes and register observations to template models [8]. [10, 19, 23, 24, 31, 53, 71] warp, align, and fuse scans of generic scenes. To model non-rigid deformations, state-of-the-art methods [10, 23, 71, 121] use Embedded Deformation Graphs [89], where dense transformations over the space are modeled with a sparse set of basis transformations. In *MoSca*, we extend classic Embedded Graphs to connect priors from 2D foundation models to dynamic Gaussian splatting.

2D Vision Foundation Models. Recent years have witnessed great progress in large-scale pretrained vision foundation models [9, 47, 72, 73, 80] that serve various downstream tasks, ranging from image-level tasks such as visual question answering [62, 63, 72] to pixel-level tasks including segmentation [47], dense tracking [32, 40], and monocular depth estimation [6, 76, 109]. These models encode strong data priors particularly useful in monocular video-based dynamic reconstruction, as they help disambiguate partial observations. While most previous methods [18, 29, 51, 56, 58, 64, 86, 99, 118] directly use the 2D priors for regularization in image space, and often in isolation from each other, we propose to lift these 2D priors to 3D and fuse them in a coordinated way.

3. Method

Overview. Given a casual monocular video of a dynamic scene with T frames $\mathcal{I} = [I_1, I_2, \dots I_T]$, our fully automatic system reconstructs the geometry and appearance of the scene with a set of dynamic Gaussians and recovers the focal length and poses of the camera if they are unknown .Our key idea is to lift the 2D video input to a novel 4D dynamic scene representation, which we name Motion Scaffolds (*MoSca*), where all the observations are fused **globally** and **geometrically**. Fig. 2 provides an overview of our approach. We first introduce the deformation representation *MoSca* in Sec.3.1 and then, detail each step of our reconstruction system in Sec. 3.2.

3.1. Deformation Representation with *MoSca*

A fundamental challenge in real-world 4D reconstruction is the high dimensionality of the potential solution space compared to the extremely limited spatiotemporal observations. However, real-world motion typically behaves rigidly, smoothly, and compactly, meaning that the actual solution is low-rank and driven by a few key "eigen" motions. With this insight, we model the underlying deformation of the scene using an explicit, compact, and structured graph $(\mathcal{V}, \mathcal{E})$, named 4D Motion Scaffold (*MoSca*), which encodes these local "eigen" motions and interpolates the dense deformation field.

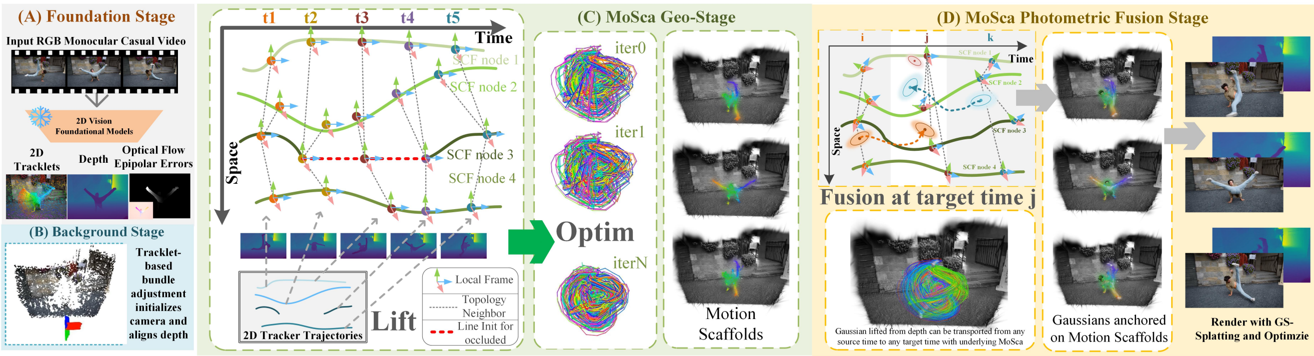

Figure 2. **Overview**: (A) Given a monocular casual video, we infer pre-trained 2D vision foundation models (Sec. 3.2.1). (B) The camera intrinsics and poses are initialized using tracklet-based bundle adjustment (Sec. 3.2.2). (C) Our proposed Motion Scaffold (*MoSca*) is lifted from 2D predictions and optimized with physics-inspired regularizations (Sec. 3.2.3). (D) Gaussians are initialized from all timesteps, deformed with *MoSca* (Sec. 3.1), and fused globally to model the dynamic scene. The entire representation is rendered with Gaussian Splatting and optimized with photometric losses (Sec. 3.2.4).

Motion Scaffold Graph Definition. Intuitively, the *MoSca* graph nodes $\mathcal{V} = \{\mathbf{v}^{(m)}\}_{m=1}^{M}$ are 6-DoF trajectories that capture the underlying low-rank, smooth motion of the scene. The number of nodes M is significantly smaller (e.g., see Tab. 7) than the number of points required to represent the scene. Specifically, each node $\mathbf{v}^{(m)} \in \mathcal{V}$ consists of per-timestep rigid transformations $\mathbf{Q}_t^{(m)}$ and a global control radius $r^{(m)}$, which parameterizes a radial basis function (RBF) describing its influence on nearby space:

$$\mathbf{v}^{(m)} = ([\mathbf{Q}_1^{(m)}, \mathbf{Q}_2^{(m)}, \dots, \mathbf{Q}_T^{(m)}], r^{(m)}), \tag{1}$$

where $\mathbf{Q}^{(m)} = [\mathbf{R}^{(m)}, \mathbf{t}^{(m)}] \in SE(3)$ and $r^{(m)} \in \mathbb{R}^+$ is the radius. To properly interpolate the node-encoded trajectories and regularize the deformation, we organize the nodes $\mathbf{v}^{(m)}$ into a topology. We define the *MoSca* graph edges $\mathcal{E}$ as:

$$\mathcal{E}(m) = \mathrm{KNN}_{n\in\{1,\dots,M\}}\left[D_{\text{curve}}(m,n)\right],$$
$$D_{\text{curve}}(m,n) = \max_{t=1,2,\dots,T} \|\mathbf{t}_t^{(m)} - \mathbf{t}_t^{(n)}\|, \tag{2}$$

where KNN denotes the K-nearest neighbors under the curve distance metric D_{curve}. This metric captures the global proximity between trajectories across all timesteps and accounts for topological changes (e.g., opening a door does not connect the door and wall).

$\mathbf{SE(3)}$ Deformation Field. Given *MoSca* $(\mathcal{V}, \mathcal{E})$, we can derive a dense deformation field by interpolating motions from nodes near the query point. We use Dual Quaternion Blending (DQB) [43] to mix multiple $SE(3)$ elements on the $SE(3)$ manifold. Similar to the unit quaternion representation of $SO(3)$, the unit dual quaternion represents $SE(3)$ using eight numbers by including a dual part. Please refer to [20, 38, 43] for details. Given L rigid transformations $\mathbf{Q}_i \in SE(3)$ and their blending weights w_i, the interpolated motion is:

$$\mathrm{DQB}(\{(w_i, \mathbf{Q}_i)\}_{i=1}^{L}) = \frac{\sum_{i=1}^{L} w_i \hat{\mathbf{q}}_i}{\|\sum_{i=1}^{L} w_i \hat{\mathbf{q}}_i\|_{DQ}} \in SE(3), \tag{3}$$

where $\hat{\mathbf{q}}$ is the dual quaternion representation of $\mathbf{Q}$ and $|\cdot|_{DQ}$ denotes the dual norm [43]. Unlike linear blend skinning (LBS), DQB is a manifold interpolation that always produces an interpolated element in $SE(3)$. Consider any query position $\mathbf{x}$ in 3D space at time t_{src}. Denote its nearest node at t_{src} as $\mathbf{v}^{(m^*)}$ where $m^* = \arg\min_m ||\mathbf{t}_{t_{src}}^{(m)} - \mathbf{x}||$ and $\mathbf{t}_{t_{src}}^{(m)}$ is the translation part of node m's transformation at time t_{src}.

We can efficiently compute its $SE(3)$ deformation to the query time t_{dst} using nodes in the neighborhood of $v^{(m^*)}$. Formally, the deformation field $\mathcal{W}$ from time t_{src} to time t_{dst} is:

$$\mathcal{W}(\mathbf{x}, \mathbf{w}; t_{\text{src}}, t_{\text{dst}}) = \mathrm{DQB}\left(\{w_i, \Delta\mathbf{Q}^{(i)}\}_{i\in\mathcal{E}(m^*)}\right), \tag{4}$$

where $\Delta\mathbf{Q}^{(i)} = \mathbf{Q}_{t_{\text{dst}}}^{(i)} (\mathbf{Q}_{t_{\text{src}}}^{(i)})^{-1}$ and $\mathbf{w} = \{w_i\}$ are skinning weights computed from RBFs parameterized by radius $r^{(i)}$:

$$w_i(\mathbf{x}, t_{\text{src}}) = \exp\left(-\|\mathbf{x} - \mathbf{t}_{t_{\text{src}}}^{(i)}\|_2^2 / 2r^{(i)}\right) \in \mathbb{R}^+. \tag{5}$$

In summary, *MoSca* $(\mathcal{V}, \mathcal{E})$ encodes the deformation field through skinning on a structured, sparse trajectory graph. In the following sections, we will demonstrate how to reconstruct *MoSca* and attach Gaussians onto it to produce the final 4D reconstruction.

3.2. Reconstruction System

3.2.1 Leveraging Priors from 2D Foundation Models

4D reconstruction from monocular videos is highly ill-posed; therefore, it is essential to leverage prior knowledge to constrain the solution space. In the first step of our system, we exploit the priors provided by large vision foundation models pre-trained on massive datasets. Specifically, we utilize off-the-shelf pre-trained models to obtain: 1) Depth estimations [34, 36, 76] $\mathcal{D} = [D_1, D_2, \dots, D_T]$ that are relatively consistent across frames; 2) Long-term 2D pixel trajectories [22, 41, 106] $\mathcal{T} = \{\tau^{(i)} = [(p_1^{(i)}, v_1^{(i)}), (p_2^{(i)}, v_2^{(i)}), \dots, (p_T^{(i)}, v_T^{(i)})]\}_i$, where $p_t^{(i)}$ and

$v_t^{(i)}$ represent the i-th trajectory's 2D image coordinate and visibility at frame t; 3) Per-frame epipolar error maps $\mathcal{M} = [E_1, E_2, \ldots, E_T]$ [66] computed from RAFT[91] dense optical flow predictions, which indicate the likelihood of being in the dynamic foreground. These inferred results provide critical cues about geometry and correspondence. However, such raw information is partial, local, and noisy, and does not constitute a complete solution. We are going to fuse and optimize these initial cues to produce a coherent and global 4D reconstruction.

3.2.2 Camera Initializaition

To enable 4D reconstruction in the wild, our system must operate on dynamic scene videos with unknown camera parameters. Therefore, in the second step of our reconstruction pipeline, we propose a tracklet-based bundle adjustment to robustly initialize the camera focal lengths and poses. Given the inferred 2D tracks $\mathcal{T}$ and epipolar error maps $\mathcal{M}$, we first compute the maximum epipolar error of each tracklet as $e(\tau) = \max_{t=1\ldots T} E_t[p_t] \cdot v_t$ across visible timesteps. We identify confident background tracklets by thresholding $e(\tau)$ with a predefined small threshold. Starting with a pre-defined initial camera focal length, we optimize the camera poses and intrinsics jointly by minimizing the reprojection errors on these confident static tracks:

$$\mathcal{L}_{proj} = \sum_{i \in |\mathcal{T}_{\text{static}}|} \sum_{a,b \in [1,T]} (v_a^{(i)} v_b^{(i)}) \cdot \left\| \pi_{\mathbf{K}} \left(\mathbf{W}_b^{-1} \mathbf{W}_a \pi_{\mathbf{K}}^{-1}(p_a^{(i)}, D_a[p_a^{(i)}]) \right) - p_b^{(i)} \right\|, \tag{6}$$

where p_a and p_b are pixel locations, $\pi_{\mathbf{K}}$ denotes projection with intrinsics $\mathbf{K}$, and $\mathbf{W}_t$ is the camera pose at time t. To account for errors in the depth estimation—particularly scale misalignment—we jointly optimize a correction to the depth $D_a[p_a]$, which consists of per-frame global scaling factors and small per-pixel corrections, using a depth alignment loss:

$$\mathcal{L}_z = \sum_{i \in |\mathcal{T}_{\text{static}}|} \sum_{a,b \in [1,T]} (v_a^{(i)} v_b^{(i)}) D_{\text{scale-inv}} \left(\left[\mathbf{W}_b^{-1} \mathbf{W}_a \pi_{\mathbf{K}}^{-1}(p_a^{(i)}, D_a[p_a^{(i)}]) \right]_z, D_b[p_b^{(i)}] \right), \tag{7}$$

where $[\cdot]_z$ takes the z coordinate, and $D_{\text{scale-inv}}(x, y) = |x/y - 1| + |y/x - 1|$. The overall bundle adjustment loss is $\mathcal{L}_{\text{BA}} = \lambda_{\text{proj}} \mathcal{L}_{proj} + \lambda_{\text{z}} \mathcal{L}_z$, and the solved camera poses $\mathbf{W}_t$ will be refined during later rendering phases. While camera solving is not our primary contribution, our system achieves state-of-the-art camera pose accuracy on dynamic videos (Sec. 4.2); more details are provided in the Supplemental Material.

3.2.3 Geometric Optimization of *MoSca*

After inferring the 2D foundational models and initializing the camera, we are ready to geometrically construct *MoSca* $(\mathcal{V}, \mathcal{E})$ in the third step of our system. A key contribution of this paper is the seamless integration of *MoSca* with powerful 2D foundational models. Specifically, the long-term 2D tracking $\mathcal{T}$, together with the depth estimates $\mathcal{D}$, provide strong cues for constructing $\mathcal{V}$. However, there is still a gap due to missing information when tracks are invisible and because the local rotation component of *MoSca* is also unknown. We address this gap by incorporating physics-inspired regularization into the optimization of *MoSca*.

3D Lift and Initialization. Similar to the camera initialization, we identify foreground 2D tracks by thresholding the maximum epipolar error $e(\tau)$ of each tracklet. We then lift the foreground tracklets into 3D using depth estimates $\mathcal{D}$ at visible timesteps and linearly interpolate between nearby observations at occluded timesteps. Formally, we compute the lifted 3D position $\mathbf{h}_t$ at timestep t from the 2D track $\tau = [(p_t, v_t)]_{t=1}^T$ as

$$\mathbf{h}_t = \begin{cases} \mathbf{W}_t \pi_{\mathbf{K}}^{-1}(p_t, D_t[p_t]), & \text{if} \quad v_t = 1, \\ \text{LinearInterp}(\mathbf{h}_{\text{left}}, \mathbf{h}_{\text{right}}), & \text{if} \quad v_t = 0, \end{cases} \tag{8}$$

where $\pi_{\mathbf{K}}^{-1}$ refers to back-projection with camera intrinsics $\mathbf{K}$, $\mathbf{W}_t$ refers to the camera pose, and $\mathbf{h}_{\text{left}}, \mathbf{h}_{\text{right}}$ refer to the lifted 3D positions from the nearest visible timesteps before and after t. From each track, we initialize a *MoSca* node $\mathbf{v}^{(i)}$ using the lifted positions $\mathbf{h}_t$ as the translation part and the identity as the rotation, i.e., $\mathbf{Q}_t^{(i)} = [\mathbf{I}, \mathbf{h}_t^{(i)}]$, along with a predefined control radius r_{init}. In practice, we retain only a subset of the densely inferred 2D tracklets by uniformly resampling nodes based on the curve distance (Eq. 2).

Geometry Optimization. Starting from the initialized rotations and the invisible lines, we propagate the visible information to the unknowns through the *MoSca* topology $\mathcal{E}$ by optimizing a physics-inspired as-rigid-as-possible (ARAP) loss. Given two timesteps separated by a time interval Δ, we define the ARAP loss $\mathcal{L}_{\text{arap}}$ as:

$$\mathcal{L}_{\text{arap}} = \sum_{t=1}^{T} \sum_{m=1}^{M} \sum_{n \in \hat{\mathcal{E}}(m)} \lambda_{\text{l}} \left| \|\mathbf{t}_t^{(m)} - \mathbf{t}_t^{(n)}\| - \|\mathbf{t}_{t+\Delta}^{(m)} - \mathbf{t}_{t+\Delta}^{(n)}\| \right| + \lambda_{\text{c}} \left\| \mathbf{Q}_t^{-1\,(n)} \mathbf{t}_t^{(m)} - \mathbf{Q}_{t+\Delta}^{-1\,(n)} \mathbf{t}_{t+\Delta}^{(m)} \right\|, \tag{9}$$

where $\hat{\mathcal{E}}$ refers to a multi-level sub-sampled topology pyramid from $\mathcal{E}$ in *MoSca* (detailed in the Supplemental Material). The first term encourages the preservation of local distances in the neighborhood, and the second term preserves the local coordinates by involving the local frame $\mathbf{Q}$ in the optimization. We also enforce the temporal smoothness of the deformation by regularizing the velocity and acceleration:

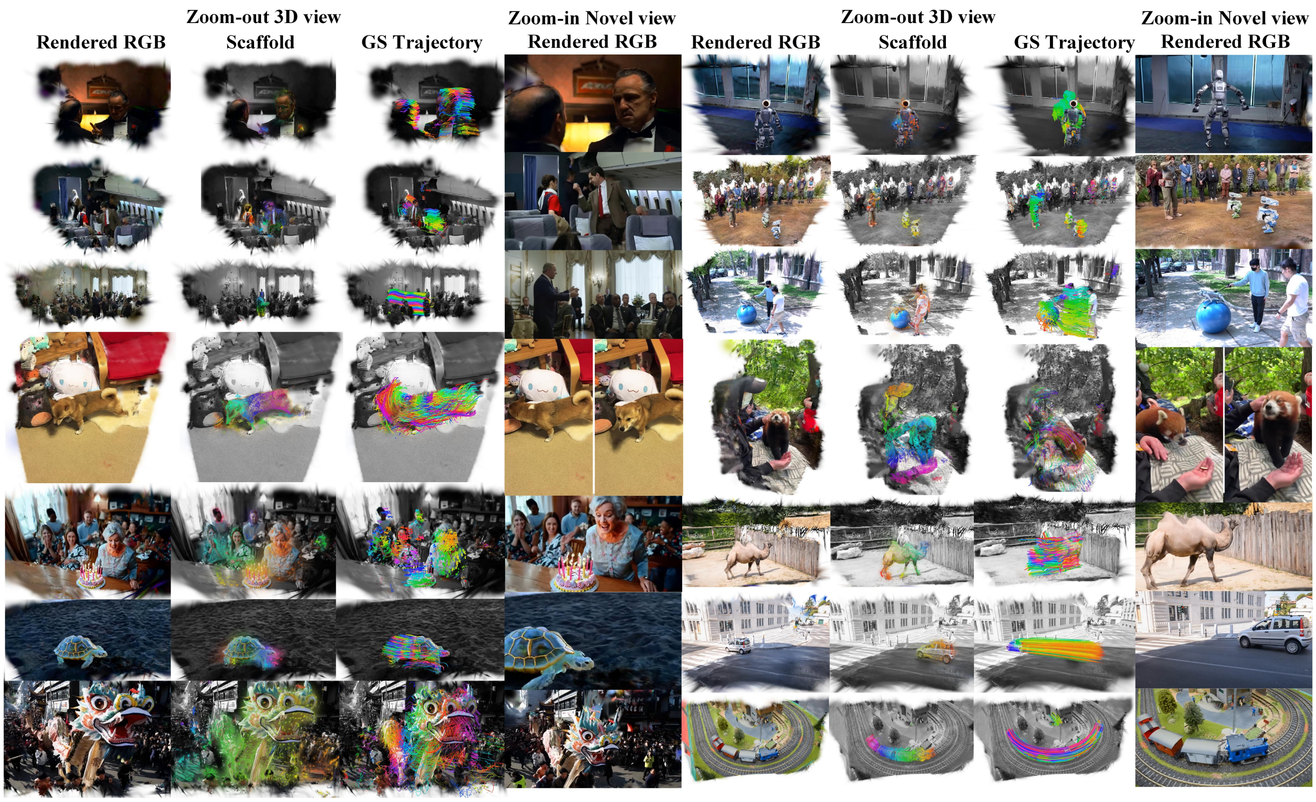

Figure 3. **In-the-wild videos**: *MoSca* can process a list of RGB frames and reconstruct the 4D scene from various types of videos.

$$\mathcal{L}_{\text{vel}} = \sum_{t=1}^{T}\sum_{m=1}^{M} \|\mathbf{t}_t^{(m)} - \mathbf{t}_{t+1}^{(m)}\| + \|\log(\mathbf{R}_t^{(m)}\mathbf{R}_{t+1}^{-1\,(m)})\|_F$$

$$\mathcal{L}_{\text{acc}} = \sum_{t=1}^{T}\sum_{m=1}^{M} \|\mathbf{t}_t^{(m)} - 2\mathbf{t}_{t+1}^{(m)} + t_{t+2}^{(m)}\| \tag{10}$$
$$+ \left|\|\log(\mathbf{R}_t^{(m)}\mathbf{R}_{t+1}^{-1\,(m)})\|_F - \|\log(\mathbf{R}_{t+1}^{(m)}\mathbf{R}_{t+2}^{-1\,(m)})\|_F\right|,$$

where $\|\log(\cdot)\|_F$ refers to the Frobenius norm of rotation logarithm (the axis-angle of the rotation). In summary, the objective of this geometric optimization in the third step of our system is $\mathcal{L}_{\text{geo}} = \lambda_{\text{arap}}\mathcal{L}_{\text{arap}} + \lambda_{\text{acc}}\mathcal{L}_{\text{acc}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}}$, and we only optimize rotations and invisible 3D translations, leaving the visible 3D positions unchanged to prevent degeneration.

3.2.4 Photometric Optimization of *MoSca*

Dynamic Scene Representation. An important feature of *MoSca* is that its global deformation field can transform points at any time globally, enabling the fusion of all observed video frames into a single coherent representation. In the final step of the system, the optimized *MoSca* collects 3D Gaussians initialized from back-projected foreground depth points at all timesteps. Formally:

$$\mathcal{G} = \{(\mu_j, R_j, s_j, o_j, c_j; t_j^{\text{ref}}, \Delta\mathbf{w}_j)\}_{j=1}^{N}, \tag{11}$$

where the first five attributes are the center, rotation, non-isotropic scales, opacity, and spherical harmonics of 3DGS [44], and the latter two are tailored for *MoSca*. Specifically, t_j^{ref} is the reference timestep—that is, the timestep at which the Gaussian is initialized from the back-projected depth; and $\Delta\mathbf{w}_j \in \mathbb{R}^K$ is the per-Gaussian learnable skinning weight correction. To obtain the complete geometry at a query timestep t, Gaussians from all timesteps are deformed to the query time t and fused:

$$\mathcal{G}(t) = \{(\mathbf{T}_j(t)\mu_j, \mathbf{T}_j(t)R_j, s_j, o_j, c_j) \mid$$
$$\mathbf{T}_j(t) = \mathcal{W}(\mu_j, \mathbf{w}(\mu_j, t_j^{\text{ref}}) + \Delta\mathbf{w}_j; t_j^{\text{ref}}, t)\}_{j=1}^{N} \tag{12}$$

where $\mathcal{W}$ is the deformation field defined in Eq.4, and $\mathbf{w}$ is the base RBF skinning weight defined in Eq.5. The static background is also represented as a standard 3DGS $\mathcal{H} = (\mu_j, R_j, s_j, o_j, c_j)_{j=1}^{H}$, which can be initialized by back-projecting the depth map using known camera parameters. Therefore, the final renderable dynamic scene at time t can be approximated by the union $\mathcal{G}(t) \cup \mathcal{H}$.

Photometric Optimization. The Gaussians described above can be rendered using a Gaussian Splatting-based differentiable renderer and optimized with depth and RGB rendering losses, along with the regularization losses from Sec. 3.2.3. To fully exploit the inferred tracklets, we also render a flow/track map by rasterizing the XYZ coordinates (replacing the RGB color with XYZ values) of each

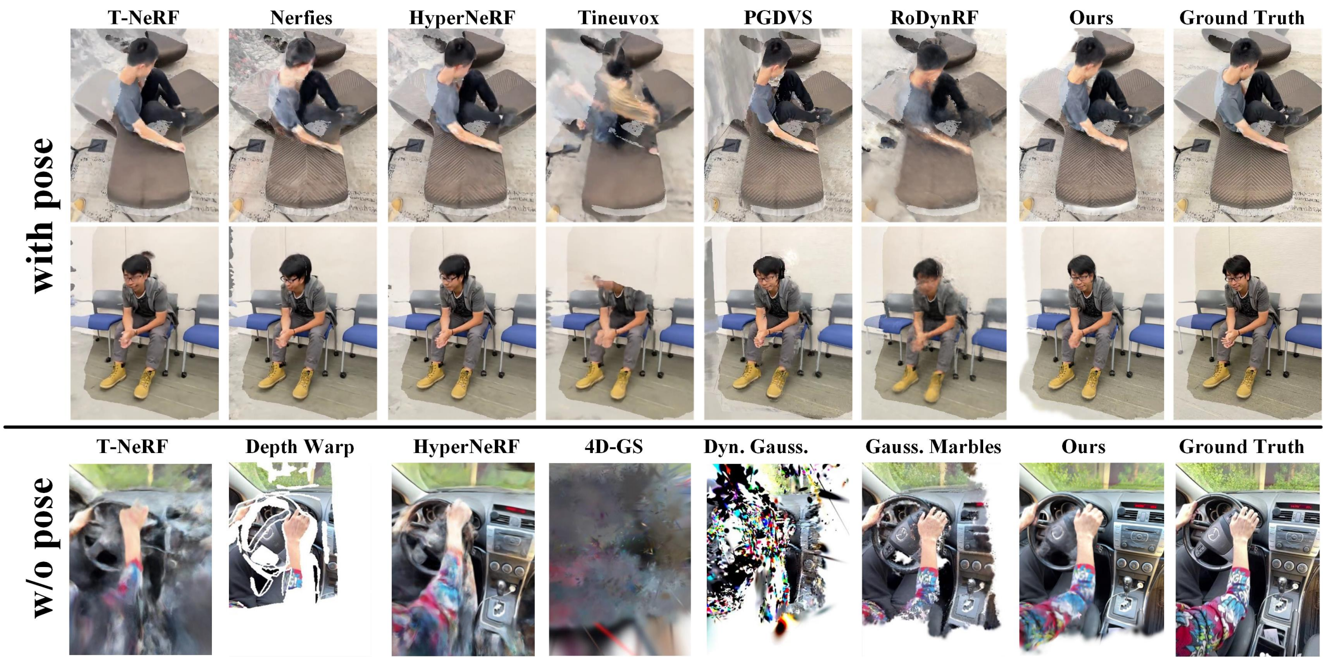

Figure 4. Visual comparison on DyCheck [30] under the settings with or without camera pose.

Gaussian at different timesteps. We supervise the flow/track map with the inferred 2D tracklets as a regularization loss $\mathcal{L}_{\text{track}}$ [99]. The final photometric step has a total objective:

$$\mathcal{L} = \lambda_{\text{rgb}}\mathcal{L}_{\text{rgb}} + \lambda_{\text{dep}}\mathcal{L}_{\text{dep}} + \lambda_{\text{track}}\mathcal{L}_{\text{track}} + \lambda_{\text{arap}}\mathcal{L}_{\text{arap}} + \lambda_{\text{acc}}\mathcal{L}_{\text{acc}} + \lambda_{\text{vel}}\mathcal{L}_{\text{vel}}. \quad (13)$$

Node Control. Similar to standard 3DGS Gaussian control techniques including gradient-based densification and reset-pruning simplification, we propose a novel control policy over the proposed *MoSca* nodes. To periodically densify nodes, we select Gaussians with high tracking-loss $\mathcal{L}_{\text{track}}$ induced gradients, subsample them, and convert them into new *MoSca* nodes. To clean the representation and prune the structure, we also periodically copy the dynamic foreground Gaussians from a randomly selected timestep into the static background and reset the foreground Gaussians to a low opacity. This simplifies unnecessary foreground Gaussians. We then prune nodes whose skinning weights toward all Gaussians fall below a threshold, indicating a limited contribution to deformation modeling.

Table 1. Comparison on DyCheck [30], group w-pose and w/o-pose means with or without camera pose and are averaged over all 7 scenes on the standard 2x resolution. Group SOM-5-1x means using the 5 scenes and 1x res. as in Shape-of-Motion [99].

	Method	mPSNR↑	mSSIM↑	mLPIPS↓
w-pose	T-NeRF [30]	16.96	0.577	0.379
	NSFF [56]	15.46	0.551	0.396
	Nerfies [74]	16.45	0.570	0.339
	HyperNeRF [75]	16.81	0.569	0.332
	PGDVS [118]	15.88	0.548	0.340
	DyPoint [119]	16.89	0.573	-
	DpDy [98]	-	0.559	0.516
	Dyn.Gauss. [67]	7.29	-	0.692
	4D GS [103]	13.64	-	0.428
	Gauss.Marbles [86]	16.72	-	0.413
	DyBluRF [11]	17.37	0.591	0.373
	CTNeRF [68]	17.69	0.531	-
	D-NPC [39]	16.41	0.582	0.319
	Shape-of-Motion [99]	17.32	0.598	0.296
	Ours	**19.32**	**0.706**	**0.264**
w/o-pose	RobustDynrf [66]	17.10	0.534	0.517
	Dyn.Gaussians [67]	7.60	-	0.704
	4D GS [103]	13.11	-	0.726
	Gaussian Marbles [86]	15.79	-	0.430
	Ours	18.84	0.676	0.289
	Ours (w. focal)	**19.02**	**0.683**	**0.279**
SOM-5-1x	Shape-of-Motion [99]	16.72	0.63	0.45
	Ours	**18.40**	**0.67**	**0.42**

4. Experiments

4.1. Novel View Synthesis

In-the-wild. One of the most significant results of *MoSca* is demonstrating that such an automatic dynamic rendering system can work effectively in real-world scenarios. In Fig. 3, we showcase reconstruction results on diverse in-the-wild **monocular** videos—including movie clips, internet videos, SORA-generated videos, and DAVIS[77] videos—demonstrating the effectiveness of *MoSca*.

DyCheck. To quantitatively evaluate our rendering results, we compare our method to others on the currently most challenging dataset – the iPhone DyCheck [30]. DyCheck features generic, diverse dynamic scenes captured with a handheld iPhone using realistic camera motions for train-

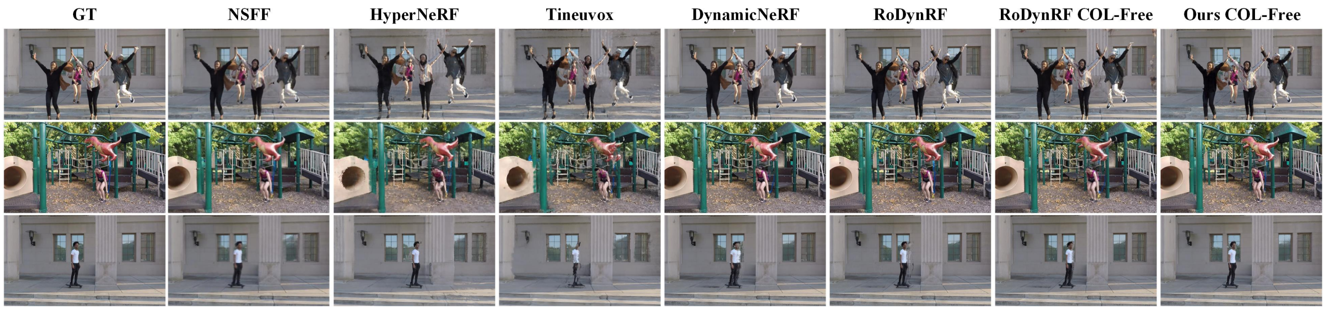

Figure 5. Visual comparison on NVIDIA dataset [112].

ing, and utilizes two static cameras at significantly different poses from the training views for testing. For a fair comparison with previous methods that exploit noisy LiDAR depth from the dataset, we use the iPhone's noisy LiDAR depth as the metric depth $\mathcal{D}$ and employ BootsTAPIR [22] for tracking. Since the camera parameters are optimized during training, during inference, we fix the scene representation and adjust the test camera poses to find the correct viewpoints. The quantitative results are reported in Tab. 1, and qualitative results are shown in Fig. 1. Due to the large deviation of the testing views from the training camera trajectory, most per-frame depth warping methods fail directly (e.g., see Fig.10 of Casual-FVS [51]). Similarly, local fusion methods exhibit large missing areas (e.g., PGDVS [118], Gaussian Marbles [86]), even though these missing areas are visible in other time steps. Some recent Gaussian-based methods like 4D-GS [103] also fail because they depend on strong multi-view stereo cues to reconstruct the scene. As shown in Tab. 1, we outperform all other methods by a large margin. We attribute this improvement to two factors: firstly, by leveraging powerful pre-trained 2D long-term trackers, our *MoSca* representation models long-term motion trajectories, enabling the global aggregation of observations across all timesteps, which leads to a more complete reconstruction. Secondly, the structured sparse motion graph design of *MoSca* facilitates optimization. Compared to dense Gaussian geometries, its compact and smoothly interpolated motion nodes significantly reduce the optimization space. Its topology enables the effective propagation of information to unobserved regions through ARAP regularization. Note that our system still performs well under the pose-free setup, as shown in the bottom group of Tab. 1.

Table 2. Comparison on NVIDIA [112], averaged over all scenes. "w/o" means without camera pose.

Method	PSNR	LPIPS	Method	PSNR	LPIPS
D-NeRF [78]	21.49	0.232	CTNeRF [68]	26.13	0.082
NR-NeRF [96]	19.69	0.323	DynPoint [119]	26.53	<u>0.068</u>
TiNeuVox [27]	19.74	0.285	D-NPC [39]	25.64	0.109
HyperNeRF [75]	17.60	0.367	RoDynRF [66]	25.89	**0.067**
NSFF [56]	24.33	0.199	RoDynRF [66] w/o	25.38	0.079
DynNeRF [29]	26.10	0.082	GaussianMarbles [86]	22.32	0.129
MonoNeRF [94]	25.62	0.106	**Ours**	**26.72**	0.070
4DGS [103]	21.45	0.199	**Ours w/o**	<u>26.54</u>	0.073
Casual-FVS [51]	24.57	0.081			

NVIDIA. We also evaluate *MoSca* on the widely used NVIDIA video dataset [112], following the protocol in RoDynRF [66]. As reported in Tab. 2 and Fig. 5, we achieve high PSNR and very competitive LPIPS results. Since the facing-forward, the small-baseline setting is relatively easier compared to the realistic DyCheck dataset, where most areas of the dynamic scene are visible in neighboring time frames, reducing the need for strong regularization and fusion of information in occluded areas – the advantages of *MoSca* are not fully showcased on NVIDIA videos.

4.2. Camera and Correspondence

Camera Pose. Another advantage of *MoSca* is its natural integration of camera solving, both geometrically through tracklet-based bundle adjustment and photometrically through rendering-based refinement. We quantitatively evaluate the camera pose estimation, a byproduct of our system, following MonST3R [115] on the SLAM dataset TUM-dynamics [88] and the synthetic Sintel dataset [12]. The camera pose errors are shown in Table 3. Although camera pose estimation is not the main focus of

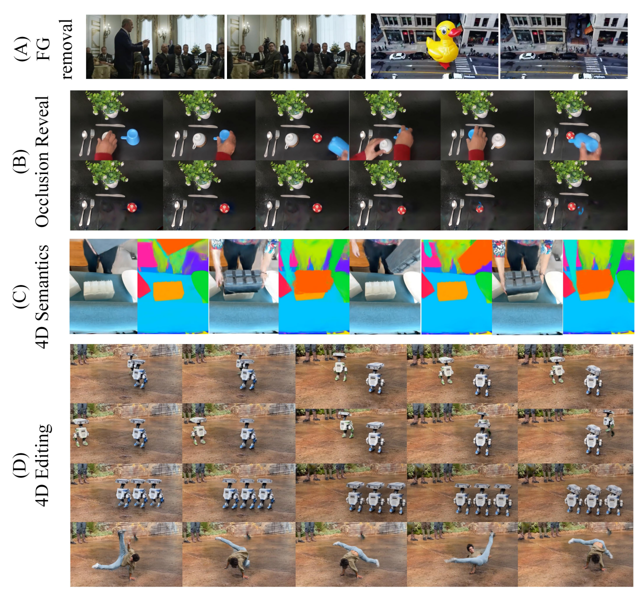

Figure 6. Application of *MoSca* reconstructed 4D scenes.

Table 3. Camera pose accuracy (* requires ground truth camera intrinsics as input)

Method	Sintel [12] ATE ↓	Sintel [12] RPE trans ↓	Sintel [12] RPE rot ↓	TUM-dynamics [88] ATE ↓	TUM-dynamics [88] RPE trans ↓	TUM-dynamics [88] RPE rot ↓
DROID-SLAM* [92]	0.175	0.084	1.912	-	-	-
DPVO* [93]	0.115	0.072	1.975	-	-	-
ParticleSfM [117]	0.129	**0.031**	0.535	-	-	-
LEAP-VO* [15]	**0.089**	0.066	1.250	0.068	**0.008**	1.686
Robust-CVD [49]	0.360	0.154	3.443	0.153	0.026	3.528
CasualSAM [116]	0.141	0.035	0.615	0.071	0.010	1.712
DUSt3R [100] w/ mask	0.417	0.250	5.796	0.083	0.017	3.567
MonST3R [115]	0.108	0.042	0.732	0.063	0.009	1.217
Ours	0.090	0.034	**0.312**	**0.031**	0.011	**0.426**

Full No M-level No DQB No Photo No GEO 8-nn Fuse

Figure 7. Visual comparison of ablation.

MoSca, it still achieves comparable or even superior performance compared to camera-pose-tailored SLAM-based and DuST3R-based methods. Notably, some of the SLAM systems in the table require known camera intrinsics, whereas *MoSca* does not.

Correspondence. One feature of *MoSca* is its ability to perform global fusion and provide dense correspondence. We quantitatively evaluate the correspondence tracking accuracy following DyCheck [30] and Gaussian Marbles [86]. Tab. 4 shows our state-of-the-art accuracy. Notably, *MoSca* is optimized starting from BootsTAPIR [22] on DyCheck, and we observe a significant improvement over the raw tracker after reconstruction optimization.

4.3. Ablation Study

We assess the effects of different components in our system in Tab. 5 and Fig. 7. We observe that both the geometric optimization and photometric optimization phases are critical. DQB contributes to smooth results, the multi-level topology pyramid enhances global rigidity and shape, and node control along with learnable skinning further improves the expressiveness of our system. Additionally, our system benefits from the global fusion of observations from every frame. We also verify the effectiveness of the tracking loss $\mathcal{L}$track. When $\mathcal{L}_{\text{track}}$ is not used, the PCK-T accuracy decreases from 0.824 to 0.737. In Tab. 6, we study how different foundation models affect performance. Note that Metric3D-v2 [34] and UniDepth [76] are entirely RGB-based and do not use LiDAR sensor information, leading to a reasonable decrease in performance. We report more specifications of our system in Tab. 7, where we observe near real-time inference FPS and the compactness of the *MoSca* nodes compared to the actual foreground GS used to model the scene.

Table 4. Correspondence on DyCheck [30] with PCK-T @0.05%

Methods	Nerfies[74]	HyperNeRf[75]	Dyn. Gauss. [67]	4D Gauss. [103]
PCK-T ↑	0.4	0.453	0.079	0.073
Methods	CoTracker[40]	Gauss.Marbles[86]	BootsTAPIR [22]	**Ours**
PCK-T ↑	0.803	0.806	0.779	**0.824**

Table 5. Ablation study on different components of the system.

Components	mPSNR	mSSIM	mLPIPs
Full model	**19.32**	0.706	**0.264**
No node control	19.28	**0.707**	0.267
No learnable skinning correction	19.27	**0.707**	0.267
No dual quaternion blending	19.18	0.701	0.276
No multi-level topology	19.14	0.701	0.270
No geometric optimizaiton stage	18.85	0.693	0.287
No photometric optimization stage	13.71	0.480	0.763
Only fuse 4 neighboring frames	16.96	0.663	0.344
Only fuse 8 neighboring frames	17.26	0.664	0.346

Table 6. Ablation study on different priors on DyCheck [30].

Tracker	**BootsTAPIR** [22]		CoTracker-v3 [41]		SpaTracker [106]	
Depth	mPSNR	mLPIPs	mPSNR	mLPIPs	mPSNR	mLPIPs
LIDAR	19.32	0.264	19.55	0.243	19.32	0.259
Metric3D-v2 [34]	17.05	0.331	17.02	0.320	17.60	0.301
UniDepth [76]	17.12	0.323	17.42	0.299	17.61	0.300

Table 7. More specs of *MoSca* on DyCheck [30] (averaged)

FPS (2x res)	Num of fg GS	Num of nodes	Ratio: #GS/#nodes
37.823	106596	3177	46.105

4.4. Applications

In-the-wild 4D reconstruction enables many interesting applications, as shown in Fig. 6. For example, we can remove the moving foreground (Figure 6-A), or remove occluders in an extremely challenging cup-game video to look through and see where the ball goes (Figure 6-B). Video object segmentation from DEVA [17] can be lifted and baked into 4D to produce novel view semantic videos (Figure 6-C). Finally, the 4D video can be edited in flexible ways, as shown in Figure 6-D. We believe that *MoSca* will provide the community with many more possibilities for future applications.

5. Limitations and Conclusion

Limitations. While *MoSca* achieves state-of-the-art performance on standard benchmarks and can operate on some in-the-wild videos, several limitations remain. (1) Our method relies on accurate 2D long-term tracks and depth estimation, indicating that improvements in these areas are crucial for enhancing our performance. (2) Our current framework only reconstructs areas that are visible at some point in the video; it would be advantageous to incorporate large-scale 2D/video diffusion priors to hallucinate areas that are never visible. (3) Another important issue for future work is the correct modeling of lighting effects such as shadows, reflections, liquids, and changes in exposure. These effects cannot be explained by deformation alone and may cause artifacts in the background.

In summary, this paper takes a step toward reconstruction and rendering from monocular in-the-wild casual videos We hope this small step could inspire future exploration toward understanding our dynamic physical world.

Acknowledgements. The authors appreciate the support of the gift from AWS AI to Penn Engineering's ASSET Center for Trustworthy AI; and the support of the following grants: NSF IIS-RI 2212433, NSF FRR 2220868 awarded to UPenn, ARL grant W911NF-21-2-0104 and a Vannevar Bush Faculty Fellowship awarded to Stanford University.

The authors thank Minh-Quan Viet Bui and the authors of DyBluRF, Xiaoming Zhao and the authors of PGDVS for providing their per-scene evaluation metrics on DyCheck dataset.

References

[1] ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20364–20373, 2022. 2

[2] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[3] Aayush Bansal, Minh Vo, Yaser Sheikh, Deva Ramanan, and Srinivasa Narasimhan. 4d visualization of dynamic events from unconstrained multi-view videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[4] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2

[5] Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. X-fields: Implicit neural view-, light- and time-image interpolation. *SIGGRAPH Asia*, 2020. 2

[6] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2

[7] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3d faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, pages 187–194, 1999. 2

[8] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *European Conference on Computer Vision*, pages 561–578. Springer, 2016. 2

[9] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 2

[10] Aljaz Bozic, Pablo Palafox, Michael Zollöfer, Angela Dai, Justus Thies, and Matthias Nießner. Neural non-rigid tracking. In *Advances in Neural Information Processing Systems*, pages 18765–18775, 2020. 2

[11] Minh-Quan Viet Bui, Jongmin Park, Jihyong Oh, and Munchurl Kim. Dyblurf: Dynamic deblurring neural radiance fields for blurry monocular video. *arXiv preprint arXiv:2312.13528*, 2023. 2, 6

[12] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*, pages 611–625. Springer, 2012. 7, 8

[13] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. *arXiv preprint arXiv:2301.09632*, 2023. 2

[14] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 333–350. Springer, 2022. 2

[15] Weirong Chen, Le Chen, Rui Wang, and Marc Pollefeys. Leap-vo: Long-term effective any point tracking for visual odometry. 2024. 8

[16] Yufan Chen, Lizhen Wang, Qijing Li, Hongjiang Xiao, Shengping Zhang, Hongxun Yao, and Yebin Liu. Monogaussianavatar: Monocular gaussian point-based head avatar. *arXiv preprint arXiv:2312.04558*, 2023. 2

[17] Ho Kei Cheng, Seoung Wug Oh, Brian Price, Alexander Schwing, and Joon-Young Lee. Tracking anything with decoupled video segmentation. In *ICCV*, 2023. 8

[18] Wen-Hsuan Chu, Lei Ke, and Katerina Fragkiadaki. Dreamscene4d: Dynamic multi-object scene generation from monocular videos. *arXiv preprint arXiv:2405.02280*, 2024. 2

[19] Brian Curless and Marc Levoy. A volumetric method for building complex models from range images. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 303–312, 1996. 2

[20] Konstantinos Daniilidis. Hand-eye calibration using dual quaternions. *The International Journal of Robotics Research*, 18(3):286–298, 1999. 3

[21] Devikalyan Das, Christopher Wewer, Raza Yunus, Eddy Ilg, and Jan Eric Lenssen. Neural parametric gaussians for monocular non-rigid object reconstruction. *arXiv preprint arXiv:2312.01196*, 2023. 2

[22] 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. 3, 7, 8

[23] Mingsong Dou, Jonathan Taylor, Henry Fuchs, Andrew Fitzgibbon, and Shahram Izadi. 3d scanning deformable objects with a single rgbd sensor. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 493–501, 2015. 2

[24] 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. IEEE Computer Society, 2021. 2

[25] Yuanxing Duan, Fangyin Wei, Qiyu Dai, Yuhang He, Wenzheng Chen, and Baoquan Chen. 4d gaussian splatting: Towards efficient novel view synthesis for dynamic scenes. *arXiv preprint arXiv:2402.03307*, 2024. 2

[26] Bardienus P Duisterhof, Zhao Mandi, Yunchao Yao, Jia-Wei Liu, Mike Zheng Shou, Shuran Song, and Jeffrey Ichnowski. Md-splatting: Learning metric deformation from 4d gaussians in highly deformable scenes. *arXiv preprint arXiv:2312.00583*, 2023. 2

[27] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 2, 7

[28] Sara Fridovich-Keil, Giacomo Meanti, Frederik Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. *arXiv preprint arXiv:2301.10241*, 2023. 2

[29] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021. 2, 7

[30] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. *Advances in Neural Information Processing Systems*, 35:33768–33780, 2022. 1, 2, 6, 8

[31] Wei Gao and Russ Tedrake. Surfelwarp: Efficient non-volumetric single view dynamic reconstruction. In *Robotics: Science and Systems (RSS)*, 2018. 2

[32] Adam W. Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. 2022. 2

[33] Liangxiao Hu, Hongwen Zhang, Yuxiang Zhang, Boyao Zhou, Boning Liu, Shengping Zhang, and Liqiang Nie. Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. *arXiv preprint arXiv:2312.02134*, 2023. 2

[34] Mu Hu, Wei Yin, Chi Zhang, Zhipeng Cai, Xiaoxiao Long, Hao Chen, Kaixuan Wang, Gang Yu, Chunhua Shen, and Shaojie Shen. Metric3d v2: A versatile monocular geometric foundation model for zero-shot metric depth and surface normal estimation. 2024. 3, 8

[35] Shoukang Hu and Ziwei Liu. Gauhuman: Articulated gaussian splatting from monocular human videos. *arXiv preprint arXiv:2312.02973*, 2023. 2

[36] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. 3

[37] 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. *arXiv preprint arXiv:2312.14937*, 2023. 2

[38] Yan-Bin Jia. Dual quaternions. 3

[39] Moritz Kappel, Florian Hahlbohm, Timon Scholz, Susana Castillo, Christian Theobalt, Martin Eisemann, Vladislav Golyanik, and Marcus Magnor. D-npc: Dynamic neural point clouds for non-rigid view synthesis from monocular video. *arXiv preprint arXiv:2406.10078*, 2024. 6, 7

[40] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker: It is better to track together. *arXiv preprint arXiv:2307.07635*, 2023. 2, 8

[41] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. 2024. 3, 8

[42] Kai Katsumata, Duc Minh Vo, and Hideki Nakayama. An efficient 3d gaussian representation for monocular/multi-view dynamic scenes. *arXiv preprint arXiv:2311.12897*, 2023. 2

[43] Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Skinning with dual quaternions. In *Proceedings of the 2007 symposium on Interactive 3D graphics and games*, pages 39–46, 2007. 3

[44] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. 2023. 2, 5

[45] Leonid Keselman and Martial Hebert. Approximate differentiable rendering with algebraic surfaces. In *European Conference on Computer Vision*, pages 596–614. Springer, 2022.

[46] Leonid Keselman and Martial Hebert. Flexible techniques for differentiable rendering with 3d gaussians. *arXiv preprint arXiv:2308.14737*, 2023. 2

[47] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 2

[48] Muhammed Kocabas, Jen-Hao Rick Chang, James Gabriel, Oncel Tuzel, and Anurag Ranjan. Hugs: Human gaussian splats. *arXiv preprint arXiv:2311.17910*, 2023. 2

[49] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 8

[50] Agelos Kratimenos, Jiahui Lei, and Kostas Daniilidis. Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. *arXiv preprint arXiv:2312.00112*, 2023. 2

[51] Yao-Chih Lee, Zhoutong Zhang, Kevin Blackburn-Matzen, Simon Niklaus, Jianming Zhang, Jia-Bin Huang, and Feng Liu. Fast view synthesis of casual videos. *arXiv preprint arXiv:2312.02135*, 2023. 1, 2, 7

[52] Jiahui Lei, Yufu Wang, Georgios Pavlakos, Lingjie Liu, and Kostas Daniilidis. Gart: Gaussian articulated template models. *arXiv preprint arXiv:2311.16099*, 2023. 2

[53] Hao Li, Robert W Sumner, and Mark Pauly. Global correspondence optimization for non-rigid registration of depth scans. *Computer Graphics Forum*, 27(5):1421–1430, 2008. 2

[54] Mingwei Li, Jiachen Tao, Zongxin Yang, and Yi Yang. Human101: Training 100+ fps human gaussians in 100s from 1 view. *arXiv preprint arXiv:2312.15258*, 2023. 2

[55] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[56] 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*, pages 6498–6508, 2021. 2, 6, 7

[57] Zhan Li, Zhang Chen, Zhong Li, and Yi Xu. Spacetime gaussian feature splatting for real-time dynamic view synthesis. *arXiv preprint arXiv:2312.16812*, 2023. 2

[58] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering, 2023. 2

[59] Yiqing Liang, Numair Khan, Zhengqin Li, Thu Nguyen-Phuoc, Douglas Lanman, James Tompkin, and Lei Xiao. Gaufre: Gaussian deformation fields for real-time dynamic novel view synthesis. *arXiv preprint arXiv:2312.11458*, 2023. 2

[60] Haotong Lin, Sida Peng, Zhen Xu, Tao Xie, Xingyi He, Hujun Bao, and Xiaowei Zhou. High-fidelity and real-time novel view synthesis for dynamic scenes. In *SIGGRAPH Asia Conference Proceedings*, 2023. 2

[61] Youtian Lin, Zuozhuo Dai, Siyu Zhu, and Yao Yao. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. *arXiv preprint arXiv:2312.03431*, 2023. 2

[62] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 2

[63] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[64] Qingming Liu, Yuan Liu, Jiepeng Wang, Xianqiang Lv, Peng Wang, Wenping Wang, and Junhui Hou. Modgs: Dynamic gaussian splatting from causually-captured monocular videos. *arXiv preprint arXiv:2406.00434*, 2024. 2

[65] Xinqi Liu, Chenming Wu, Jialun Liu, Xing Liu, Jinbo Wu, Chen Zhao, Haocheng Feng, Errui Ding, and Jingdong Wang. Gva: Reconstructing vivid 3d gaussian avatars from monocular videos, 2024. 2

[66] 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*, pages 13–23, 2023. 1, 2, 4, 6, 7

[67] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. *arXiv preprint arXiv:2308.09713*, 2023. 2, 6, 8

[68] Xingyu Miao, Yang Bai, Haoran Duan, Yawen Huang, Fan Wan, Yang Long, and Yefeng Zheng. Ctnerf: Cross-time transformer for dynamic neural radiance field from monocular video. *arXiv preprint arXiv:2401.04861*, 2024. 2, 6, 7

[69] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2

[70] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 2

[71] Richard A Newcombe, Dieter Fox, and Steven M Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 343–352, 2015. 2

[72] OpenAI. Gpt-4 technical report, 2023. https://openai.com/research/gpt-4. 2

[73] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 2

[74] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2, 6, 8

[75] 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. *arXiv preprint arXiv:2106.13228*, 2021. 2, 6, 7, 8

[76] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. *arXiv preprint arXiv:2403.18913*, 2024. 2, 3, 8

[77] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 6

[78] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2, 7

[79] Zhiyin Qian, Shaofei Wang, Marko Mihajlovic, Andreas Geiger, and Siyu Tang. 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. *arXiv preprint arXiv:2312.09228*, 2023. 2

[80] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2

[81] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *European Conference on Computer Vision*, pages 573–586. Springer, 2012. 2

[82] Alfredo Rivero, ShahRukh Athar, Zhixin Shu, and Dimitris Samaras. Rig3dgs: Creating controllable portraits from casual monocular videos. *arXiv preprint arXiv:2402.03723*, 2024. 2

[83] Jenny Seidenschwarz, Qunjie Zhou, Bardienus Duisterhof, Deva Ramanan, and Laura Leal-Taixé. Dynomo: Online point tracking by dynamic online monocular gaussian reconstruction. *arXiv preprint arXiv:2409.02104*, 2024. 2

[84] Zhijing Shao, Zhaolong Wang, Zhuang Li, Duotun Wang, Xiangru Lin, Yu Zhang, Mingming Fan, and Zeyu Wang. Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. *arXiv preprint arXiv:2403.05087*, 2024. 2

[85] 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*, 2023. 2

[86] Colton Stearns, Adam W. Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In *ArXiv*, 2024. 2, 6, 7, 8

[87] Timo Stich, Christian Linz, Georgia Albuquerque, and Marcus Magnor. View and time interpolation in image space. *Computer Graphics Forum*, 2008. 2

[88] Jürgen Sturm, Nikolas Engelhard, Felix Endres, Wolfram Burgard, and Daniel Cremers. A benchmark for the evaluation of rgb-d slam systems. In *2012 IEEE/RSJ international conference on intelligent robots and systems*, pages 573–580. IEEE, 2012. 7, 8

[89] Robert W Sumner, Johannes Schmid, and Mark Pauly. Embedded deformation for shape manipulation. In *ACM siggraph 2007 papers*, pages 80–es. 2007. 2

[90] David Svitov, Pietro Morerio, Lourdes Agapito, and Alessio Del Bue. Haha: Highly articulated gaussian human avatars with textured mesh prior. *arXiv preprint arXiv:2404.01053*, 2024. 2

[91] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 4

[92] Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021. 8

[93] Zachary Teed, Lahav Lipson, and Jia Deng. Deep patch visual odometry. *Advances in Neural Information Processing Systems*, 36, 2024. 8

[94] Fengrui Tian, Shaoyi Du, and Yueqi Duan. Mononerf: Learning a generalizable dynamic radiance field from monocular videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17903–17913, 2023. 2, 7

[95] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018. 2

[96] 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. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 2, 7

[97] Chaoyang Wang, Ben Eckart, Simon Lucey, and Orazio Gallo. Neural trajectory fields for dynamic novel view synthesis. *arXiv preprint arXiv:2105.05994*, 2021. 2

[98] Chaoyang Wang, Peiye Zhuang, Aliaksandr Siarohin, Junli Cao, Guocheng Qian, Hsin-Ying Lee, and Sergey Tulyakov. Diffusion priors for dynamic view synthesis from monocular videos. *arXiv preprint arXiv:2401.05583*, 2024. 6

[99] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. 2024. 2, 6

[100] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024. 8

[101] Jing Wen, Xiaoming Zhao, Zhongzheng Ren, Alexander G Schwing, and Shenlong Wang. Gomavatar: Efficient animatable human modeling from monocular video using gaussians-on-mesh. *arXiv preprint arXiv:2404.07991*, 2024. 2

[102] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16210–16220, 2022. 2

[103] 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. *arXiv preprint arXiv:2310.08528*, 2023. 2, 6, 7, 8

[104] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. D^2 nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. *arXiv preprint arXiv:2205.15838*, 2022. 2

[105] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021. 2

[106] Yuxi Xiao, Qianqian Wang, Shangzhan Zhang, Nan Xue, Sida Peng, Yujun Shen, and Xiaowei Zhou. Spatialtracker:

Tracking any 2d pixels in 3d space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 3, 8

[107] Gengshan Yang, Deqing Sun, Varun Jampani, Daniel Vlasic, Forrester Cole, Huiwen Chang, Deva Ramanan, William T Freeman, and Ce Liu. Lasr: Learning articulated shape reconstruction from a monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4925–4935, 2021. 2

[108] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22247–22257, 2022. 2

[109] 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*, pages 10371–10381, 2024. 2

[110] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. *arXiv preprint arXiv:2309.13101*, 2023. 2

[111] Zeyu Yang, Hongye Yang, Zijie Pan, Xiatian Zhu, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*, 2023. 2

[112] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5336–5345, 2020. 2, 7

[113] Meng You and Junhui Hou. Decoupling dynamic monocular videos for dynamic view synthesis. *arXiv preprint arXiv:2304.01716*, 2023. 2

[114] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient and compact surface reconstruction in unbounded scenes. *arXiv preprint arXiv:2404.10772*, 2024. 2

[115] Junyi Zhang, Charles Herrmann, Junhwa Hur, Varun Jampani, Trevor Darrell, Forrester Cole, Deqing Sun, and Ming-Hsuan Yang. Monst3r: A simple approach for estimating geometry in the presence of motion. *arXiv preprint arxiv:2410.03825*, 2024. 7, 8

[116] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T Freeman. Structure and motion from casual videos. In *European Conference on Computer Vision*, pages 20–37. Springer, 2022. 8

[117] Wang Zhao, Shaohui Liu, Hengkai Guo, Wenping Wang, and Yong-Jin Liu. Particlesfm: Exploiting dense point trajectories for localizing moving cameras in the wild. In *European conference on computer vision (ECCV)*, 2022. 8

[118] Xiaoming Zhao, Alex Colburn, Fangchang Ma, Miguel Angel Bautista, Joshua M. Susskind, and Alexander G. Schwing. Pseudo-generalized dynamic view synthesis from a video, 2024. 2, 6, 7

[119] Kaichen Zhou, Jia-Xing Zhong, Sangyun Shin, Kai Lu, Yiyuan Yang, Andrew Markham, and Niki Trigoni. Dynpoint: Dynamic neural point for view synthesis. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 6, 7

[120] C. Lawrence Zitnick, Sing Bing Kang, Matthew Uyttendaele, Simon Winder, and Richard Szeliski. High-quality video view interpolation using a layered representation. *ACM Transactions on Graphics (TOG)*, 2004. 2

[121] Michael Zollhöfer, Matthias Nießner, Shahram Izadi, Christoph Rehmann, Christopher Zach, Matthew Fisher, Chenglei Wu, Andrew Fitzgibbon, Charles Loop, Christian Theobalt, et al. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics (ToG)*, 33 (4):1–12, 2014. 2