

MotionScale: Reconstructing Appearance, Geometry, and Motion of Dynamic Scenes with Scalable 4D Gaussian Splatting

Haoran Zhou Gim Hee Lee

Department of Computer Science, National University of Singapore

haoran.zhou@u.nus.edu, gimhee.lee@nus.edu.sg



Figure 1. Visualization of a reconstructed dynamic scene and extracted moving objects. Given a single monocular video as input, **MotionScale** reconstructs a 4D scene representation that effectively captures photorealistic appearance, accurate 3D geometry, diverse human motion. Refer to the supplementary material for video results and additional examples.

Abstract

Realistic reconstruction of dynamic 4D scenes from monocular videos is essential for understanding the physical world. Despite recent progress in neural rendering, existing methods often struggle to recover accurate 3D geometry and temporally consistent motion in complex environments. To address these challenges, we propose *MotionScale*, a 4D Gaussian Splatting framework that scales efficiently to large scenes and extended sequences while maintaining high-fidelity structural and motion coherence. At the core of our approach is a scalable motion field parameterized by cluster-centric basis transformations that adaptively expand to capture diverse and evolving motion patterns. To ensure robust reconstruction over long durations, we introduce a progressive optimization strategy comprising two decoupled propagation stages: 1) A background extension stage that adapts to newly visible regions, refines camera poses, and explicitly models transient shadows; 2) A foreground propagation stage that en-

forces motion consistency through a specialized three-stage refinement process. Extensive experiments on challenging real-world benchmarks demonstrate that *MotionScale* significantly outperforms state-of-the-art methods in both reconstruction quality and temporal stability. Project page: <https://hrzhou2.github.io/motion-scale-web/>.

1. Introduction

Understanding and reconstructing dynamic 4D scenes is a pivotal challenge in computer vision, essential for enabling machines to perceive and interact with the physical world. Recently, the emergence of geometric foundation models has revolutionized the recovery of 3D structures from unconstrained images. Frameworks such as DUST3R [45] and VGGT [42] have demonstrated remarkable capability in estimating dense correspondences and inferring underlying 3D geometry from sparse or monocular observations. These models provide powerful geometric and motion pri-

ors that generalize across diverse environments, underpinning a wide range of downstream applications, including autonomous driving and motion forecasting [8, 38], AR/VR content creation [37] and immersive telepresence [11].

Despite the strong priors provided by these 2D models, high-fidelity 4D reconstruction, which aims to recover synchronized appearance, geometry, and motion across dynamic environments, remains an open challenge. Existing approaches [2, 9, 24, 29, 30, 34, 49, 54] based on Neural Radiance Fields (NeRFs) [25] and 3D Gaussian Splatting (3DGS) [18] have achieved remarkable success in photorealistic synthesis for static or mildly dynamic settings with dense multi-view supervision. However, scaling these representations to unconstrained, in-the-wild environments remains a significant bottleneck. More recent works [12, 28, 43, 46, 48] have begun to integrate efficient 4D Gaussian Splatting with 2D geometric and motion priors to reconstruct scenes from casual monocular video. While these methods yield plausible view synthesis from observed viewpoints, they frequently suffer from geometric distortion and lack temporal coherence across extended sequences. We identify two primary bottlenecks hindering current state-of-the-art approaches: 1) *Underconstrained Geometry*: Supervision predominantly relies on view-dependent appearance signals, which lack the capacity to enforce strict 3D structural consistency for dynamic objects. 2) *Accumulated Temporal Drift*: Motion models often rely on 2D tracking priors that lack 3D awareness. Over long-duration sequences, the representation inevitably accumulate errors, resulting in geometric collapse and inconsistent motion trajectories.

To address these challenges, we propose MotionScale, a scalable 4D Gaussian Splatting framework designed for the high-fidelity reconstruction of large-scale dynamic scenes. Unlike prior methods that rely on global deformation fields or fixed-capacity architectures, MotionScale introduces a cluster-centric motion representation that adaptively expands to capture diverse motion patterns across space and time. This design allows our motion field to scale with scene complexity while maintaining both spatio-temporal consistency and computational efficiency. Furthermore, we develop a progressive optimization strategy that scales seamlessly to unseen frames by effectively incorporating newly visible regions and refining motion trajectories across extended sequences. We evaluate MotionScale on several challenging real-world benchmarks, demonstrating that it achieves state-of-the-art reconstruction quality and motion consistency, significantly outperforming existing 4D Gaussian Splatting methods. Our main contributions are:

- A scalable 4D Gaussian Splatting framework, MotionScale, that effectively reconstructs large-scale dynamic scenes with accurate geometry, photorealistic appearance, and coherent motion across long sequences.

- A cluster-centric motion field that adaptively expands to model complex and unbounded motion patterns.
- A progressive optimization strategy featuring decoupled foreground and background propagation stages that ensure stable convergence and temporal coherence.

2. Related Work

2D vision models for 3D scene understanding. Recent progress in 2D foundation models has greatly advanced 3D scene understanding, which provides strong per-frame priors for tasks such as depth estimation, point tracking, and semantic segmentation [19, 36]. In monocular depth estimation, large-scale pre-trained models now deliver high-quality per-frame depth with strong zero-shot generalization [1, 3, 35, 51, 52]. Among these methods, Depth Anything V2 [52] is a large-scale depth model with up to 1.3B parameters. Trained on 595K synthetic labeled images and over 62M unlabeled real images, it produces significantly finer and more robust depth predictions than its predecessor [51]. However, although Depth Anything models produce high-quality depth maps for individual images, they do not enforce consistent scene geometry across views. This limitation has motivated methods that recover metric-scale 3D geometry from single-view observations [23, 32, 33, 44]. Specifically, MoGe [44] addresses single-view geometry by directly predicting a dense 3D point map from a single image, which yields accurate and fine-grained reconstruction of scene structure.

For correspondence and motion, foundation models now enable tracking of arbitrary points throughout the entire video. This goes beyond traditional optical flow methods, which are limited to short-term frame-to-frame correspondences [7, 13, 14, 40, 41]. The Tracking Any Point (TAP) paradigm was first introduced by TAP-Vid [4], which established a benchmark and proposed TAP-Net as a baseline model for point tracking. Building on this foundation, many methods have extended the TAP paradigm [5, 6, 15, 16, 57]. Among these methods, TAPIR [5] improves point tracking with a two-stage network that matches and refines query point positions across video frames. Another line of work is CoTracker [15], which introduces a transformer-based architecture for joint tracking of dense point sets over long video sequences and uses proxy tokens to improve efficiency. CoTracker3 [16] further simplifies the architecture and adopts a semi-supervised training strategy on real unlabeled videos. It reaches state-of-the-art performance while using substantially less training data. Furthermore, SpatialTracker [50] extends point tracking into 3D and jointly estimates depth and motion to maintain temporally consistent point trajectories in dynamic scenes. Beyond single tasks, joint feed-forward geometry models such as VGGT, π^3 and MapAnything infer camera poses, depth or point-maps and multi-view structure directly from images [17, 42, 47].

4D Scene Reconstruction. Reconstructing dynamic 3D scenes and generating novel views (NVS) is an active research area in computer vision. Neural Radiance Fields (NeRF) [25] pioneered high-fidelity view synthesis for static scenes, and subsequent dynamic NeRFs [2, 9, 21, 26, 29, 30, 34] extend this framework with canonical-space fields and time-varying deformations to model dynamic scenes. However, NeRF-based approaches require intensive volumetric optimization and are computationally expensive and inefficient for real-time rendering. This limitation has motivated point-based radiance representations that improve efficiency. 3D Gaussian Splatting (3DGS) [18] marked a major milestone by representing scenes as a set of 3D Gaussians and achieving real-time photorealistic rendering. Recent work builds on the expressive 3DGS representation and extends it to dynamic scenes [12, 24, 28, 43, 46, 48, 49, 54] with the capability of modeling complex geometry and motion from casual in-the-wild videos. These approaches significantly improve training and rendering efficiency as well as visual quality for dynamic novel view synthesis, yet they still face challenges in recovering accurate geometry, handling complex motion, and scaling to large-scale scenes and long video sequences.

3. Method

Overview. Fig. 2 provides an overview of MotionScale. Building on 3D Gaussian Splatting (3DGS) (*cf.* Sec. 3.1), we represent dynamic scenes as a set of canonical 3D Gaussians governed by a scalable motion field that models complex and diverse motion patterns while maintaining spatio-temporal consistency (*cf.* Sec. 3.2). The motion field leverages cluster-based basis transformations to parameterize the dynamics of local regions and employs an adaptive control strategy that autonomously expands or prunes clusters to achieve scalability. MotionScale adopts a progressive optimization strategy that jointly refines appearance, geometry, and motion while scaling seamlessly to long video sequences (*cf.* Sec. 3.3). The optimization proceeds via temporal propagation across frames, where a *background extension* stage refines unseen regions, camera poses, and shadows, while a *foreground propagation* stage employs a three-stage refinement to enforce long-term motion coherence.

3.1. Preliminary: 3D Gaussian Splatting

We build upon 3D Gaussian Splatting (3DGS) [18] as the foundational representation of our method. Standard 3DGS models a static scene as a collection of N Gaussian primitives defined in a canonical coordinate frame, $g_i^0 = \{\mu_i^0, \mathbf{R}_i^0, \mathbf{s}_i, o_i, \mathbf{c}_i\}$, where $i = 1, \dots, N$. Here, $\mu_i^0 \in \mathbb{R}^3$ and $\mathbf{R}_i^0 \in \text{SO}(3)$ represent the center position and orientation of the i -th Gaussian in 3D space. The attributes $\mathbf{s}_i \in \mathbb{R}^3$, $o_i \in \mathbb{R}$, and $\mathbf{c}_i \in \mathbb{R}^3$ denote its scale, opacity, and color, respectively. During rendering, for a given pixel $\mathbf{p}$ in

the image view $\mathbf{I}$ with extrinsic matrix $\mathbf{E}$ and intrinsic matrix $\mathbf{K}$, the pixel color $\mathbf{I}(\mathbf{p})$ is obtained by alpha-blending the colors of all intersected 3D Gaussians as:

$$\mathbf{I}(\mathbf{p}) = \sum_{i \in H(\mathbf{p})} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (1)$$

where $H(\mathbf{p})$ denotes the set of Gaussians projected at pixel $\mathbf{p}$. Each Gaussian contributes to the final color according to its opacity term, $\alpha_i = o_i \cdot \exp(-\frac{1}{2}(\mathbf{p} - \mu_i^{2d})^\top (\Sigma_i^{2d})^{-1}(\mathbf{p} - \mu_i^{2d}))$, where μ_i^{2d} and Σ_i^{2d} represent projected 2D mean and covariance on the image plane. Building on this, we represent a dynamic scene as a set of canonical 3D Gaussians $\{g_i^0\}$ and a time-dependent motion field that maps each Gaussian to its state at time t .

3.2. Scalable Motion Field

A fundamental challenge in large-scale dynamic reconstruction lies in designing a motion representation that is both expressive and scalable. Conventional deformation fields often rely on fixed-capacity architectures, such as global MLPs or a predefined set of temporal bases, which struggle to resolve localized, diverse object motions. In contrast, we propose a *cluster-based motion field* that enables an adaptive and scalable allocation of model capacity. Specifically, the dynamic Gaussians $\mathcal{G}_d = \{g_i\}_{i=1}^{N_d}$ are partitioned into K disjoint clusters $\{\mathcal{C}_k\}_{k=1}^K$. For each cluster $\mathcal{C}_k$, we define a hierarchical motion model consisting of a global transformation and a set of local refinement bases. The global transformation $\mathbf{G}_k^t = [\mathbf{R}_{k,g}^t \mid \mathbf{t}_{k,g}^t] \in \text{SE}(3)$ captures the global rigid movement of the entire cluster. To model localized non-rigid deformations, we define B fine-grained basis transformations $\mathcal{B}_k^t = \{(\mathbf{r}_{k,b}^t, \mathbf{t}_{k,b}^t)\}_{b=1}^B$, where $\mathbf{r}_{k,b}^t$ is a continuous rotation representation (e.g., 6D [58]) and $\mathbf{t}_{k,b}^t \in \mathbb{R}^3$. Each Gaussian $g_i \in \mathcal{C}_k$ is assigned a learnable coefficient vector $\mathbf{w}_i = [w_{i,1}, \dots, w_{i,B}]^\top$, where $\sum_{b=1}^B w_{i,b} = 1$. The local transformation $\mathbf{L}_{i,\ell}^t = [\mathbf{R}_{i,\ell}^t \mid \mathbf{t}_{i,\ell}^t]$ is computed by blending the cluster bases:

$$\mathbf{R}_{i,\ell}^t = \mathcal{R}\left(\sum_{b=1}^B w_{i,b} \mathbf{r}_{k,b}^t\right), \quad \mathbf{t}_{i,\ell}^t = \sum_{b=1}^B w_{i,b} \mathbf{t}_{k,b}^t, \quad (2)$$

where $\mathcal{R}(\cdot)$ denotes the mapping to $\text{SO}(3)$. The final state of Gaussian g_i at time t is the composition of global and local transformations applied to the canonical state $(\mu_i^0, \mathbf{R}_i^0)$:

$$\mu_i^t = \mathbf{R}_{k,g}^t (\mathbf{R}_{i,\ell}^t \mu_i^0 + \mathbf{t}_{i,\ell}^t) + \mathbf{t}_{k,g}^t, \quad \mathbf{R}_i^t = \mathbf{R}_{k,g}^t \mathbf{R}_{i,\ell}^t \mathbf{R}_i^0. \quad (3)$$

Adaptive control. To dynamically adjust representation capacity, we introduce an adaptive control mechanism inspired by the densification strategy of 3D Gaussians [18]. This scheme splits or culls clusters based on their fidelity in representing local dynamics. Since each cluster is designed

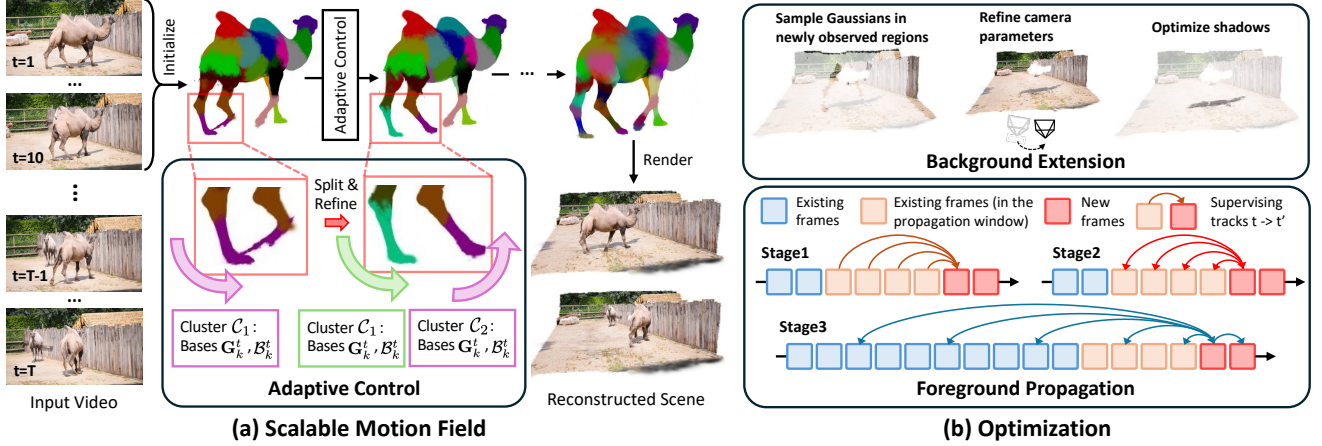


Figure 2. **Overview of MotionScale.** Our method adopts a scalable motion field that progressively captures object motions through an adaptive control mechanism, enabling efficient splitting and refinement of motion components. For optimization, the background is updated through region sampling, camera refinement, and shadow handling, while the foreground propagation employs a three-stage refinement to propagate motion across long temporal windows for consistent 4D reconstruction.

to model a roughly rigid body segment, any significant non-rigid or inconsistent motion within a cluster indicates that the current representation is insufficient. To resolve this, as illustrated in Fig. 2(a), we partition the cluster’s Gaussians $\{g_i \in \mathcal{C}_k\}$ into two groups exhibiting distinct motion patterns (detailed in Sec. 3.3). To ensure optimization stability, we duplicate the original motion parameters for both new clusters. This initialization allows the two clusters to diverge in parameter space while fully preserving the original motion state at the moment of splitting, ultimately enabling the motion field to capture finer, localized variations.

Efficiency and scalability. The cluster-based motion field is designed for both computational efficiency and architectural scalability. Because each Gaussian is influenced by a single cluster, the computational cost remains nearly constant even as the motion field expands. Furthermore, the representation is highly memory-efficient, as each cluster maintains a compact parameter set that scales marginally relative to the total Gaussian count. Beyond efficiency, this architectural flexibility also ensures high expressiveness. As illustrated in Fig. 2(a), the splitting process isolates distinct or non-rigid movements into independent clusters, allowing the motion field to resolve previously ambiguous dynamics. By progressively optimizing these new clusters, our method can capture intricate details and cover all meaningful parts of an object (*cf.* Fig. 2(a), top row). Consequently, this adaptive control enables our framework to scale seamlessly to larger scenes and longer sequences while maintaining both coherent geometry and precise motion tracking.

3.3. Optimization Strategy

Given a video sequence of T images $\{I_t\}_{t=1}^T$ with unknown camera parameters, our goal is to reconstruct a 4D scene

representation that captures high-quality geometry, photorealistic appearance, and accurate motion. To provide a robust initialization for both geometry and dynamics, we leverage pre-trained 2D models to extract a suite of priors, including monocular depth maps $\{\mathbf{D}_t\}_{t=1}^T$, foreground masks $\{\mathbf{M}_t\}_{t=1}^T$, and dense 2D point tracks $\{\mathbf{U}_t\}_{t=1}^T$. Additionally, we estimate initial camera poses using vision-based geometry frameworks [47], which are jointly refined with the scene representation during optimization.

Initialization. We represent the scene as a combination of a static background and a dynamic foreground that are jointly rendered to produce the final image [12, 43]. The optimization begins on an initial temporal window $\{I_t\}_{t=1}^{T_{\text{init}}}$. To initialize the background, we back-project 3D points from the monocular depth maps $\{\mathbf{D}_t\}$ to create an initial 3DGS point cloud. For the dynamic motion field, we sample 3D trajectories from the 2D tracks $\{\mathbf{U}_t\}$ and the corresponding depth maps. We then apply K-means clustering to these 3D points in the canonical frame to define the initial spatial extent of the K clusters. The global transformations $\mathbf{G}_k^t$ are then initialized via Procrustes analysis to estimate the rigid transformations between point clouds across the initial frames, while local refinement bases $\mathbf{B}_k^t$ are set to identity.

Progressive optimization. To handle long video sequences efficiently while maintaining temporal coherence, we adopt a progressive optimization strategy (*cf.* Fig. 2(b)). Once the initial window is optimized, we incrementally incorporate a subset of T_{new} additional frames. This iterative process consists of two primary steps: background extension and foreground adaptation.

Background extension. To incorporate a batch of T_{new} additional frames, we first extend the static background recon-

struction. Since new views often reveal previously occluded regions or areas beyond the initial image boundaries, we identify and systematically populate these unobserved regions to maintain background completeness. Specifically, we project all existing background Gaussians onto the new image planes, marking pixels without coverage as unobserved areas that require new samples. In these regions, we initialize additional Gaussians by sampling 3D points from the monocular depth maps $\{\mathbf{D}_t\}$. We then perform a targeted optimization on the background pixels of the new frames to refine the added geometry and ensure photometric consistency. Simultaneously, we jointly refine the camera extrinsics via an end-to-end gradient-based approach. This lightweight refinement corrects sub-pixel misalignments in the initial poses [47] through direct photometric loss, providing a simple yet effective alternative to explicit SLAM-based tracking or global bundle adjustment.

Foreground propagation. Following the background update, we extend the motion field to the newly introduced frames. The objective is to estimate the global transformations $\mathbf{G}_k^t$ and the local bases $\mathcal{B}_k^t$ that parameterize the scene dynamics at these new timestamps. To ensure temporal continuity, we initialize the motion bases for the T_{new} frames based on the optimized transformations from the most recent frames. The estimation is supervised using 2D tracking priors. For each pixel p in a query frame t , we compute its 3D trajectory at a target time t' by accumulating the 3D positions of the Gaussians via alpha-blending:

$$\mathbf{X}_{t \rightarrow t'}(\mathbf{p}) = \sum_{i \in H(\mathbf{p})} \mu_i^{t'} \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (4)$$

where $H(\mathbf{p})$ denotes the set of Gaussians projected at pixel p , and $\mu_i^{t'}$ is the position of Gaussian g_i at time t' , derived from its associated cluster transformations $\mathcal{T}_k^{t'}$. The 3D position $\mathbf{X}_{t \rightarrow t'}(\mathbf{p})$, representing the estimated world coordinates for pixel p at time t' , is then projected onto the image plane to obtain the estimated 2D track $\hat{\mathbf{U}}_{t \rightarrow t'}(\mathbf{p})$ and depth $\hat{\mathbf{D}}_{t \rightarrow t'}(\mathbf{p})$. The motion field is then optimized by minimizing the tracking and depth consistency losses:

$$L_{\text{track}} = \frac{1}{|I_t|} \sum_{p \in I_t} \|\hat{\mathbf{U}}_{t \rightarrow t'}(\mathbf{p}) - \mathbf{U}_{t \rightarrow t'}(\mathbf{p})\|, \quad (5)$$

$$L_{\text{depth}} = \frac{1}{|I_t|} \sum_{p \in I_t} \|\hat{\mathbf{D}}_{t \rightarrow t'}(\mathbf{p}) - \mathbf{D}_{t \rightarrow t'}(\mathbf{p})\|. \quad (6)$$

where $\mathbf{U}_{t \rightarrow t'}(\mathbf{p})$ is the 2D track prior, denoting the target location in frame t' for a pixel p from frame t , and $\mathbf{D}_{t \rightarrow t'}(\mathbf{p}) = \mathbf{D}_{t'}(\mathbf{U}_{t \rightarrow t'}(\mathbf{p}))$ represents the monocular depth prior evaluated at that tracked location.

To mitigate noise in the 2D tracking priors and ensure temporal consistency, we optimize the motion field through a *three-stage refinement process* (cf. Fig. 2(b)):

1. *Initial Alignment.* We first define a short propagation window (denoted as orange frames in Fig. 2(b)) covering the T_{prop} most recent frames. To avoid polluting the well-optimized previous frames with unaligned new data, we employ a one-directional tracking loss using $\mathbf{U}_{t \rightarrow t'}(\mathbf{p})$, where t belongs to the established frames and t' represents the new frames. In this stage, we optimize only the motion bases for the new frames while freezing all other scene parameters, including Gaussian attributes and existing transformations.
2. *Short-term Consistency.* Once the new frames are roughly aligned, we transition to a *bi-directional tracking loss* computed over arbitrary frame pairs (t, t') within the propagation window. This stage enforces local temporal consistency and allows gradients to flow back into recent frames, refining their states with information from the new views. Stages 1 and 2 are repeated iteratively as new frames are incorporated. This stage remains efficient by limiting the optimization to the motion field and leveraging the sparsity of the 2D tracking priors.
3. *Long-term Refinement.* After a sufficient number of frames have been integrated, we perform long-term motion refinement by sampling frame pairs (t, t') across the entire optimized sequence. This global supervision resolves accumulated drift and occlusions. At this stage, we jointly optimize all parameters and allow for Gaussian densification. This joint optimization objective combines the previously defined tracking and depth losses with a photometric (RGB) loss and as-rigid-as-possible (ARAP) regularization terms.

Adaptive control. During the long-term refinement in Stage 3, we trigger the adaptive control mechanism to dynamically update the cluster topology. We first identify and partition clusters that fail to accurately represent the coherent motion of their associated Gaussians. Specifically, for each cluster, we extract the 3D trajectories of its Gaussians over the propagation window to serve as feature descriptors for HDBSCAN clustering. Within the resulting density-based sub-clusters, we apply Agglomerative Clustering to isolate two candidate groups and compute the distance between their centroids. If this distance exceeds a predefined threshold, the original cluster is identified as motion-inconsistent. As illustrated in Fig. 2(a), this approach effectively isolates Gaussians with divergent dynamics. Following the splitting protocol in Sec. 3.2, new clusters and motion bases are initialized to accommodate these new dynamics. Finally, we prune undersized clusters by eliminating their associated Gaussians and removing the corresponding motion field entries, followed by a remapping of the remaining cluster indices to ensure a compact representation.

Shadow Gaussians. To realistically reconstruct the tran-

sient lighting effects caused by moving objects, we introduce a dedicated set of “shadow Gaussians” within the background representation. Unlike static background primitives, these Gaussians are coupled with a dynamic motion field, allowing them to track with the objects casting the shadows. Recognizing that shadows lack well-defined 3D geometry and that their masks are often temporally inconsistent, we deliberately omit geometric and motion supervision for these primitives. Instead, shadow Gaussians are optimized primarily via the photometric (RGB) loss, complemented by a segmentation constraint to prevent spatial overlap with the foreground object. Initialized from coarse shadow masks in the starting frames, these primitives are refined jointly with the static background during the propagation process. This streamlined strategy effectively captures complex shadow dynamics while maintaining the overall simplicity of the pipeline (*cf.* Sec. 4.5).

4. Experiments

4.1. Experimental Setup

Implementation details. We implement MotionScale in PyTorch and conduct all experiments on a single NVIDIA RTX 4090 GPU. By default, for in-the-wild sequences, we utilize π^3 [47] for monocular depth and camera poses, and SAM2 [36] for foreground masks. 2D point tracks are generated by sampling a dense grid within these masks and tracking them via CoTracker3 [16]. To ensure fair comparison on established benchmarks (e.g., DyCheck), we strictly follow their provided evaluation protocols and utilize the standard input priors. The scene is initialized with canonical 3D Gaussians following [18]. Optimization follows the progressive optimization strategy described in Sec. 3. Further details on hyperparameters and training schedules are provided in the supplementary material.

Datasets. We evaluate our method on three diverse benchmarks: (1) *DAVIS* [31], featuring single-view, casually recorded videos with complex non-rigid motions and natural camera movement; (2) *DyCheck* [10], containing 14 real-world dynamic scenes, including additional synchronized views from two static cameras and metric LiDAR depth. We utilize its sparse keypoint annotations (5-15 per sequence) to evaluate long-term 3D tracking; and (3) *NVIDIA Dynamic Scenes* [55], which provides calibrated multi-view sequences of diverse human and object activities. These datasets allow us to assess MotionScale’s ability to recover consistent 4D representations from both casual monocular recordings and controlled multi-view rigs.

Evaluation metrics. For rendering and NVS, we report standard metrics: PSNR, SSIM, and LPIPS [56]. For 3D tracking, we report the 3D End-Point-Error (EPE) and the percentage of points within distance thresholds $\delta_{3D}^{0.5}$ and $\delta_{3D}^{1.0}$ (5cm and 10cm). For 2D tracking, we follow standard pro-

Table 1. Comparison of novel view synthesis results on DyCheck [10] and NVIDIA [55] datasets.

Method	DyCheck NVS			NVIDIA NVS		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
T-NeRF [10]	15.60	0.55	0.55	20.76	0.59	0.17
HyperNeRF [30]	15.99	0.59	0.51	20.05	0.57	0.18
DynIBaR [22]	13.41	0.48	0.55	-	-	-
Deformable-GS [53]	11.92	0.49	0.66	-	-	-
4D Gaussians [49]	13.11	0.39	0.73	17.69	0.48	0.38
Dynamic GM [39]	15.79	0.59	0.44	22.36	0.66	0.15
4D-Fly [48]	17.03	0.60	0.37	22.52	0.69	0.14
Shape of Motion [43]	16.72	0.63	0.45	23.37	0.75	0.10
MotionScale	17.98	0.70	0.40	26.75	0.78	0.07

ocols [4] and report Average Jaccard (AJ), average position accuracy (δ_{avg}), and Occlusion Accuracy (OA).

4.2. 4D Scene Reconstruction

We evaluate MotionScale on the monocular DAVIS benchmark [31], which presents significant challenges due to its single-view perspective and complex, non-rigid deformations. We compare our approach against state-of-the-art dynamic reconstruction methods, including Shape-of-Motion [43] and GFlow [46]. As illustrated in the top portion of Fig. 3, we render dynamic objects from two novel viewpoints across multiple timestamps to assess geometric and temporal consistency. While GFlow frequently produces “cloud-like” artifacts and Shape-of-Motion exhibits visible geometric drift and temporal flickering, our cluster-based motion field effectively constrain Gaussian trajectories, preserving sharp surfaces and textural details. Furthermore, in large-scale scenes characterized by extensive camera movement and rapid object dynamics (Fig. 3, bottom), competing methods struggle with motion discontinuities during large displacements. Conversely, our progressive optimization strategy ensures a stable global structure and detailed local motion by grounding the motion field in the established background geometry. Overall, these results demonstrate that MotionScale achieves superior 4D fidelity, delivering clearer geometry and smoother motion across diverse real-world scenarios.

4.3. Novel View Synthesis

We evaluate novel view synthesis (NVS) performance on the DyCheck [10] and NVIDIA Dynamic Scenes [55] benchmarks. As established in Sec. 4.1, we strictly follow the standard evaluation protocols and utilize provided input priors to ensure a direct comparison with existing baselines. Quantitative results are summarized in Tab. 1, where MotionScale consistently outperforms all prior methods on both datasets. On the DyCheck benchmark, our method’s substantial gains in PSNR and lower LPIPS scores reflect superior photometric and perceptual fidelity. These

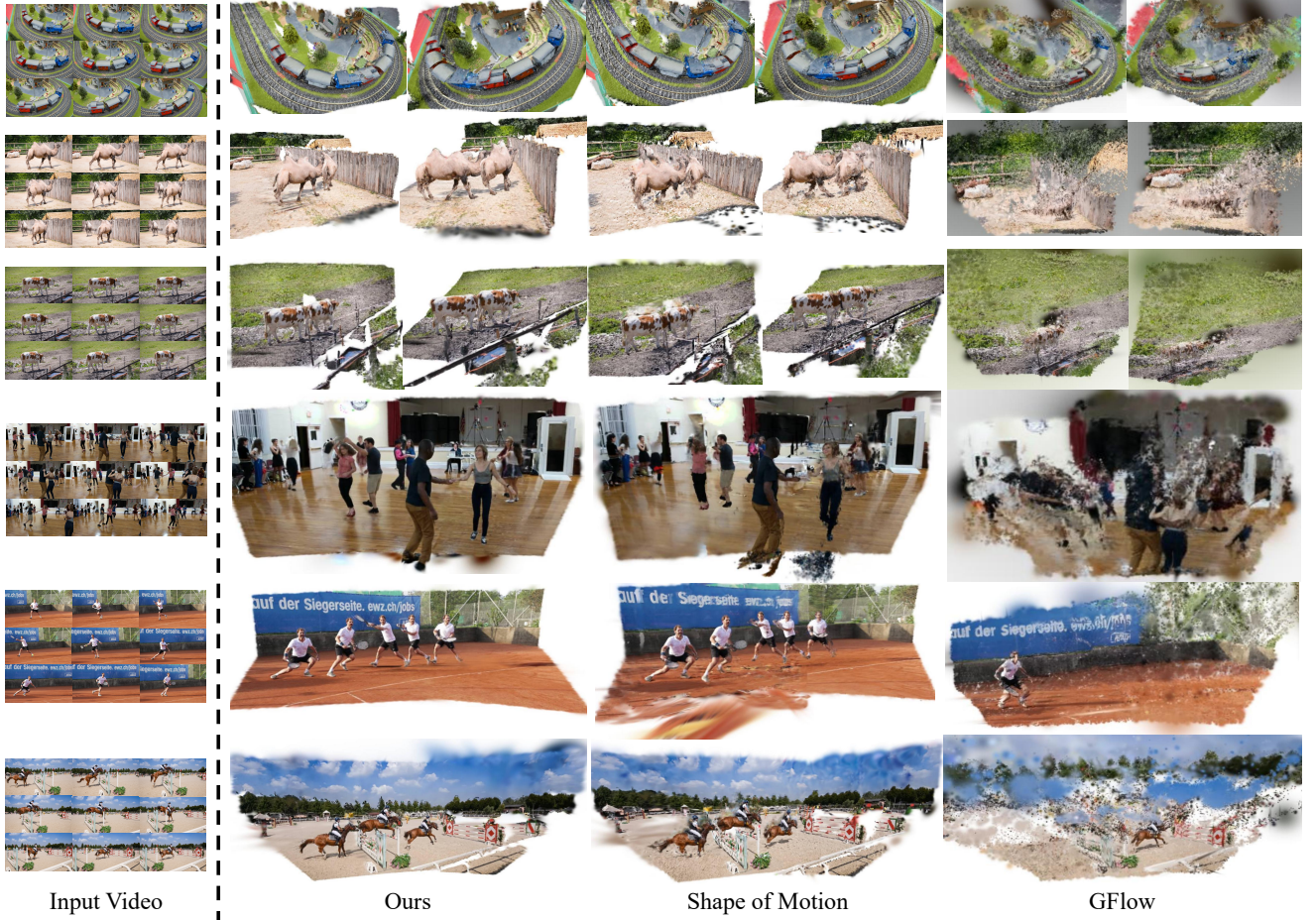


Figure 3. Comparison of dynamic scene reconstruction results on challenging real-world videos from DAVIS dataset. We compare MotionScale with Shape of Motion [43] and GFlow [46] on several dynamic scenes containing complex object motions, occlusions, and large appearance variations. For the top rows, we show rendered results under two different viewpoints for each compared method.

improvements are particularly pronounced in dynamic regions involving large non-rigid motions, where competing approaches frequently produce blur or ghosting artifacts. Similarly, on NVIDIA Dynamic Scenes, MotionScale effectively preserves fine-grained motion details and maintains strict temporal coherence across synthesized views. These results validate that our proposed method successfully resolves complex temporal dynamics, delivering high-quality novel view synthesis in both casual monocular and calibrated multi-view settings.

4.4. 3D Point Tracking

We evaluate the accuracy of our recovered motion field by performing point-based tracking on the DyCheck benchmark [10]. We evaluate 3D trajectories by lifting sparse 2D keypoints using LiDAR depth, alongside standard 2D tracking metrics. Quantitative results in Tab. 2 show that MotionScale significantly outperforms existing dynamic reconstruction baselines across all metrics. For 3D tracking,

Table 2. Comparison of point-based tracking performance on the DyCheck [10] dataset.

Method	3D Point Tracking			2D Point Tracking		
	EPE ↓	$\delta_{3D}^{0.5} \uparrow$	$\delta_{3D}^{10} \uparrow$	AJ ↑	$< \delta_{avg} \uparrow$	OA ↑
HyperNeRF [30]	0.182	28.4	45.8	10.1	19.3	52.0
DynIBaR [22]	0.252	11.4	24.6	5.4	8.7	37.7
Deformable-GS [53]	0.151	33.4	55.3	14.0	20.9	63.9
DynMF [20]	0.188	22.9	53.8	5.5	9.5	60.5
CoTracker [15]+DA [51]	0.202	34.3	57.9	24.1	33.9	73.0
TAPIR [5]+DA [51]	0.114	38.1	63.2	27.8	41.5	67.4
DELTA [27]	0.159	32.5	55.3	24.7	34.1	68.9
SpatialTracker [50]	0.125	37.7	63.9	24.9	36.9	73.5
Shape of Motion [43]	0.082	43.0	73.3	34.4	47.0	86.6
MotionScale	0.070	47.0	76.4	37.7	50.6	87.4

our method achieves the lowest End-Point Error (EPE) and the highest accuracy under both $\delta_{3D}^{0.5}$ and δ_{3D}^{10} thresholds, demonstrating a robust ability to maintain 3D correspon-

Table 3. Ablation studies on the DyCheck dataset.

Method	PSNR $\uparrow$	SSIM $\uparrow$	LPIPS $\downarrow$	AJ $\uparrow$	$\delta_{\text{avg}} \uparrow$	OA $\uparrow$
Ours (Full)	17.98	0.70	0.40	37.7	50.6	87.4
Global Bases	16.70	0.63	0.45	34.2	46.6	86.1
w/o Adaptive Control	17.21	0.67	0.42	34.9	47.0	86.6
w/o Pose Ref.	17.45	0.67	0.41	-	-	-
w/o Shadow	16.26	0.60	0.50	-	-	-
w/o FG Propagation	16.97	0.64	0.42	34.4	46.9	86.4

dences despite rapid non-rigid deformations. In the 2D evaluation, MotionScale attains notable gains in Average Jaccard (AJ) and Occlusion Accuracy (OA). These improvements suggest that our tracking loss and cluster-based motion propagation successfully resolve long-term dependencies and visibility changes, producing more stable and temporally coherent trajectories than methods relying on global or limited motion representations.

4.5. Ablation Studies

We conduct ablation experiments on the DyCheck benchmark [10] to evaluate the contribution of our key framework components. Quantitative results for NVS and 2D tracking are summarized in Tab. 3.

Scalable motion field. To verify our motion representation, we compare against a baseline using a fixed number of global motion bases shared by all Gaussians (*Global Bases*), similar to [43]. Our cluster-based design significantly outperforms this baseline, demonstrating that localizing motion bases to specific Gaussian clusters provides the necessary degrees of freedom to capture fine-grained, non-rigid deformations that global bases tend to over-smooth.

Adaptive Control. We ablate our dynamic cluster management by disabling the splitting and culling mechanisms (*w/o Adaptive Control*). In this variant, cluster assignments remain fixed to the initialization. The performance drop indicates that the ability to topologically adapt the motion field and develop divergent dynamics is crucial for maintaining motion accuracy as the scene complexity evolves.

Pose Refinement. We evaluate the impact of joint pose refinement by fixing camera trajectories to initial π^3 [47] estimates (*w/o Pose Ref.*). As shown qualitatively in Fig. 4 (top), this leads to degraded photometric alignment. The subtle geometric drift causes noticeable blurring in sharp textures (red box), confirming that high-quality monocular priors require spatial refinement during Gaussian optimization to ensure strict consistency.

Shadow Gaussians. We further ablate the role of our dedicated shadow primitives by disabling this subset in the background model (*w/o Shadow*). Quantitatively, this leads to a significant decrease in PSNR for background rendering (Tab. 3). Visually, the absence of these primitives prevents

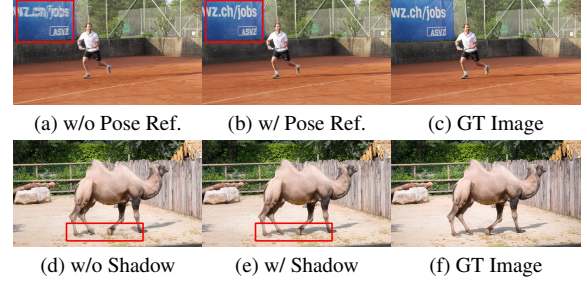


Figure 4. Visual comparison of ablation results.

the model from accurately reconstructing transient lighting on the ground (*cf.* Fig. 4, bottom). More critically, without a dedicated representation, the optimization frequently forces foreground Gaussians to over-extend to capture shadow regions. This leads to unintended geometric dilation and ghosting artifacts on the dynamic objects themselves. Our design effectively decouples foreground geometry from transient shadows, ensuring both clean background textures and high-fidelity object reconstruction.

Foreground Propagation. Finally, we assess our foreground propagation by replacing it with a global optimization approach initialized from full-frame tracks (*w/o FG Propagation*). This variant fails to maintain temporal coherence over long sequences and is prone to optimization instability. This highlights that our three-stage refinement is essential for scaling 4D reconstruction to complex videos.

5. Conclusion

In this paper, we introduced **MotionScale**, a scalable framework designed to recover high-fidelity 4D representations from casual monocular video. While existing methods often achieve impressive view-dependent rendering, they frequently struggle to maintain geometric and temporal consistency over long sequences due to their reliance on short-horizon 2D priors and rigid motion representations. We addressed these challenges by proposing an adaptive, cluster-based motion field that dynamically adjusts its capacity through splitting and culling. Coupled with a progressive optimization strategy, MotionScale effectively bridges the gap between noisy 2D tracks and coherent 3D geometry by incrementally incorporating temporal context. Our comprehensive evaluation across the DAVIS, DyCheck, and NVIDIA benchmarks demonstrates that MotionScale significantly outperforms state-of-the-art approaches in rendering quality, motion accuracy, and geometric stability.

Acknowledgment. This research / project is supported by the National Research Foundation (NRF) Singapore, under its NRF-Investigatorship Programme (Award ID. NRF-NRFI09-0008).

References

- [1] Shariq Farooq Bhat, Reiner Birkel, Diana Wofk, Peter Wonka, and Matthias Müller. Zoedepth: Zero-shot transfer by combining relative and metric depth. *arXiv preprint arXiv:2302.12288*, 2023. 2
- [2] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pages 333–350. Springer, 2022. 2, 3
- [3] Sili Chen, Hengkai Guo, Shengnan Zhu, Feihu Zhang, Zilong Huang, Jiashi Feng, and Bingyi Kang. Video depth anything: Consistent depth estimation for super-long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22831–22840, 2025. 2
- [4] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. TAP-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022. 2, 6
- [5] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10061–10072, 2023. 2, 7
- [6] Carl Doersch, Pauline Luc, Yi Yang, Dilara Gokay, Skanda Koppula, Ankush Gupta, Joseph Heyward, Ignacio Rocco, Ross Goroshin, Joao Carreira, et al. Bootstrap: Bootstrapped training for tracking-any-point. In *Proceedings of the Asian Conference on Computer Vision*, pages 3257–3274, 2024. 2
- [7] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015. 2
- [8] Scott Ettinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R Qi, Yin Zhou, et al. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9710–9719, 2021. 2
- [9] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2, 3
- [10] 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. 6, 7, 8
- [11] Donald M Hilty, Karan Randhawa, Marlene M Maheu, Alastair JS McKean, Richard Pantera, Matthew C Mishkind, and Albert “Skip” Rizzo. A review of telepresence, virtual reality, and augmented reality applied to clinical care. *Journal of Technology in Behavioral Science*, 5(2):178–205, 2020. 2
- [12] Yi-Hua Huang, Yang-Tian Sun, Ziyi Yang, Xiaoyang Lyu, Yan-Pei Cao, and Xiaojuan Qi. Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4220–4230, 2024. 2, 3, 4
- [13] Zhaoyang Huang, Xiaoyu Shi, Chao Zhang, Qiang Wang, Ka Chun Cheung, Hongwei Qin, Jifeng Dai, and Hongsheng Li. Flowformer: A transformer architecture for optical flow. In *European conference on computer vision*, pages 668–685. Springer, 2022. 2
- [14] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017. 2
- [15] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-tracker: It is better to track together. In *European conference on computer vision*, pages 18–35. Springer, 2024. 2, 7
- [16] Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-tracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6013–6022, 2025. 2, 6
- [17] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauß, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 2
- [18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 2, 3, 6
- [19] 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
- [20] Agelos Kratimenos, Jiahui Lei, and Kostas Daniilidis. Dynmf: Neural motion factorization for real-time dynamic view synthesis with 3d gaussian splatting. In *European Conference on Computer Vision*, pages 252–269. Springer, 2024. 7
- [21] 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. 3
- [22] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4273–4284, 2023. 6, 7
- [23] Jiahao Lu, Tianyu Huang, Peng Li, Zhiyang Dou, Cheng Lin, Zhiming Cui, Zhen Dong, Sai-Kit Yeung, Wenping Wang, and Yuan Liu. Align3r: Aligned monocular depth estimation

- for dynamic videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 22820–22830, 2025. 2
- [24] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 800–809. IEEE, 2024. 2, 3
- [25] 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, 3
- [26] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4):1–15, 2022. 3
- [27] Tuan Duc Ngo, Peiye Zhuang, Chuang Gan, Evangelos Kalogerakis, Sergey Tulyakov, Hsin-Ying Lee, and Chaoyang Wang. Delta: Dense efficient long-range 3d tracking for any video. *arXiv preprint arXiv:2410.24211*, 2024. 7
- [28] Jongmin Park, Minh-Quan Viet Bui, Juan Luis Gonzalez Bello, Jaeho Moon, Jihyong Oh, and Munchurl Kim. Splines: Robust motion-adaptive spline for real-time dynamic 3d gaussians from monocular video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26866–26875, 2025. 2, 3
- [29] 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, 3
- [30] 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, 3, 6, 7
- [31] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 6
- [32] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10106–10116, 2024. 2
- [33] Luigi Piccinelli, Christos Sakaridis, Yung-Hsu Yang, Mattia Segu, Siyuan Li, Wim Abbeloos, and Luc Van Gool. Unidepthv2: Universal monocular metric depth estimation made simpler. *arXiv preprint arXiv:2502.20110*, 2025. 2
- [34] 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, 3
- [35] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020. 2
- [36] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 2, 6
- [37] Alexander Schäfer, Gerd Reis, and Didier Stricker. A survey on synchronous augmented, virtual, and mixed reality remote collaboration systems. *ACM Computing Surveys*, 55(6):1–27, 2022. 2
- [38] Jianxin Shi, Jinhao Chen, Yuandong Wang, Li Sun, Chunyang Liu, Wei Xiong, and Tianyu Wo. Motion forecasting for autonomous vehicles: a survey. *arXiv preprint arXiv:2502.08664*, 2025. 2
- [39] Colton Stearns, Adam Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 6
- [40] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. Pwc-net: Cnns for optical flow using pyramid, warping, and cost volume. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8934–8943, 2018. 2
- [41] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European conference on computer vision*, pages 402–419. Springer, 2020. 2
- [42] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025. 1, 2
- [43] Qianqian Wang, Vickie Ye, Hang Gao, Weijia Zeng, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9660–9672, 2025. 2, 3, 4, 6, 7, 8
- [44] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5261–5271, 2025. 2
- [45] 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. 1
- [46] Shizun Wang, Xingyi Yang, Qihong Shen, Zhenxiang Jiang, and Xinchao Wang. Gflow: Recovering 4d world from monocular video. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7862–7870, 2025. 2, 3, 6, 7

- [47] Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. π^3 : Scalable permutation-equivariant visual geometry learning, 2025. [2](#), [4](#), [5](#), [6](#), [8](#)
- [48] Diankun Wu, Fangfu Liu, Yi-Hsin Hung, Yue Qian, Xiaohang Zhan, and Yueqi Duan. 4d-fly: Fast 4d reconstruction from a single monocular video. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16663–16673, 2025. [2](#), [3](#), [6](#)
- [49] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20310–20320, 2024. [2](#), [3](#), [6](#)
- [50] 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*, pages 20406–20417, 2024. [2](#), [7](#)
- [51] 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](#), [7](#)
- [52] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. [2](#)
- [53] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20331–20341, 2024. [6](#), [7](#)
- [54] Zeyu Yang, Hongye Yang, Zijie Pan, and Li Zhang. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In *International Conference on Learning Representations (ICLR)*, 2024. [2](#), [3](#)
- [55] 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. [6](#)
- [56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. [6](#)
- [57] Yang Zheng, Adam W Harley, Bokui Shen, Gordon Wetstein, and Leonidas J Guibas. Pointodyssey: A large-scale synthetic dataset for long-term point tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19855–19865, 2023. [2](#)
- [58] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. [3](#)