KineticGS: Momentum-driven Coherent 4D Gaussian Splatting for Monocular Dynamic Scene Reconstruction

Yunqing Wang[1], Baoyao Yang[⋆1], Si-Qi Liu[⋆2,3], Chong Yin[4], Yihua Shao[5], and Hao Tang[6,7]

[1] Guangdong University of Technology, China
[2] National Health Data Institute, Shenzhen, China
[3] Shenzhen Research Institute of Big Data, China
[4] Hainan University, China
[5] Institute of Automation Chinese Academy of Sciences, China
[6] School of Computer Science, Peking University, China
[7] Beijing Academy of Artificial Intelligence, China

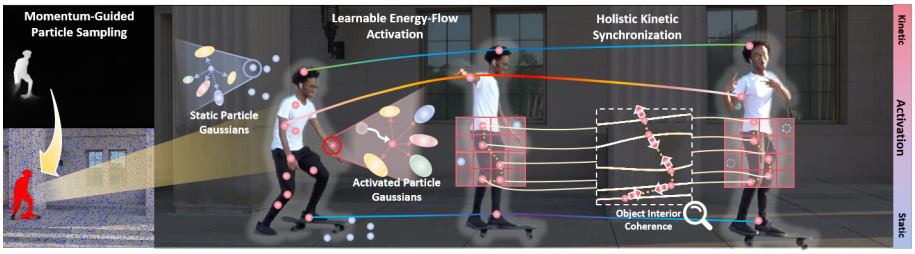


Fig. 1: KineticGS exploits a hierarchical momentum strategy (Initialization, Sustenance, Synchronization) to maintain energy coherence in 4DGS, preventing dynamic object fragmentation and ensuring robust reconstruction from monocular videos.

Abstract. We present KineticGS, a novel approach for reconstructing dynamic scenes from monocular videos through momentum-driven 4D Gaussian Splatting. While recent advances in motion modeling have improved dynamic reconstruction, they often produce temporal inconsistencies and non-physical deformations due to the lack of physical priors constraining motion trajectories. KineticGS addresses this with a physically grounded momentum hierarchy that models Gaussian particles as momentum-carrying entities, thereby establishing a hierarchical motion representation from pixel observations to object-level coherence. Specifically, Gaussian particles serve as momentum carriers, with dynamics driven by local motion energy. We perform momentum-guided particle sampling, concentrating representation in high-energy regions to capture fine-grained non-rigid motions. A learnable energy-flow module then predicts per-particle temporal activation, ensuring continuous and physically plausible motion propagation. Further, a holistic kinetic

⋆ Corresponding author

synchronization mechanism enforces consistent motion evolution among correlated particles, preserving structural integrity without explicit disintegration. Experiments on dynamic scene datasets show that KineticGS improves reconstruction fidelity, temporal coherence, and physical realism, achieving +0.91 dB average PSNR gain in dynamic-object regions over the second-best.

1 Introduction

Dynamic scene reconstruction has become a fundamental problem in computer vision, aiming to recover both the 3D geometry and temporal evolution of real-world environments from video observations. The ability to model dynamic scenes with high spatial and temporal fidelity unlocks numerous applications, including immersive AR/VR experiences, autonomous robotic perception, and realistic digital avatars [1, 11, 28, 30, 32]. Among various input settings, monocular video reconstruction is particularly appealing due to its low hardware cost, flexible deployment, and compatibility with common imaging devices such as mobile phones. Accurately reconstructing dynamic scenes from a single-view video remains highly challenging, as it requires disentangling complex motions, occlusions, and non-rigid deformations from limited observations [9, 35].

Early breakthroughs in Neural Radiance Fields (NeRF) [24] established the foundation for high-fidelity novel view synthesis through volumetric neural representations. However, extending NeRF to dynamic or monocular settings remains challenging. Its implicit volumetric representation relies heavily on dense multi-view supervision and extensive sampling, and its slow rendering hinders practical deployment in real-world applications [6, 7, 25–27]. The recent emergence of 3D Gaussian Splatting (GS) [14] has introduced a more explicit and efficient paradigm for dynamic scene reconstruction. By representing scenes through spatially adaptive Gaussian primitives optimized over position, scale, and color, 3DGS enables faster differentiable rendering and more interpretable geometric structures, naturally supporting 4D extensions [12, 41, 48]. Yet, current 4DGS variants predominantly rely on appearance-based supervision without incorporating physical priors, often resulting in fragmented trajectories, unstable motion estimation, and diminished temporal coherence in monocular settings characterized by complex dynamics or frequent occlusions.

To overcome these limitations, several methods incorporate pre-computed 2D cues such as depth or optical flow [21, 38, 44, 52] to provide coarse geometric and motion guidance. However, these cues are typically noisy and incomplete in single-view videos, capturing only approximate structures that rapidly deteriorate under occlusions, motion blur, or pestimation inaccuracies [31]. This leads to error accumulation across frames, manifesting as temporal flickering, motion

** Code will be available at: `https://github.com/wyunqing/KineticGS.git`

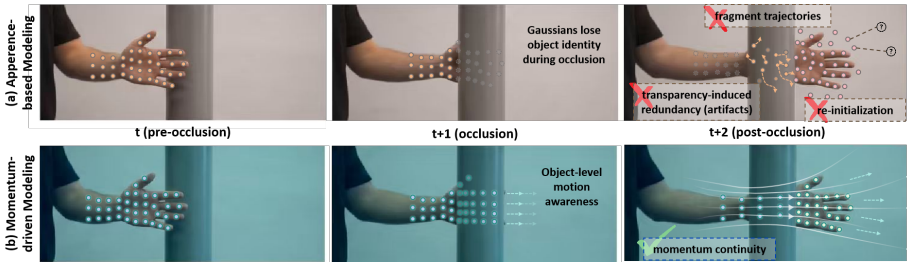


Fig. 2: Comparison in Gaussian Modeling. (a) Existing methods typically rely on appearance-based Gaussians, which leads to disconnected motion modeling before and after occlusion and produces fragmented trajectories. (b) In contrast, our method preserves a momentum-carrying Gaussian: its velocity is guided by visible regions, and an activation curve governs its temporal updates, ensuring continuous motion.

drift, and geometric distortion in non-rigid regions. Recent efforts, such as PhysGaussian [43], have attempted to embed Newtonian dynamics into Gaussian kernels to enable physically consistent simulation and rendering. Nevertheless, these simulation-oriented formulations require precise knowledge of material properties and external forces, parameters that are typically unavailable or unidentifiable in unconstrained monocular scenarios. Thus, existing methods remain kinematically blind, modeling how objects *appear* in different frames rather than how they *move*. The absence of sufficient explicit motion modeling leads to fragmented representations and temporal incoherence under occlusion: as shown in Fig. 2 (a), the reappearing arm exhibits fragment trajectories and abrupt transparency changes, resulting in physically implausible dynamics. This observation motivates the search for an object-aware modeling framework that explicitly preserves momentum continuity, allowing motion to propagate coherently even in complex scenes with large and rapid motions, as shown in Fig. 2 (b).

To achieve comprehensive motion awareness, this paper proposes KineticGS, a hierarchical Gaussian Splatting method driven by momentum. Inspired by physical systems in which momentum governs motion evolution, we model 3D points as momentum-carrying particles, Gaussian primitives that serve as dynamic anchors to propagate and synchronize kinetic information across time. This formulation establishes a hierarchical representation spanning particle and object levels, enabling unified modeling of dynamic scenes with fine-grained motion awareness and structural coherence, consistent with recent evidence supporting hierarchical inductive biases in dynamic reconstruction [3,17,20]. At the core of KineticGS are three tightly integrated components that collectively ensure physically grounded 4D reconstruction. KineticGS first performs momentum-guided particle sampling, which adaptively allocates particles to regions with high motion energy, granting dynamically active zones denser and more expressive representations. Each particle then undergoes learnable energy-flow activation, modulated by a momentum continuity field that captures kinetic energy propagation over time, thereby promoting smooth and plausible motion trajec-

tories. Finally, holistic kinetic synchronization regularizes the temporal behavior of particles within the same motion entities, which we regard as objects in this paper, enforcing coherent motion patterns that preserve structural integrity while mitigating drift and fragmentation. By coupling momentum-guided particle sampling, learnable energy-flow activation, and object-aware kinetic synchronization, KineticGS achieves physically consistent 4D reconstruction, effectively bridging photometric accuracy with physical motion realism. Our contributions are summarized as follows:

- We propose KineticGS, a momentum-driven Gaussian Splatting method that embeds kinetic principles into 4D scene modeling for temporally continuous and physically coherent 4D scene reconstruction from monocular videos.
- We design a hierarchical paradigm that achieves coherent 4D reconstruction through threefold momentum intervention: adaptively allocating motion-aware particles, ensuring continuous energy propagation, and synchronizing object-level dynamics without explicit segmentation.
- Extensive experiments on monocular dynamic scene datasets demonstrate that KineticGS achieves superior performance, delivering a +0.91 dB PSNR improvement over the second-best method in dynamic-object regions while maintaining competitive storage efficiency.

2 Related Works

2.1 Dynamic Scene Reconstruction

Dynamic scene reconstruction aims to recover the evolving geometry, appearance, and motion of real-world scenes. Early methods primarily rely on multi-view or RGB-D inputs [13, 25, 26, 41, 48], where synchronized camera views or depth signals provide strong spatial cues for resolving geometric ambiguities. These approaches typically establish a canonical scene representation and capture temporal dynamics through explicit warping fields or neural deformation networks [25–27, 31]. When multi-view or depth data are sufficiently available, they achieve high-quality reconstructions, offering a modern alternative to traditional multi-view stereo pipelines [37, 50]. A paradigm shift was brought by 3D Gaussian Splatting (3DGS) [14], which represents scenes using explicit point-based primitives to enable real-time differentiable rendering. Building upon this foundation, dynamic extensions [19, 41, 45, 48] adapt Gaussian primitives to model time-varying motion and appearance. Recent advancements further improve efficiency and compactness through compressed motion representations [5, 22], while sparse-anchor-based formulations [3, 15, 17, 20] enhance structural coherence by explicitly regularizing spatial configurations. Despite these developments, most methods still lack physically grounded motion modeling. Though PhysGaussian [43] incorporates Newtonian dynamics, it relies on predefined material or force parameters, which are seldom available in real-world videos. This motivates our investigation into physics-aware constraints for motion modeling, aiming to achieve stable and robust scene reconstruction.

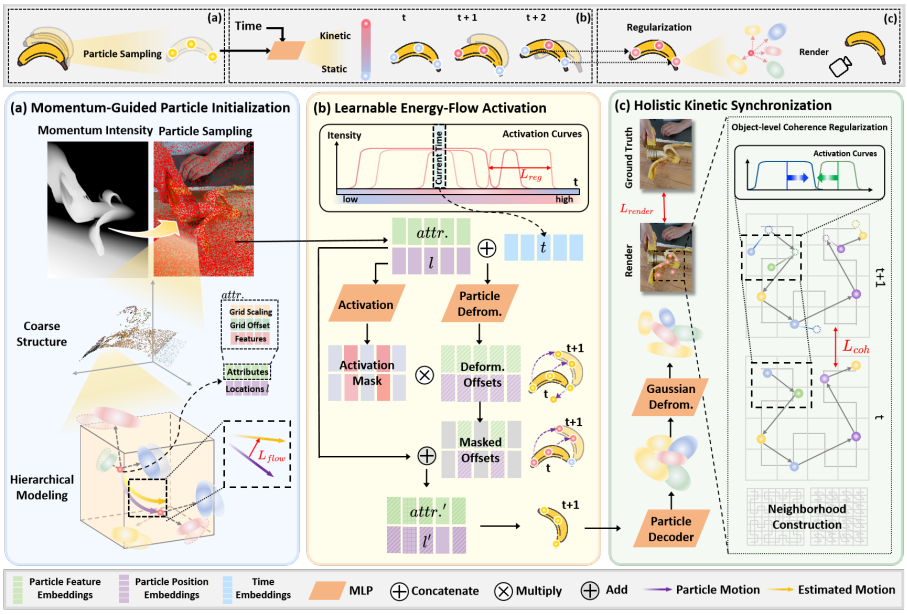


Fig. 3: Overview. KineticGS reconstructs dynamic scenes from monocular videos using a hierarchical representation of Gaussians and particles. Particle activations are modulated by energy-aware masking, and coherence regularization enforces consistent motion trajectories and aligned activation times, producing temporally stable and physically plausible reconstructions.

2.2 Monocular 4D Reconstruction

Among dynamic scene reconstruction tasks, the monocular setting presents a particularly ill-posed inverse problem. It aims to infer 3D structure and temporal evolution from a single-view video, where the absence of explicit depth information leads to inherent ambiguity [2, 34]. Neural implicit representations, particularly dynamic NeRF methods, have advanced monocular reconstruction by jointly modeling geometry, appearance, and motion [8, 27, 31, 49], but often demand dense sampling and entail high computational costs. Recent explicit representations, such as 3D Gaussian Splatting [14] and its dynamic extensions [41], offer more efficient alternatives by directly parameterizing scene content through point-based primitives. Other monocular methods leverage 2D cues, including optical flow, 2D keypoints, and learned image features [18,20,27,39,44,52], which are used to mitigate depth ambiguity and guide motion estimation. However, these cues are often incomplete or inaccurate under occlusions, motion blur, or textureless regions, which compromises their reliability for consistent 3D reconstruction [31,36]. This implies that achieving physically plausible motion in such monocular scenarios necessitates modeling momentum continuity, which enables coherent propagation even through occluded areas.

3 Preliminary: Gaussian Splatting

3DGS [14] is initially introduced for static scene reconstruction, representing a scene with a collection of Gaussian primitives G_i, each defined by its center, scale, opacity, rotation, and color. During rendering, each Gaussian is projected onto the image plane and composited in depth order via an efficient tile-based rasterization strategy, achieving real-time performance. To enhance memory efficiency, a variant, namely Scaffold-GS [23], proposes an anchor-based representation that organizes the scene into a hierarchical structure. It employs a sparse set of anchors, typically sampled from the point cloud or voxel centers, each storing locations, along with learnable features and the corresponding grid offsets and scales that facilitate voxel-based densification during training. Each anchor $\mathbf{a}_i$ then generates K neural Gaussians via a lightweight MLP through its feature $\mathbf{h}_i$, denoted as $G_{i,k} = \mathcal{N}(\mathbf{a}'_{i,k}, \Sigma_{i,k})$, with their centers defined as learnable offsets from the anchor:

$$\mathbf{a}'_{i,k} = \mathbf{a}_i + \Delta\mathbf{a}_{i,k}, \quad k = 0, \ldots, K-1. \tag{1}$$

where $\Delta\mathbf{a}_{i,k} = MLP(\mathbf{h}_i; k)$ denotes the k-th offset derived from MLP. This formulation decouples spatial storage from Gaussian attributes, enabling compact scene modeling with preserved fine-grained geometry and appearance.

For dynamic scenes, a deformation field is introduced to evolve canonical Gaussians over time while maintaining the hierarchical structure between anchors and their associated Gaussians. Let a deformation network $\mathcal{F}$ predict per-Gaussian transformation offsets; the Gaussian parameters at time t are then updated as:

$$\tilde{G}^t_{i,k} = G^0_{i,k} + \mathcal{F}(G^0_{i,k}, t), \quad k = 0, \ldots, K-1, \tag{2}$$

where $G^0_{i,k}$ denotes the Gaussian corresponding to the initial anchor $\mathbf{a}'_{i,k}$ defined in Eq. (1), and $\tilde{G}^t_{i,k}$ represents its deformed counterpart at time t. This continuous deformation of anchor-associated Gaussians ensures spatiotemporal coherence across frames. Notably, the same deformation principle (Eq. (1) and (2)) can also be extended to particle deformation, allowing both its location and other attributes to evolve according to a learned motion field.

4 Method

KineticGS reconstructs dynamic scenes from monocular videos through a physically grounded hierarchy of Gaussians, particles, and dynamic objects, as illustrated in Fig. 3. We introduce momentum-guided particles as motion-aware primitives that bridge Gaussian splats across time and capture the flow of kinetic information (Sec. 4.1). These particles adaptively distribute according to regional motion energy and evolve under energy-flow activation, enabling continuous and physically plausible dynamics (Sec. 4.2). At the object level, coherent motion among correlated particles is synchronized to preserve structural consistency and suppress temporal drift (Sec. 4.3). Through this hierarchical momentum-driven

formulation, KineticGS achieves physically consistent and temporally coherent 4D reconstruction from monocular videos. All components are jointly optimized under a unified objective, with a warm-up stage gradually activating momentum-related losses (Sec. 4.4).

4.1 Momentum-Guided Particle Initialization

Inspired by classical mechanics, where momentum is defined as $p = m\mathbf{v}$ with m denoting mass and $\mathbf{v}$ representing velocity, we reinterpret monocular 4D reconstruction as the process of estimating and propagating momentum-carrying particles in dynamic scenes. In this formulation, the scene is represented as a field of particles whose motion is governed by their intrinsic momentum, allowing dynamic evolution to be modeled in a physically grounded manner.

Since mass and velocity are not directly observable, we approximate them from visual cues. We interpret mass as the resilience of a region to dynamic reconstruction updates, which we derive from geometric confidence based on depth: regions with reliable depth estimates exhibit stronger spatial consistency and are less prone to abrupt optimization changes. Depth thereby serves as a mass-weighting factor, assigning higher inertia to geometrically stable surfaces. Velocity $\mathbf{v}$ is estimated from residual optical flow, which captures non-rigid motion beyond geometry-induced rigid flow. Formally, given the observed optical flow $\mathbf{f}(u, v)$ and the rigid flow $\mathbf{f}_{\text{geom}}(u, v)$ computed from estimated geometry, the momentum field is

$$\mathbf{p}(u, v) = m(u, v) \otimes \big(\mathbf{f}(u, v) - \mathbf{f}_{\text{geom}}(u, v)\big), \tag{3}$$

where $\otimes$ denotes element-wise multiplication. The residual $\mathbf{f}(u, v) - \mathbf{f}_{\text{geom}}(u, v)$ approximates the local non-rigid velocity, while the mass term $m(u, v)$ is estimated from depth-based geometric confidence that provides a measure of geometric reliability for each pixel (details in Sec. 5.1). Their combination yields an implicit momentum distribution over the image domain, where momentum naturally concentrates in regions supported by both strong geometry and strong dynamic evidence.

Particle Sampling. Based on this momentum distribution, we sample dynamic particles to focus capacity on regions with significant non-rigid motion. For each pixel (u, v), we compute the energy density as the sampling probabilities:

$$E(u, v) = \frac{|\mathbf{p}(u, v)|^2}{\sum_{u,v} |\mathbf{p}(u, v)|^2}, \tag{4}$$

where higher momentum yields denser sampling. Pixels drawn from this distribution are back-projected to 3D space using the canonical depth map $\mathbf{d}$ (typically the first frame, to capture motion from the start). For each particle $\mathbf{p}_i$, its 3D location is denoted as

$$\boldsymbol{l}_i = \pi^{-1}\big((u_i, v_i), \mathbf{d}(u_i, v_i)\big). \tag{5}$$

where $\pi^{-1}(\cdot)$ denotes the inverse camera projection. (u_i, v_i) is the pixel coordinate of $\mathbf{p}_i$ and $\mathbf{d}(u_i, v_i)$ is the projected depth.

Hierarchical Modeling. While the above initialization enables dynamic modeling through motion-aware anchors, implicit motion learning alone remains susceptible to drift and temporal inconsistency. We therefore introduce explicit physical guidance: during early training, we align each particle's motion with residual optical flow via a flow–depth consistency loss:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{i,t}\big[\ \big\|\boldsymbol{l}_i^{t\to t+1} - \hat{\boldsymbol{l}}_i^{t+1}\big\|_2\ \big] \tag{6}$$

where $\hat{\boldsymbol{l}}_i^{t+1}$ is the predicted particle location at time $t+1$, and $\boldsymbol{l}_i^{t\to t+1}$ is the target propagated from flow and depth. Each particle thus acts as an adaptive motion carrier, with positions and attributes evolving consistently with local momentum, forming a physically grounded initialization for subsequent optimization.

Leveraging this physics-inspired formulation, KineticGS allocates representational capacity according to the scene's intrinsic kinetic energy rather than heuristic priors. Each momentum-carrying particle $\mathbf{p}_i$—with location $\boldsymbol{l}_i$ and physical attributes including features $\mathbf{h}_i$, grid offsets $\mathbf{o}_i$, and scales $\mathbf{s}_i$—serves as a dynamic counterpart of the static anchors introduced in Sec. 3, evolving them into motion-aware entities that structurally organize surrounding Gaussians while preserving spatiotemporal coherence. This establishes a physically grounded scaffold for subsequent motion modeling.

4.2 Learnable Energy-Flow Activation

Guided by the principle of momentum persistence, the motion state of a momentum-carrying particle should not change abruptly across frames. Instead, its momentum should evolve smoothly over time, even under temporary occlusion. This observation motivates a learnable mechanism that governs the contribution of each particle's stored momentum to scene evolution. We introduce a Learnable Energy-Flow Activation module, which links a particle's temporal visibility to its intrinsic kinetic energy. Each particle maintains a latent momentum vector whose magnitude reflects its local kinetic energy, and its temporal activation is governed by a smoothly varying conservation function. Formally, we define a continuous activation window for each particle $\mathbf{p}_i^t$ at time t as:

$$\mathbf{w}_i^t = \sigma\Big(\frac{t - (t_i^{ct} - t_i^{cr})}{\tau}\Big) \cdot \sigma\Big(\frac{(t_i^{ct} + t_i^{cr}) - t}{\tau}\Big), \tag{7}$$

where $\tau > 0$ controls transition smoothness, and $[t_i^{ct} - t_i^{cr}, t_i^{ct} + t_i^{cr}]$ defines the learnable temporal activation interval. The center time $t_i^{\text{ct}} = \frac{T}{2} + \frac{T}{2} \cdot \Delta t_i^{\text{ct}}$ and temporal span $t_i^{\text{cr}} = \frac{T}{4} + (\frac{T}{4} - \epsilon) \cdot \Delta t_i^{\text{cr}}$ are parameterized via a lightweight MLP conditioned on the particle's spatial coordinates $\boldsymbol{l}_i$:

$$[\Delta t_i^{ct}, \Delta t_i^{cr}] = \tanh(\text{MLP}(\boldsymbol{l}_i)), \tag{8}$$

where T is the video length and ϵ prevents degenerate spans.

The double-sigmoid formulation in Eq. (7) ensures that particle activation naturally aligns with local kinetic states, enabling coherent evolution of dynamic

regions while avoiding abrupt frame-wise gating. Through learnable temporal offsets, particles can adaptively represent motion patterns ranging from brief bursts to sustained activities, effectively capturing the natural variation in dynamic scenes. A distributional regularization is designed and applied to the learned temporal offsets, which promotes balanced and consistent particle activations over time. Specifically, we construct two empirical histograms across all particles: $\mathbf{b}^{\mathrm{ct}}$ that captures the distribution of activation centers over time, and $\mathbf{b}^{\mathrm{cr}}$ that reflects variations in activation spans. Both are regularized toward uniform reference distributions $\mathbf{r}^{\mathrm{ct}}$ and $\mathbf{r}^{\mathrm{cr}}$ via Kullback–Leibler (KL) divergence [16]:

$$\mathcal{L}_{\mathrm{reg}} = D_{\mathrm{KL}}\big(\mathbf{b}^{\mathrm{ct}} \parallel \mathbf{r}^{\mathrm{ct}}\big) + D_{\mathrm{KL}}\big(\mathbf{b}^{\mathrm{cr}} \parallel \mathbf{r}^{\mathrm{cr}}\big), \tag{9}$$

During rendering, the activation window $\mathbf{w}_i^t$ regulates the release of particle momentum over time. The activation window determines whether this momentum contributes to deformation at time t. The deformed particle state is updated as:

$$\mathbf{p}_i^{t\prime} = \mathbf{p}_i^t + \mathcal{H}(\mathbf{w}_i^t) \cdot \Delta\mathbf{p}_i^t, \tag{10}$$

where $\Delta\mathbf{p}_i^t$ denotes the learned momentum-induced deformation offset, and $\mathcal{H}(\mathbf{w}_i^t)$ acts as a temporal momentum gate that modulates whether the particle's inertial state influences its update. For location $\boldsymbol{l}_i^t$, $\mathcal{H}(\mathbf{w}_i^t)$ acts as a binary gate:

$$\mathcal{H}(\mathbf{w}_i^t) = \begin{cases} 1 & \text{if } \ \mathbf{w}_i^t > 0 \\ 0 & \text{otherwise} \end{cases} \tag{11}$$

ensuring that positional displacement occurs only when sufficient dynamic evidence is present, consistent with momentum being activated within its temporal support. For other attributes—including features $\mathbf{h}_i^t$, grid offsets $\mathbf{o}_i^t$, and grid scaling $\mathbf{s}_i^t$—it becomes $\mathcal{H}(\mathbf{w}_i^t) = \mathbf{w}_i^t$, enabling soft, continuous modulation of visual properties. This formulation confines particle updates to their active energy windows, maintaining temporal and momentum-aware coherence while preventing non-physical artifacts.

4.3 Holistic Kinetic Synchronization

Beyond temporal persistence, physically plausible motion demands spatial coherence: particles belonging to the same object should exhibit consistent dynamics. We therefore introduce a Holistic Kinetic Synchronization mechanism that enforces momentum coherence among neighboring particles. Unlike purely positional smoothness, our formulation constrains momentum alignment, reducing velocity divergence across particles and stabilizing mass-weighted motion propagation. As the highest level of our momentum-aware hierarchy, this object-scale constraint complements our momentum-guided initialization and learnable activation modules, collectively ensuring temporally stable and physically grounded 4D reconstruction.

Neighborhood Construction. Establishing meaningful particle neighborhoods is essential for kinetic synchronization, yet particularly challenging in monocular settings where geometrically proximate particles may belong to distinct motion fields. To mitigate spatial mixing in KNN that often connects geometrically close but semantically distinct regions, we serialize particles along a space-filling curve such as Hilbert or Z-order to ensure locally consistent neighborhood organization [42]. These curves preserve spatial locality while providing continuous 1D orderings of 3D positions, enabling neighborhoods that better respect object boundaries and motion coherence. During training, we dynamically project particles onto both Hilbert and Z-orderings, randomly alternating between them across iterations. This stochastic serialization exposes particles to diverse yet structurally consistent neighborhoods, allowing momentum constraints to propagate robustly through space and time. As a result, particles within the same object maintain coherent motion evolution under challenging conditions such as partial occlusion or rapid movement, while effectively preventing spurious inter-object interactions.

Object-Level Coherence. To enforce coherent motion at the object level, we constrain coherence over particle neighborhoods:

$$\mathcal{L}_{\text{coh}} = \sum_{(i,j)\in\text{neighbors}} \|\Delta\boldsymbol{l}_i - \Delta\boldsymbol{l}_j\|^2 + \|t_i^{ct} - t_j^{ct}\|^2, \tag{12}$$

where $\Delta\boldsymbol{l}_i$ and t_i^{ct} denote the location offset and activation center time of particle $\mathbf{p}_i$, respectively. This formulation promotes local momentum continuity by jointly constraining spatial energy distribution and temporal continuity, providing a discrete approximation of momentum conservation that suppresses drift and maintains coherent object dynamics under monocular viewing ambiguity.

4.4 Overall Objective

The full training objective combines conventional rendering supervision with our physically grounded constraints:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{render}} + \lambda_{\text{flow}}\mathcal{L}_{\text{flow}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} + \lambda_{\text{coh}}\mathcal{L}_{\text{coh}}, \tag{13}$$

where $\mathcal{L}_{\text{render}}$ enforces photometric consistency, with formulation following prior work [41]. $\mathcal{L}_{\text{flow}}$ aligns particle motion with residual optical flow (Sec. 4.1). $\mathcal{L}_{\text{reg}}$ and $\mathcal{L}_{\text{coh}}$ ensure smooth temporal activation (Sec. 4.2) and object-level kinetic consistency (Sec. 4.3), respectively. λ_{flow}, λ_{reg}, and λ_{coh} are hyperparameters that balance different losses.

The combination of these terms forms a comprehensive momentum-based objective that stabilizes 4D reconstruction while preserving physically plausible and coherent motion propagation within each object.

Table 1: Quantitative comparison on dynamic object regions of the Dycheck-iPhone dataset [9]. Each group reports (mPSNR(dB)↑, mSSIM↑, mLPIPS↓). Best results in Red and second-best in Yellow.

Method	Apple			Block			Paper-windmill			Space-out		
	mPSNR↑	mSSIM↑	mLPIPS↓	mPSNR↑	mSSIM↑	mLPIPS↓	mPSNR↑	mSSIM↑	mLPIPS↓	mPSNR↑	mSSIM↑	mLPIPS↓
Deform. 3DGS [48]	19.46	0.924	0.095	15.65	0.825	0.169	18.73	0.849	0.133	18.76	0.857	0.115
SC-GS [12]	20.03	0.923	0.089	16.32	0.832	0.156	19.89	0.850	0.109	18.72	0.856	0.116
4DGaussians [41]	21.98	0.928	0.067	18.29	0.830	0.140	20.03	0.857	0.102	17.17	0.852	0.123
MoDec-GS [17]	23.96	0.941	0.071	19.62	0.836	0.143	20.58	0.856	0.104	17.64	0.856	0.121
KineticGS (Ours)	24.98	0.947	0.068	20.55	0.843	0.137	22.06	0.872	0.101	18.63	0.860	0.118

Method	Spin			Teddy			Wheel			Average		
	mPSNR↑	mSSIM↑	mLPIPS↓	mPSNR↑	mSSIM↑	mLPIPS↓	mPSNR↑	mSSIM↑	mLPIPS↓	mPSNR↑	mSSIM↑	mLPIPS↓
Deform. 3DGS [48]	15.35	0.813	0.172	17.09	0.832	0.159	15.51	0.667	0.217	17.22	0.807	0.151
SC-GS [12]	18.01	0.797	0.156	18.36	0.832	0.154	15.40	0.679	0.237	18.10	0.824	0.145
4DGaussians [41]	17.90	0.798	0.145	18.67	0.831	0.146	14.96	0.684	0.238	18.43	0.826	0.137
MoDec-GS [17]	17.61	0.796	0.149	19.72	0.840	0.153	15.33	0.698	0.235	19.21	0.832	0.139
KineticGS (Ours)	18.14	0.801	0.148	20.53	0.853	0.152	15.98	0.706	0.228	20.12	0.840	0.136

5 Experiments

5.1 Experimental Setups

Implementation Details. Our framework is built upon 4DGaussians [41] and Scaffold-GS [4] codebase, with most parameters unchanged. We use DepthAnythingV2 [46] for per-frame depth estimation and Unimatch [47] for optical flow computation as 2D priors. All experiments are conducted on a single NVIDIA RTX 3090 GPU, with training and inference performed under identical settings to ensure fairness. Additional implementation details and hyperparameter analysis are provided in Suppl. A and B.

Datasets and Metrics. KineticGS is evaluated on monocular video datasets: iPhone dataset [9], which features handheld sequences with complex motion and illumination variation, and the HyperNeRF dataset [26], which provides controlled videos with continuous non-rigid deformations. We adopt the interp subset of HyperNeRF to emphasize smooth deformation patterns, enabling focused analysis of temporal coherence, physical plausibility, and geometric consistency under regulated motion. These datasets encompass both realistic and controlled scenarios, enabling a comprehensive assessment of reconstruction robustness, physical accuracy, and motion generalization. Following prior works, we report PSNR [10], SSIM [40], and LPIPS [51], where PSNR and SSIM quantify photometric and structural accuracy and LPIPS reflects perceptual realism. We report metrics at two levels: globally across the scene, and within dynamic object regions extracted by U^2-Net [29], to evaluate overall reconstruction quality and motion fidelity, respectively.

Comparison Methods. We compare KineticGS with representative baselines, including Deformable 3DGS [48], 4DGaussians [41], and MoDec-GS [17], which are the primary Gaussian-based methods commonly used for comparison in this line of research [17, 33]. All models are trained using official implementations

Table 2: Overall quantitative comparison on the Dycheck-iPhone [9] and HyperNeRF-interp [26] datasets. Average results across all scenes are reported, with the best and second-best results highlighted in Red and Yellow, respectively.

Method	(a) Dycheck-iPhone [9]			(b) HyperNeRF-interp [26]		
	mPSNR ↑	mSSIM ↑	mLPIPS ↓	mPSNR ↑	mSSIM ↑	mLPIPS ↓
Deformable 3DGS [48]	12.60	0.303	0.590	25.90	0.764	0.291
4DGaussians [41]	14.24	0.325	0.532	27.59	0.802	0.287
Modec-GS [17]	14.42	0.336	0.524	27.78	0.827	0.219
KineticGS (Ours)	14.65	0.342	0.511	27.92	0.828	0.221

under the same settings for fair comparison. In the main paper, we focus on methods based on Gaussian representations; additional comparisons with NeRF-based approaches are provided in the Suppl. E.

5.2 Quantitative Comparison

Table 1 summarizes the quantitative comparison on reconstruction quality within dynamic object regions. KineticGS consistently achieves the best or second-best performance across all DyCheck-iPhone sequences [9] and ranks first on average. One-stage deformation methods such as Deformable 3DGS [48] and 4DGaussians [41] struggle to jointly capture structural and fine-grained motions using a single MLP during inference, leading to less competitive results. While the hierarchical approach MoDec-GS [17] improves performance, it lacks physically guided initialization and explicit structural constraints, resulting in less coherent reconstructions and degraded novel-view synthesis.

Table 2 reports averaged results over the DyCheck-iPhone [9] and HyperNeRF-interp [26] datasets. On the challenging DyCheck-iPhone dataset, our method achieves the best performance across all reported metrics. While our mLPIPS score is slightly higher than that of Modec-GS [17] in the HyperNeRF-interp dataset, we consistently outperform all other baselines in mPSNR and mSSIM, metrics that directly reflect photometric accuracy and structural fidelity. This demonstrates KineticGS's strong capability in maintaining structural integrity and visual quality in dynamic scenes. Notably, our physically-grounded representation also offers high compactness, requiring only about 25% of the storage consumed by 4DGaussians [41]. Due to page constraints, we relegate the comprehensive analysis of storage efficiency and dynamic region accuracy to Suppl. C.

Effectiveness. While KineticGS achieves meaningful progress in monocular 4D reconstruction—improving mPSNR by ∼0.9dB over the strongest baselines, sequence-averaged metrics may not fully capture its distinctive strengths, as they are inherently biased toward static frames or low-motion regions. To intuitively demonstrate the effectiveness of KineticGS designed for fine-grained motion modeling, we examine per-frame PSNR curves (Fig. 4). Our method delivers its largest improvements precisely on frames with significant object motion—evident in the knife blade region (00:04–00:17), where both fidelity and

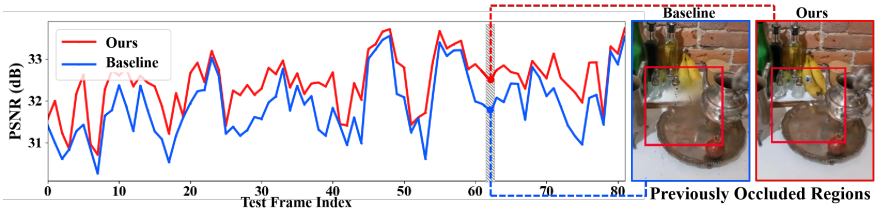


Fig. 4: Per-frame PSNR comparison across a sequence.

temporal consistency are visibly enhanced. This gain in dynamic regions stems from two key mechanisms designed to handle occlusion and reappearance: 1) momentum-driven activation, where each particle learns a smooth temporal energy window rather than relying on frame-wise visibility, and 2) kinetic synchronization, which enforces motion coherence within local neighborhoods. As a result, particles retain continuous latent momentum even when temporarily occluded, avoiding abrupt disappearance and unstable re-initialization. Meanwhile, visible neighboring particles propagate consistent motion cues through the coherence regularizer, allowing occluded regions to inherit structurally aligned dynamics. Together, temporal continuity and spatial synchronization provide a momentum-consistent mechanism that suppresses drift and enables stable reappearance of previously hidden regions, as demonstrated in Fig. 4.

5.3 Qualitative Comparison

As shown in Fig. 5, KineticGS produces sharper and more temporally consistent reconstructions than previous approaches, especially in challenging regions such as object boundaries and moving areas. KineticGS captures sharp object boundaries, exemplified by the knife edge in the *slice-banana* scene and facial outlines in the *space-out* scene. This improvement stemd from momentum-guided particle initialization, which concentrates Gaussians in high-motion areas to better resolve fine structures, while coherence constraints (Eq. (12)) maintain geometric consistency during rapid motion. Furthermore, KineticGS preserves stronger structural integrity, evident in the *broom* scene where the thin broomstick retains clear and stable edges despite fast movement. This is attributed to the joint enforcement of energy-flow activation and holistic kinetic synchronization, which enhances spatial-temporal coherence within object interiors. Additional visual results (in video format) are provided in the supplementary material, demonstrating consistent performance gains across varied motion types.

5.4 Ablation Studies

The effectiveness of each proposed component in KineticGS is evaluated via ablation experiments on dynamic object regions, using a hierarchical baseline with

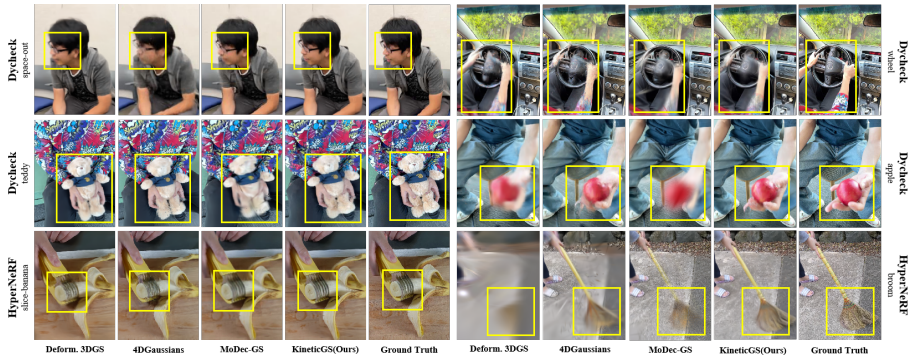
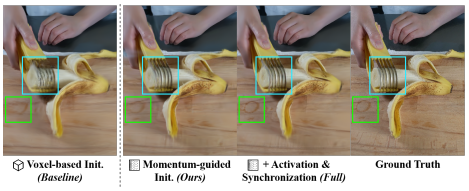


Fig. 5: Qualitative Comparison. KineticGS produces sharper and more physically coherent reconstructions than methods lacking comprehensive motion representation—evident in fine geometric details and consistent textures, particularly along object boundaries. *(Yellow boxes highlight dynamic regions with noticeable motion.)*

Components			mPSNR↑	mSSIM↑	mLPIPS↓
Initialization	Learnable Act.	Kinetic Sync.			
Voxel-based			19.17	0.829	0.141
Momentum-Guided			19.68	0.836	0.138
Momentum-Guided	✓		19.76	0.837	0.134
(Ours)	✓	✓	20.12	0.840	0.136

(a) Quantitative Comparison.



(b) Qualitative Comparisons.

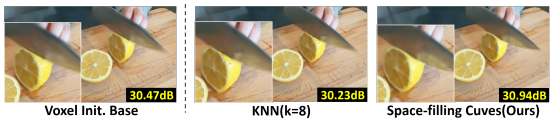
Fig. 6: Ablation study. Key components of KineticGS incrementally improve performance within dynamic object regions.*(Blue and green boxes highlight regions with noticeable motion and texture differences.)*

conventional voxel-based initialization. As shown in Table 6a, our momentum-guided initialization brings a +0.74 gain in mPSNR over the baseline, underscoring its contribution to structural fidelity. This improvement stems from concentrating particles in high-momentum regions, which enhances the capture of fast or non-rigid motion. In the blue boxed region of Fig. 6b, for instance, the moving knife edges exhibits clearer structure with momentum-guided initialization compared to the baseline. The Learnable Energy-flow Activation module further enhances performance by adaptively controlling particle activation over time. This mechanism encourages temporally coherent motion propagation and enables implicit identification of dynamic regions, reducing visual artifacts and improving perceptual quality in the reconstructed sequences as shown in the green boxed region of Fig. 6b. The incorporation of Holistic Kinetic Synchronization yields an additional +0.36 mPSNR gain by enforcing momentum consistency among neighboring particles. This enhances spatial coherence, thereby reducing fragmentation and structural noise, leading to the highest reconstruction quality.

We further analyze the effect of neighborhood construction on motion coherence, which plays a critical role in determining how particle interactions are

Method	PSNR↑	SSIM↑
Voxel Init. Base	14.41	0.335
KNN ($k = 8$)	14.53	0.337
Space-filling Curves (ours)	**14.65**	**0.342**

(a) Qualitative Comparison.



(b) Qualitative Comparison.

Fig. 7: Additional ablation results on neighborhood construction.

organized within motion entities. As shown in Fig. 7, both quantitative and qualitative results indicate that, compared to KNN-based grouping, our curve-based construction provides a more structured spatial ordering of particles. This structured organization leads to more coherent motion boundaries within dynamic regions, as visualized in Fig. 7b, where object structures are more clearly delineated.

6 Conclusion

We present KineticGS, a momentum-driven Gaussian Splatting method for reconstructing dynamic scenes from monocular videos. By modeling Gaussian primitives using momentum-carrying particles organized within a physically grounded hierarchy, KineticGS captures complex motions while preserving fine-grained spatiotemporal dynamics. Through momentum-guided particle initialization, learnable activation for energy-aware regions, and kinetic synchronization within holistic objects, KineticGS achieves high geometric accuracy and physical plausibility in 4D reconstruction. Experiments show strong performance in fidelity and temporal coherence, outperforming existing methods on multiple benchmarks with robust structure under challenging motion.

Limitations and future work. In our approach, we adopt the first frame as a reference, following practices seen in some canonical-space methods. While effective for smoothly deforming objects, it struggles with multiple interacting entities or reciprocal motions, such as frequent object entry and exit. This motivates future work on adaptive canonical representations that update continuously with new observations for robust open-world reconstruction.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62472105; in part by the Natural Science Foundation of Guangdong Province under Grants 2025A1515011385, 2024A1515010186, and 2025A1515011556; in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2023A1515110706; in part by the Youth Scientific Talent Cultivation Program of the Guangdong Provincial Association for Science and Technology under Grant SKXRC2026513; and in part by the Research Foundation of the Guangdong-Hong Kong-Macao Applied Mathematics Center under Grant 2025A1515060016.

References

1. Adamkiewicz, M., Chen, T., Caccavale, A., Gardner, R., Culbertson, P., Bohg, J., Schwager, M.: Vision-only robot navigation in a neural radiance world. IEEE Robotics and Automation Letters **7**(2), 4606–4613 (2022)
2. Bhoi, A.: Monocular depth estimation: A survey. arXiv preprint arXiv:1901.09402 (2019)
3. Chen, J., Li, Z., Cai, Y., Jiang, H., Qian, C., Kang, J., Gao, S., Zhao, H., Mao, T., Zhang, Y.: Haif-gs: Hierarchical and induced flow-guided gaussian splatting for dynamic scene. arXiv preprint arXiv:2506.09518 (2025)
4. Cho, W.O., Cho, I., Kim, S., Bae, J., Uh, Y., Kim, S.J.: 4d scaffold gaussian splatting for memory efficient dynamic scene reconstruction. arXiv preprint arXiv:2411.17044 (2024)
5. Fan, Z., Wang, K., Wen, K., Zhu, Z., Xu, D., Wang, Z., et al.: Lightgaussian: Unbounded 3d gaussian compression with 15x reduction and 200+ fps. Advances in neural information processing systems **37**, 140138–140158 (2024)
6. Fang, J., Yi, T., Wang, X., Xie, L., Zhang, X., Liu, W., Nießner, M., Tian, Q.: Fast dynamic radiance fields with time-aware neural voxels. In: SIGGRAPH Asia 2022 Conference Papers. pp. 1–9 (2022)
7. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12479–12488 (2023)
8. Gao, C., Saraf, A., Kopf, J., Huang, J.B.: Dynamic view synthesis from dynamic monocular video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5712–5721 (2021)
9. Gao, H., Li, R., Tulsiani, S., Russell, B., Kanazawa, A.: Monocular dynamic view synthesis: A reality check. Advances in Neural Information Processing Systems **35**, 33768–33780 (2022)
10. Hore, A., Ziou, D.: Image quality metrics: Psnr vs. ssim. In: 2010 20th international conference on pattern recognition. pp. 2366–2369. IEEE (2010)
11. Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., Nie, L.: Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 634–644 (2024)
12. Huang, Y.H., Sun, Y.T., Yang, Z., Lyu, X., Cao, Y.P., Qi, X.: Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4220–4230 (2024)
13. Jang, H., Kim, D.: D-tensorf: Tensorial radiance fields for dynamic scenes. arXiv preprint arXiv:2212.02375 (2022)
14. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Trans. Graph. **42**(4), 139–1 (2023)
15. Kong, H., Yang, X., Wang, X.: Efficient gaussian splatting for monocular dynamic scene rendering via sparse time-variant attribute modeling. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 4374–4382 (2025)
16. Kullback, S., Leibler, R.A.: On information and sufficiency. The annals of mathematical statistics **22**(1), 79–86 (1951)

17. Kwak, S., Kim, J., Jeong, J.Y., Cheong, W.S., Oh, J., Kim, M.: Modec-gs: Global-to-local motion decomposition and temporal interval adjustment for compact dynamic 3d gaussian splatting. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 11338–11348 (2025)
18. Lei, J., Weng, Y., Harley, A.W., Guibas, L., Daniilidis, K.: Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 6165–6177 (2025)
19. Li, Z., Chen, Z., Li, Z., Xu, Y.: Spacetime gaussian feature splatting for real-time dynamic view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8508–8520 (2024)
20. Liang, Y., Xu, T., Kikuchi, Y.: Himor: Monocular deformable gaussian reconstruction with hierarchical motion representation. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 886–895 (2025)
21. Lin, Y., Dai, Z., Zhu, S., Yao, Y.: Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21136–21145 (2024)
22. Liu, H., Liu, Y., Li, C., Li, W., Yuan, Y.: Lgs: A light-weight 4d gaussian splatting for efficient surgical scene reconstruction. arXiv preprint arXiv:2406.16073 (2024)
23. Lu, T., Yu, M., Xu, L., Xiangli, Y., Wang, L., Lin, D., Dai, B.: Scaffold-gs: Structured 3d gaussians for view-adaptive rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20654–20664 (2024)
24. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
25. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5865–5874 (2021)
26. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. arXiv preprint arXiv:2106.13228 (2021)
27. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10318–10327 (2021)
28. Qian, Z., Wang, S., Mihajlovic, M., Geiger, A., Tang, S.: 3dgs-avatar: Animatable avatars via deformable 3d gaussian splatting. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5020–5030 (2024)
29. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U2-net: Going deeper with nested u-structure for salient object detection. Pattern recognition **106**, 107404 (2020)
30. Reiser, C., Szeliski, R., Verbin, D., Srinivasan, P., Mildenhall, B., Geiger, A., Barron, J., Hedman, P.: Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. ACM Transactions on Graphics (ToG) **42**(4), 1–12 (2023)
31. Rosinol, A., Leonard, J.J., Carlone, L.: Nerf-slam: Real-time dense monocular slam with neural radiance fields. In: 2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 3437–3444. IEEE (2023)
32. Shao, Z., Wang, Z., Li, Z., Wang, D., Lin, X., Zhang, Y., Fan, M., Wang, Z.: Splattingavatar: Realistic real-time human avatars with mesh-embedded gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1606–1616 (2024)

33. Stearns, C., Harley, A., Uy, M., Dubost, F., Tombari, F., Wetzstein, G., Guibas, L.: Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
34. Sun, J., Xie, Y., Chen, L., Zhou, X., Bao, H.: Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15598–15607 (2021)
35. Tian, F., Du, S., Duan, Y.: Mononerf: Learning a generalizable dynamic radiance field from monocular videos. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 17903–17913 (2023)
36. Uy, M.A., Martin-Brualla, R., Guibas, L., Li, K.: Scade: Nerfs from space carving with ambiguity-aware depth estimates. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16518–16527 (2023)
37. Wang, F., Zhu, Q., Chang, D., Gao, Q., Han, J., Zhang, T., Hartley, R., Pollefeys, M.: Learning-based multi-view stereo: A survey. arXiv preprint arXiv:2408.15235 (2024)
38. Wang, Q., Ye, V., Gao, H., Zeng, W., Austin, J., Li, Z., Kanazawa, A.: Shape of motion: 4d reconstruction from a single video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9660–9672 (2025)
39. Wang, S., Yang, X., Shen, Q., Jiang, Z., Wang, X.: Gflow: Recovering 4d world from monocular video. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 7862–7870 (2025)
40. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
41. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20310–20320 (2024)
42. Wu, X., Jiang, L., Wang, P.S., Liu, Z., Liu, X., Qiao, Y., Ouyang, W., He, T., Zhao, H.: Point transformer v3: Simpler faster stronger. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4840–4851 (2024)
43. Xie, T., Zong, Z., Qiu, Y., Li, X., Feng, Y., Yang, Y., Jiang, C.: Physgaussian: Physics-integrated 3d gaussians for generative dynamics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4389–4398 (2024)
44. Xu, H., Peng, S., Wang, F., Blum, H., Barath, D., Geiger, A., Pollefeys, M.: Depthsplat: Connecting gaussian splatting and depth. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16453–16463 (2025)
45. Yan, J., Peng, R., Tang, L., Wang, R.: 4d gaussian splatting with scale-aware residual field and adaptive optimization for real-time rendering of temporally complex dynamic scenes. In: Proceedings of the 32nd ACM International Conference on Multimedia. pp. 7871–7880 (2024)
46. Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. Advances in Neural Information Processing Systems **37**, 21875–21911 (2024)
47. Yang, L., Zhao, Z., Zhao, H.: Unimatch v2: Pushing the limit of semi-supervised semantic segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **47**(4), 3031–3048 (2025)
48. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: Proceedings of the

IEEE/CVF conference on computer vision and pattern recognition. pp. 20331–20341 (2024)
49. Yu, Z., Peng, S., Niemeyer, M., Sattler, T., Geiger, A.: Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. Advances in neural information processing systems **35**, 25018–25032 (2022)
50. Yuan, Z., Cao, J., Li, Z., Jiang, H., Wang, Z.: Sd-mvs: Segmentation-driven deformation multi-view stereo with spherical refinement and em optimization. In: Proceedings of the AAAI conference on artificial intelligence. vol. 38, pp. 6871–6880 (2024)
51. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
52. Zhu, R., Liang, Y., Chang, H., Deng, J., Lu, J., Yang, W., Zhang, T., Zhang, Y.: Motiongs: Exploring explicit motion guidance for deformable 3d gaussian splatting. Advances in Neural Information Processing Systems **37**, 101790–101817 (2024)