SMG: Semantic Motion Graph for Monocular Dynamic Gaussian Splatting

Haozheng Yu, Xinyu Yang, Rundong Luo,
Jennifer J. Sun, and Bharath Hariharan

Cornell University

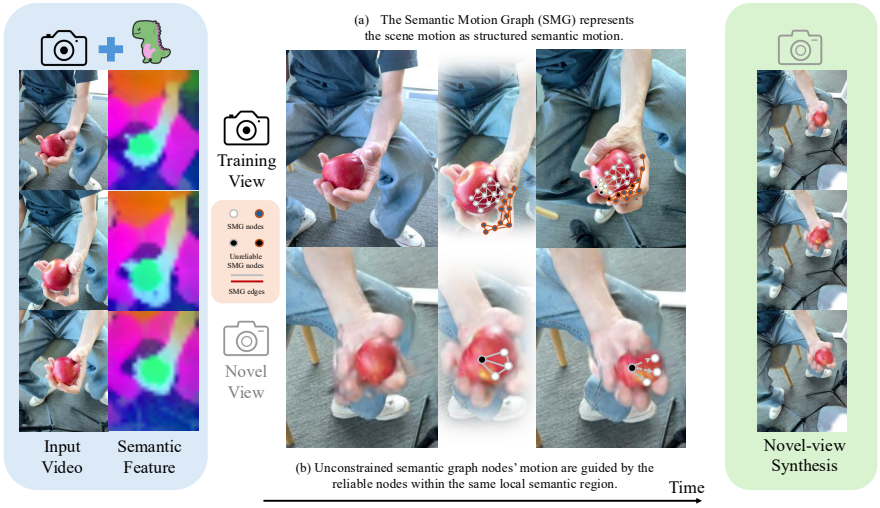


Fig. 1: We present Semantic Motion Graph (SMG), a monocular dynamic Gaussian splatting framework that models the Gaussian motion as structured semantic motion. SMG handles the uncertainty within the Gaussian optimization. (a). To handle the uncertainty of the Gaussian motion with limited observations, the reliable SMG nodes are used to guide the unreliable motion (b). Our method yields robust dynamic Gaussians that generalize to novel views under challenging scenarios.

Abstract. We study dynamic Gaussian Splatting from monocular videos. While recent advancements in dynamic Gaussian Splatting offer a promising foundation for modeling dynamic scenes, they often overfit to the training views and fail under occlusion or complex scene motion due to the lack of reliable regularization signals in under-constrained regions. We propose Semantic Motion Graph (SMG), a novel approach that models the Gaussian motion as the semantic motion. Our key insight is that the real-world scene motion is often structured by semantic coherence: regions that are spatially close and semantically related tend to exhibit consistent dynamics. To leverage this prior, we build SMG to model structured motion of the scene. The Gaussian motion is driven by the motion of SMG nodes. We further observe that the uncertainty of Gaussian

motion arises from both unreliable off-the-shelf priors and weakly constrained regions during optimization. SMG addresses this by using reliable graph nodes to guide the motion of nearby unreliable nodes. To evaluate dynamic Gaussian Splatting under challenging real-world scenarios, we introduce a new multiview dataset collected under an ego-exo setup. Extensive experiments demonstrate that SMG achieves state-of-the-art performance on monocular dynamic Gaussian Splatting across challenging real-world benchmarks. Project: https://smg-gaussian.github.io/.

1 Introduction

Reconstructing the dynamic scene from a monocular video is a crucial task for AR/VR, robotics, and motion capture, as it enables interaction, re-rendering, and 3D control. While recent advances in static 3D scene reconstruction, especially neural-field-based approaches [2, 4, 18, 43, 44, 71], have achieved remarkable photorealism, these approaches often break down when extended to dynamic scenes. The reason is fundamental: real-world videos rarely provide dense viewpoints, and object motion is inherently unpredictable. Without sufficient multiview constraints, distinguishing camera motion from object motion becomes extremely ill-posed. Occlusions and missing observations leave large portions of the scene unsupervised, causing severe overfitting, surface drifting, and inconsistent geometry when viewed from novel perspectives.

As a popular scene reconstruction representation, Gaussian Splatting also suffers under sparse viewpoints. Many Gaussians are only weakly constrained, and their colors and opacities often collude to fit the observed views while hallucinating geometry in unobserved regions, leading to floating artifacts and inconsistent reconstruction [10, 32, 70, 79]. Recent efforts [30, 38, 56, 61] introduce data-driven priors and physically-based priors [40] to regularize the dynamic Gaussian optimization under limited viewpoints. These methods assume that the irregular and complex per-frame motion can be decomposed into smooth and local rigid transformations, which regularizes dynamic Gaussian optimization under limited views. Yet, motion smoothness alone is insufficient: in many real scenes, occlusion hides entire body parts or objects, and purely geometric constraints fail to propagate correct motion into such regions.

In this work, we take the idea one step further. We observe that local rigidity is not merely a geometric assumption—it emerges from semantic consistency. Gaussians belonging to the same semantic region (e.g., a rigid object) tend to share coherent motion. Motivated by this, we introduce Semantic Motion Graph (SMG), a new mechanism that models the motion of Gaussians as SMG node deformation (Fig. 1). We first build the SMG based on lifted semantic tracks, which restrict motion propagation within local semantic groups. We also observe that the uncertainty of Gaussian motion arises from both unreliable priors and unconstrained regions during optimization. To address this issue, we propose confidence-aware as-rigid-as-possible (C-ARAP), a mechanism that propagates motion and maintains topology by guiding less reliable nodes with reliable

ones. We further design local rigid motion control (LRM), which handles uncontrollable node velocity changes when observations are limited. Together, these mechanisms provide additional regularization for SMG, enabling correct motion propagation and better generalization to challenging novel views.

To evaluate dynamic Gaussian Splatting under complex camera and object motion, we propose SMG dataset, a custom multiview dataset captured in an ego-exo setup. Together with SMG, they enable a more comprehensive understanding of dynamic Gaussian Splatting under challenging real-world conditions.

To summarize, our contributions are:

- We propose SMG, a monocular dynamic Gaussian Splatting pipeline that models the Gaussian motion as a semantic motion graph.
- We develop SMG-based motion uncertainty modeling that propagates motion from reliable nodes to less visible nodes within the same semantic group, thereby handling uncertainty during dynamic Gaussian optimization.
- We introduce SMG dataset, a challenging multiview dataset with complex camera and object motion for monocular dynamic Gaussian Splatting.
- SMG demonstrates state-of-the-art performance for novel-view synthesis on challenging scenes and can also benefit other scene understanding and 3D tasks, e.g. 3D tracking.

2 Related Work

2.1 Dynamic Scene Representation

Reconstructing and synthesizing the real world is a fundamental question in the computer vision community. NeRF [43] is one of the pioneering works in this field, which introduced the idea of learning a neural radiance field implicitly with a multi-layer perceptron (MLP). This approach also enables novel-view synthesis via volumetric rendering. However, optimization and rendering via the MLP is time-consuming, even when using GPUs. A series of follow-up works address the rendering speed [8,18,44] or the rendering quality [2–4,71] issues of NeRF. Some works further explored using NeRF to represent complex dynamic scenes, for example, by modeling a deformation field based on a canonical field [14, 15, 47, 49], directly learning a neural radiance field conditioned on time [19, 33, 48], or learning a hybrid representation to parameterize the dynamic scene [5, 17].

Recently, 3D Gaussian Splatting (3DGS) [24] has gained significant attention for its efficient rendering, fast training speed, and outstanding visual effects. Unlike NeRF, 3DGS represents the scene as a set of 3D Gaussians without the MLP. Built on the success of 3DGS [24], following works have further expanded this representation to model dynamic scenes by encoding the 3D Gaussian position alongside its translation and rotation over time [40], optimizing the deformation field of the canonical Gaussians [63, 68], or modeling time as an additional dimension [66]. However, when applied to real-world videos with sparse viewpoints, 3D Gaussians often overfit the training views and produce unreliable reconstructions from novel views [10, 32, 70, 79]. Recent and concurrent efforts

explore learning feed-forward transformer-based models to reconstruct dynamic Gaussians [65], or incorporating strong data-driven priors to regularize the optimization of dynamic Gaussians and enables reliable novel view synthesis from in-the-wild videos [30,38,46,53,56,61,64]. However, under severe occlusions or when the learned priors do not generalize, these methods may still fail. We observe that real-world motion is semantically structured, and introduce a semantic-aware Gaussian motion propagation framework that regularizes the deformation of under-constrained Gaussians, leading to more stable dynamics and improved novel-view synthesis.

2.2 3D Feature Field

With the development of vision foundation models [27, 31, 45, 51, 52], recent approaches explore to distill the high-level features into neural fields. Early effort [75] encodes semantic labels in NeRF [43] to enable semantic completion from noisy 2D labels. LERF [25] and its following works [26,37] distill CLIP [51] and DINO [6] features into NeRF to enable open-vocabulary 3D queries. Recent approaches [7, 50, 77] focus on distilling the 3D feature fields with Gaussian splatting, enabling efficient 3D segmentation and understanding.

Recent works explore distilling semantic fields on top of dynamic Gaussians to enable semantic-driven tracking and Gaussian editing [16, 28, 55]. Concurrent efforts [34, 78] further incorporate MLLMs for user-interactive 4D VQA tasks. However, these works solely treat semantic fields as an auxiliary output of the dynamic Gaussian Splatting pipelines, ignoring the fact that the dynamics in real-world monocular videos are often semantically structured. In contrast, our goal is fundamentally different: we exploit semantic cues as a strong regularization of the Gaussian deformation under sparse observations. We leverage the semantically structured nature of motion and introduce a semantic motion graph that transforms the Gaussian motion into the semantic graph motion. We further propose uncertainty modeling strategies that guide the graph motion under noisy priors or unreliable optimization, leading to more stable motion propagation and improved novel-view synthesis.

2.3 3D Scene Understanding

Our work is also related to the area of 3D scene understanding more broadly. This area includes tasks such as sparse and dense 3D grounding [1,9,11,21,67,74], 3D question answering [13,41] and 3D captioning [12,13]. 3D understanding fundamentally differs from 2D understanding by demanding a comprehensive grasp of spatial geometry, occlusion, and physical properties. However, the scarcity of 3D data and the inherent limitations of 2D data to recover 3D information pose significant challenges for 2D-trained vision foundation models in effectively understanding 3D scenes. Initialized by the 2D priors and the semantics, our work addresses this gap by learning the 3D feature fields and tracks in the process of optimizing dynamic Gaussians. Our approach, similar to other works based on 3D Gaussians [34, 78], employs test-time optimization on individual scenes.

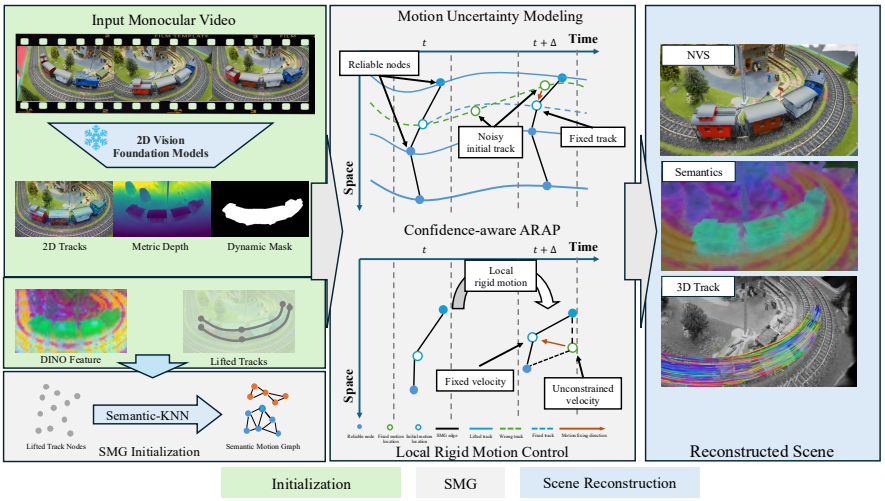


Fig. 2: Overview of SMG. Given a monocular video with 2D tracks, depth, and dynamic/static masks predicted by the off-the-shelf models, the Semantic Motion Graph is initialized with the 3D lifting field in the nearby semantic groups. To handle the uncertainty during the Gaussian optimization, we design confidence-aware ARAP (C-ARAP) and local rigid motion control (LRM) to preserve the graph topology and guide the motion of unreliable nodes by the reliable nodes, producing semantically structured and temporally coherent motion. Finally, our proposed SMG improves the robustness of novel-view synthesis and yields dynamic 3D scenes enriched with consistent semantics and 3D tracks.

We aim to bridge the dimensional gap and facilitate robust spatial understanding from monocular videos (e.g. 3D object segmentation and tracking), thereby contributing to 3D scene understanding more broadly.

3 Method

We introduce Semantic Motion Graph (SMG), a novel method that models the motion of semantic groups for dynamic Gaussian Splatting under a monocular setup. The SMG is first initialized with the 3D lifting field within nearby semantic groups, which ensures that the motion is only propagated locally (Sec. 3.1). To handle the uncertainty during the optimization, the confidence-aware as-rigid-as-possible (C-ARAP) and local-rigid-motion (LRM) controls are incorporated into SMG to ensure that the semantic motion is propagated correctly under unreliable supervision (Sec. 3.2). Finally, the dynamic Gaussians are initialized and optimized under the guidance of SMG (Sec. 3.3), yielding dynamic Gaussians that remain stable under novel views.

3.1 Semantic Motion Graph (SMG)

A major challenge in monocular dynamic Gaussian Splatting is that motion in unobserved regions is inherently ill-posed. Recent methods [30, 56, 61] utilize the strong off-the-shelf priors to construct 3D lifting fields to regularize the motion in the unconstrained regions.

We argue that resolving such ambiguity requires modeling how motion correlations are structured. In real-world scenarios, the motion coherence is structured by semantic coherence rather than mere geometric proximity. To this end, we present Semantic Motion Graph (SMG) (Fig. 2). Given the 2D tracks, visibility, and the depth estimated by the off-the-shelf models, we construct the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with the lifted tracks as nodes. We restrict connectivity to spatial local neighbors to preserve geometric continuity, while enforcing semantic consistency to avoid propagating motion across unrelated nearby objects. Formally, the edge $\mathcal{E}_{ij}$ between nodes $i, j \in \mathcal{V}$ is defined as

$$\mathcal{E}_{ij} = \mathbf{1}[j \in \mathrm{KNN}(i)] \cdot \mathbf{1}[\cos(f_i, f_j) \geq \tau] \tag{1}$$

where $\mathrm{KNN}(\cdot)$ is performed at the trajectory level, using the top-k distances computed over frames in which both i and j are visible. τ is the semantic gating threshold. $\cos(f_i, f_j)$ denotes the cosine similarity of the semantic feature of i and j. To obtain the semantic feature of each trajectory, we first extract a per-frame semantic feature map and project each trajectory point to image coordinates in its visible frames. We then aggregate these frame-wise features by top-k cosine-consistent averaging to obtain a robust trajectory-level semantic descriptor. We use DINOv3 [54] as our semantic feature extractor $\mathrm{f}(\cdot)$ in all our experiments.

3.2 Modeling Motion Uncertainty in Dynamic Gaussians

Dynamic Gaussian Splatting under monocular setups is inherently ill-posed and prone to uncertainty. Noisy off-the-shelf priors can be a source of uncertainty. During optimization, regions with limited visibility are not well constrained, which introduces additional uncertainty into the system. To address these challenges, we introduce confidence-aware as-rigid-as-possible (C-ARAP) and local rigid motion control (LRM) to handle the uncertainty within the whole optimization process.

Confidence-aware as-rigid-as-possible. Dynamic Gaussian Splatting under the monocular setup suffers from uncertainty introduced by off-the-shelf priors. Due to occlusions and complex object motions in monocular videos, the estimated depth and tracks inevitably introduce noise to the system. The proposed SMG is initialized from the lifted 2D tracks, which may lead to both reliable and unreliable motion being propagated in the graph.

We use confidence-aware as-rigid-as-possible (C-ARAP) to propagate motion within the semantic groups. Rigidity modeling of dynamic Gaussians [40] and lifting field nodes [30] has been introduced by previous works. However, modeling all the Gaussians or nodes fairly introduces uncertainty during motion propagation. We modulate each SMG edge using node-level confidence estimated from

visibility and motion consistency. This allows local motion to be propagated by reliable nodes within a local semantic group while suppressing the nodes with high uncertainty to propagate the unreliable motions. Given the per-frame node visibility $\text{viz}_{t,i} \in \{0, 1\}$ from the off-the-shelf point tracking model, the per-frame node confidence of node i is defined as

$$c_{t,i} = \text{viz}_{t,i} \cdot (c_{t,i}^{\text{motion}})^{\lambda_m} \cdot (c_{t,i}^{\text{rigid}})^{\lambda_r}, \tag{2}$$

$$c_{t,i}^{\text{motion}} = \exp\left(-\left(\frac{r_{t,i}^{v}}{\kappa_m(s_{t,i}^{v}+\varepsilon)}\right)^2\right), \quad c_{t,i}^{\text{rigid}} = \exp\left(-\left(\frac{r_{t,i}^{\ell}}{\kappa_r(s_{t,i}^{\ell}+\varepsilon)}\right)^2\right), \tag{3}$$

where $c_{t,i}^{\text{motion}}$ and $c_{t,i}^{\text{rigid}}$ are the framewise node motion and rigid confidence. r^v and r^l denote the deviation of the node velocity from the neighborhood average velocity and the abnormal change of edge length across adjacent frames, respectively; larger values indicate lower reliability due to motion inconsistency or local non-rigidity. s^v and s^l are the absolute deviation over neighbors used for adaptive normalization. κ_m and κ_r are the soft thresholds controlling the tolerance level. This framewise node confidence determines the reliability of the node motion that reflects the local motion. Once the per-frame node confidence is obtained, we define the asymmetric confidence-aware edge weight of SMG as

$$\tilde{w}_{ij}^{t} = w_{ij}^{\text{topo}} \left(\alpha + (1-\alpha)c_{t,j}\right), \tag{4}$$

where $\tilde{w}_{ij}^{t}$ denotes the asymmetric edge weight from node i to node j at time t and $\alpha \in [0, 1]$ is a lower bound that prevents the edge weight from collapsing to zero. w_{ij}^{topo} is the weight of SMG edge E_{ij} normalized by topological connectivity. The motion of nodes with low framewise confidence should be guided by nearby reliable nodes rather than vice versa. Finally, the C-ARAP is computed as

$$L_{\text{C-ARAP}} = \sum_{t,i} \sum_{j \in \mathcal{N}(i)} \tilde{w}_{ij}^{t} \left[\lambda_c \left(\left\| d_{ij}^{t+\delta} - d_{ij}^{t} \right\|\right)^2 + \lambda_l \left(\left| \|d_{ij}^{t+\delta}\| - \|d_{ij}^{t}\| \right|\right)^2\right], \tag{5}$$

$$d_{ij}^{t} = (R_i^t)^{\top}(x_j^t - x_i^t), \tag{6}$$

where $j \in \mathcal{N}(i)$ is the neighboring nodes of node i. λ_c and λ_l are the parameters that control the relative strength of the vector-consistency term and the length-preservation term. d_{ij}^{t} is computed as the multiplication of the orientation R_i^t of node i at time t and the world displacement $x_j^t - x_i^t$ from node i to node j at time t.

Local rigid motion under unconstrained regions. Beyond the uncertainty introduced by off-the-shelf priors, monocular dynamic Gaussians are also sensitive to the estimated velocity field. When geometric constraints are weak, optimization can produce locally inconsistent velocities, leading to structural drifts.

To stabilize local motion, we introduce a local rigid motion control that constrains the local velocity field to be explained by a rigid twist. Given a node

i and its neighboring nodes $j \in \mathcal{N}(i)$, the average velocity of the local region at time t, $\bar{v}_i^t$ is defined as the average velocity of j weighed by the edge weight. The local center g_i^t is defined as the weighted average location of j. We compute the residual of the motion of each j as

$$\tilde{v}_{ij}^t = v_j^t - \bar{v}_i^t, \quad \tilde{x}_{ij}^t = x_j^t - g_i^t. \tag{7}$$

We then estimate the local angular velocity ω_i^t that best represents local motion via weighted least squares:

$$\omega_i^t = \arg\min_{\omega} \sum_{j \in \mathcal{N}(i)} \tilde{w}_{ij}^t \left\| \tilde{v}_{ij}^t - \omega \times \tilde{x}_{ij}^t \right\|^2. \tag{8}$$

Finally, we penalize the local node velocities that cannot be explained by the estimated local rigid motion. The local rigid motion loss (LRM) is computed as

$$L_{\text{LRM}} = \sum_{t,i} \sum_{j \in \mathcal{N}(i)} \tilde{w}_{ij}^t \rho \left(\left\| \tilde{v}_{ij}^t - \omega_i^t \times \tilde{x}_{ij}^t \right\|_2 \right). \tag{9}$$

where $\rho(\cdot)$ denotes the Huber robust penalty.

3.3 SMG-guided Dynamic Gaussian Optimization

Gaussian initialization and deformation. After obtaining the SMG, we optimize dynamic Gaussians by coupling photometric supervision with the SMG-based semantic motion regularizations mentioned above.

Given the dynamic mask computed by an off-the-shelf flow model, we sample pixels from dynamic regions with valid depth and back-project them to 3D to initialize dynamic Gaussians. Each Gaussian is parameterized as $\mathcal{G}_i = \{\mathbf{x}_i^{ref}, \mathbf{R}_i^{ref}, \mathbf{s}_i, \alpha_i, \mathbf{f}_i, t_i^{ref}\}$, where $\mathbf{x}_i^{ref} \in \mathbb{R}^3$ is the reference position, $\mathbf{R}_i^{ref} \in SO(3)$ is the reference rotation, $\mathbf{s}_i \in \mathbb{R}^3$ is the scale, $\alpha_i \in (0,1)$ is opacity, $\mathbf{f}_i$ denotes SH coefficients and t_i^{ref} is the reference timestamp.

Each Gaussian is attached to its nearest SMG node at t_i^{ref} as a local anchor. Instead of being controlled by a single node, it is driven by K semantic neighbors of that node with distance-aware normalized weights $\{w_{ik}\}_{k=1}^K$. At timestep t, we obtain the Gaussian pose by blending node motions from t_i^{ref} to t:

$$(\mathbf{x}_i^t, \mathbf{R}_i^t) = \mathcal{W}\left(\{\mathcal{T}_k^{t_i^{ref} \to t}\}_{k=1}^K, \{w_{ik}\}_{k=1}^K, \mathbf{x}_i^{ref}, \mathbf{R}_i^{ref} \right), \quad \sum_{k=1}^K w_{ik} = 1. \tag{10}$$

where $\mathcal{T}_k^{t_i^{ref} \to t}$ is the rigid motion of the k-th controlling SMG node from reference time to target time, and $\mathcal{W}(\cdot)$ is the weighted motion blending operator. We follow MoSca [30] and implement $\mathcal{W}(\cdot)$ with dual-quaternion blending (DQB) [23] for stable and smooth articulated deformation.

Method	mPSNR ↑	mSSIM ↑	mLPIPS ↓	Method	mPSNR ↑	mSSIM ↑	mLPIPS ↓
T-NeRF [20]	16.96	0.577	0.379	Gauss.Marbles [56]	16.72	-	0.413
NSFF [35]	15.46	0.551	0.396	DyBluRF [57]	17.37	0.591	0.373
Nerfies [47]	16.45	0.570	0.339	CTNeRF [42]	17.69	0.531	-
HyperNeRF [48]	16.81	0.569	0.332	D-NPC [22]	16.41	0.582	0.319
PGDVS [73]	15.88	0.548	0.340	MoSca [30]	19.32	0.706	0.264
DyPoint [76]	16.89	0.573	-	Shape-of-Motion [61]	17.32	0.598	0.296
DpDy [60]	-	-	0.516	OriGS* [64]	19.69	0.716	0.256
Dyn.Gauss. [40]	7.29	-	0.692	OriGS [64]	19.43	0.695	0.281
4DGS [63]	13.64	-	0.428	**Ours**	19.54	0.718	0.250

Table 1: Comparison of novel view synthesis on the Dycheck (iPhone) dataset [20]. Following previous works [30], results are reported on half-resolution images. The best, second-best, and third-best results are highlighted in red, orange, and yellow. "*" indicates results reported in the OriGS paper [64].

Photometric Joint Optimization. We jointly optimize photometric rendering consistency and semantic motion consistency with the ground truth image and off-the-shelf priors. The training objective is:

$$\mathcal{L} = \lambda_{rgb}\mathcal{L}_{rgb} + \lambda_{dep}\mathcal{L}_{dep} + \lambda_{mask}\mathcal{L}_{mask} + \lambda_{track}\mathcal{L}_{track}. \tag{11}$$

During optimization, we update Gaussian parameters together with the SMG deformation. We further apply density controls to adapt the Gaussian population over time and improve reconstruction quality and stability.

4 Experiments

4.1 Implementation Details

Our model is implemented using PyTorch and trained on a single NVIDIA H100 GPU. We perform KNN search based on the 3D trajectory distance, which is aggregated over the three co-visible frames with the smallest distances. The scaffold topology is then constructed using the top-16 nearest neighbors. For C-ARAP, we use $\lambda_m = \lambda_r = 1.0$, $\kappa_m = \kappa_r = 2.5$, $\alpha = 0.6$ for all experiments. For the semantic gating parameter τ, we use $\tau = 0.75$ for Dycheck and NVIDIA experiments and $\tau = 0.85$ for SMG-Dataset experiments. For scene reconstruction experiments, we assessed the quality of our reconstructions using standard metrics: Peak Signal-to-Noise Ratio (PSNR), which quantifies pixel-level similarity; Learned Perceptual Image Patch Similarity (LPIPS) [72], which measures perceptual similarity; and Structural Similarity Index (SSIM) [62], which evaluates structural similarity.

4.2 Evaluations on Dycheck Dataset

We report our results on Dycheck for novel-view synthesis. Dycheck is an iPhone-captured dataset featuring camera pose and sensor depth. We follow the baselines [30, 56] and evaluate on all 7 scenes: Apple, Block, Paper Windmill, Space

Fig. 3: Qualitative results of novel-view synthesis on Dycheck dataset. SMG shows superior results against the baselines. SMG preserves the correct object geometry while effectively reducing floating or drifting dynamic Gaussians, leading to more stable and coherent reconstructions across viewpoints. Zoom-in to see details.

Out, Spin, Teddy, and Wheel with half resolution. We use camera view 0 for training and 1 and 2 for testing. Following the previous works, we compute the metrics on the covisible mask area. We report the quantitative results in Table 1. OriGS [64] does not provide the training hyperparameters on the iPhone dataset and the evaluation code. As their experiment follows MoSca's structure, we modify MoSca's hyperparameters and report the reproduced OriGS results. We also provide OriGS paper results as "OriGS*". Due to the wide baseline in some of the Dycheck scenes, e.g. Apple, previous methods that rely solely on photometric optimization fail under novel views. By incorporating off-the-shelf priors for regularization, 3D lifting fields [30, 61, 64] achieve much better reconstruction than their counterparts. However, they fail to handle the uncertainty introduced by the monocular priors and the optimization process. For fair comparison, we apply the same off-the-shelf priors with MoSca and OriGS and use the noisy iPhone sensor depth for training. By constructing the semantic lifting field and modeling the uncertainty within the whole optimization process, SMG outperforms the baselines on all metrics.

Qualitatively, SMG outperforms all baselines on producing less geometry drifting and more reliable object motion (Fig. 3). The Gaussian deformation field [63] produces severe artifacts and fails to reconstruct the object geometry because the deformation MLP tends to overfit to the training views when lacking

regularization. MoSca [30] renders overall reasonable scene geometry and motion under novel views when the monocular off-the-shelf priors are reliable. However, when the priors are noisy, MoSca mistakenly propagates the motion to the scaffold, leading to the breaking down of the object geometry (the human arm in the "Wheel" and "Spin" scene) and floating Gaussians (e.g. the human hand in the "Apple" scene). OriGS [64] inherits the weaknesses of MoSca while introducing more artifacts and worse geometry. Compared to the baseline methods, the semantic motion graph ensures that the motion propagation is restricted to nearby semantic regions. When the priors are not reliable, the proposed C-ARAP and LRM preserve the object geometry and motion optimized when the supervision is reliable, and reduce the negative effects of the unreliable priors to the motion graph. We observe that SMG is good at modeling the dynamic objects under the scenes with a large baseline, e.g. "Apple". SMG better preserves the objects' physical shapes. This also comes from the confidence-aware uncertainty modeling. Reliable SMG nodes provide stable guidance for unreliable nodes, helping preserve object shape through occlusions and improving temporal consistency.

Method	PSNR	LPIPS	Method	PSNR	LPIPS
D-NeRF [49]	21.49	0.232	CTNeRF [42]	26.13	0.082
NR-NeRF [59]	19.69	0.323	DynPoint [76]	26.53	0.068
TiNeuVox [15]	19.74	0.285	D-NPC [22]	25.64	0.109
HyperNeRF [48]	17.60	0.367	RoDynRF [39]	25.89	0.067
NSFF [35]	24.33	0.199	RoDynRF [39] w/o	25.38	0.079
DynNeRF [19]	26.10	0.082	GaussianMarbles [56]	22.32	0.129
MonoNeRF [58]	25.62	0.106	MoSca [30]	26.72	0.070
4DGS [63]	21.45	0.199	**Ours**	26.87	0.068
Casual-FVS [29]	24.57	0.081			

Table 2: Novel-view synthesis results on the NVIDIA Dataset [69].

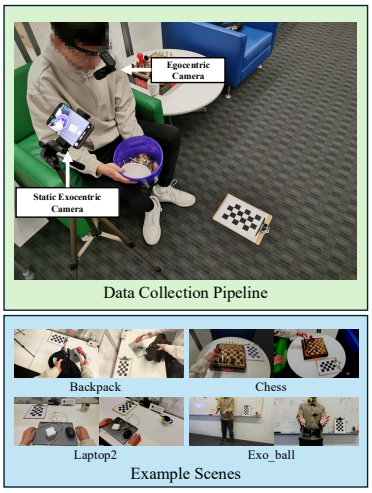
4.3 Evaluations on NVIDIA Dataset

We conduct experiments on the NVIDIA dataset for novel-view synthesis following MoSca's [30] protocol (Table 2). NVIDIA dataset is a real-world video dataset captured by front-facing cameras. Due to the narrow-baseline, the viewpoints of the training and testing frames are very similar. Most motions are fully observed and temporally stable, leaving few under-constrained regions where semantic local motion guidance plays a significant role. Nevertheless, SMG achieves comparable results to state-of-the-art methods. SMG also achieves better motion propagation and preserves cleaner geometry. As shown in Fig. 4 (the human face in "Balloon2"), the proposed semantic motion propagation method effectively prevents motion leakage and contributes to cleaner geometry even in well-constrained scenes.

4.4 Evaluations on SMG Dataset

While Dycheck provides multi-view videos with diverse real-world motion, most dynamics are dominated by simple object rotations (e.g., rotating humans, ap-

Fig. 4: Qualitative results on the NVIDIA dataset.



Scene	Method	PSNR↑	SSIM↑	LPIPS↓
backpack	MoSca [30]	10.01	0.532	0.466
	OriGS [64]	10.22	0.529	0.479
	SMG	**10.38**	**0.542**	**0.454**
chess	MoSca [30]	12.43	0.410	0.439
	OriGS [64]	12.89	0.412	0.460
	SMG	**13.02**	**0.445**	**0.410**
laptop2	MoSca [30]	13.69	0.576	0.409
	OriGS [64]	13.85	0.570	0.430
	SMG	**14.27**	**0.581**	**0.374**
exo_ball	MoSca [30]	17.10	0.652	0.366
	OriGS [64]	17.13	0.660	0.357
	SMG	**17.27**	**0.668**	**0.348**
All	MoSca [30]	13.31	0.542	0.421
	OriGS [64]	13.52	0.543	0.432
	SMG	**13.74**	**0.559**	**0.396**

Fig. 5: Left: Data collection pipeline and example scenes. Right: Quantitative comparison on the proposed SMG dataset.

ples, or steering wheels), and camera motion is relatively mild. As a result, off-the-shelf priors such as depth and tracks can typically be estimated reliably.

To evaluate the robustness of 3D lifting field methods under more challenging conditions, we introduce the **SMG Dataset**, a real-world multi-view video dataset featuring complex camera motion and human–object interactions. Unlike Dycheck, which relies on multiple casually placed cameras, our setup pairs a moving egocentric camera with a static exocentric camera (Fig. 5 (left)). The strong egocentric motion and intricate interactions produce substantially harder geometric conditions. We estimate per-frame camera poses using a PnP solver with calibrated 2D–3D correspondences, and provide metric depth maps predicted by the state-of-the-art metric depth estimator Depth Anything 3 (DA3) [36]. Since egocentric depth predictions are often noisy, we prepare 4 scenes with post-processed DA3 depth for both quantitative and qualitative evaluation, while the remaining 16 scenes are used for qualitative evaluation.

We report the quantitative comparison between state-of-the-art methods [30, 64] and SMG in Fig. 5 (right). SMG consistently outperforms MoSca and OriGS across all four scenes. The performance gap is smaller in the "exo_ball" scene,

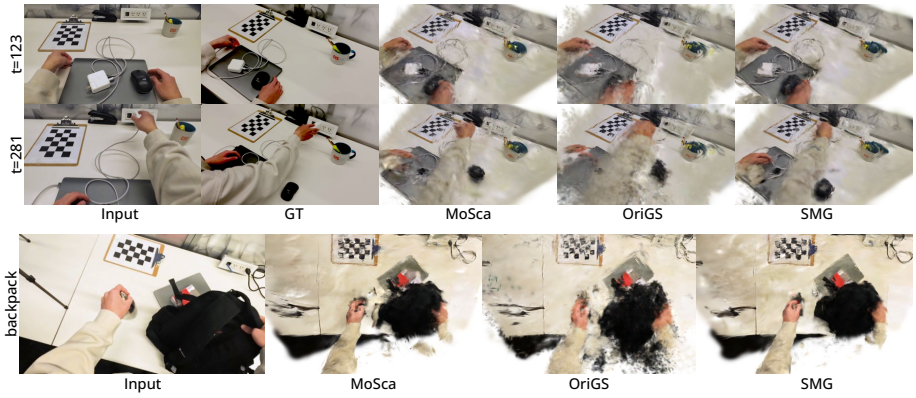


Fig. 6: Top: Qualitative comparison on the testing camera of the "laptop2" scene. Bottom: Novel-view visualization near the input cameras of the "Backpack" scene.

where the camera baseline is small and the egocentric camera remains largely front-facing; under these conditions, both the observations and off-the-shelf priors are relatively reliable. In contrast, for scenes with large viewpoint changes and heavy occlusions, SMG outperforms MoSca and OriGS by a larger margin.

Qualitative results are shown in Fig. 6. Due to the noisy predicted depth and tracks, novel-view synthesis becomes extremely challenging for dynamic Gaussian methods. We show qualitative comparisons on the test view of the "Laptop2" scene (top) and on a nearby view of the input camera of the "Backpack" scene (bottom). We highlight two representative failure cases of the previous methods. First, note the geometry and position of the mouse and the left hand over time, from frame 123 to 281 of "Laptop2". Previous methods incorrectly propagate the motion of the right arm to the static mouse, which is not visible in the training view, leading to clear drift in the rendered novel views. In contrast, SMG restricts motion propagation within local semantic groups, preventing motion leakage across unrelated objects and preserving accurate mouse geometry and position. The left hand in "Laptop2" and the laptop and backpack in "Backpack" illustrate a second failure mode. Without explicitly modeling uncertainty during optimization, the previous methods become under-constrained when observations are limited, causing the object geometry to collapse. With the proposed C-ARAP and LRM, SMG maintains the correct object geometry and position even under limited observations and recovers plausible geometry once observations become available.

Method	EPE ↓	$\delta^{3D}_{.05}$ ↑	$\delta^{3D}_{.10}$ ↑
MoSca [30]	0.055	73.1	89.6
OriGS [64]	0.057	71.9	89.7
SMG	**0.052**	**74.1**	**91.6**

Table 3: 3D tracking results on Dycheck dataset.

Fig. 7: Qualitative results of 3D tracking on Dycheck dataset. (Zoom in to see the blue, purple and pink tracks.)

4.5 Evaluations on 3D Tracking

Since SMG is initialized with the lifted 2D tracks, it naturally supports long-range 3D point tracking. We further evaluate this capability on the five scenes (Apple, Block, Spin, Teddy, and Paper-windmill) of the Dycheck dataset. We report the 3D end-point-error (EPE) and the percentage of points that fall within a given threshold of the ground truth 3D location $\delta^{3D}_{.05}$ $\delta^{3D}_{.10}$, where the threshold is 5cm and 10cm. SMG successfully models these local deformations, resulting in more accurate and semantically consistent 3D motion tracks. As shown in Table 3, SMG outperforms MoSca and OriGS on all metrics. We further show the qualitative results on the "Paper-windmill" scene in the Dycheck dataset (Fig. 7). With the proposed C-ARAP and LRM, the semantic motion graph propagates correct motion across the frames.

4.6 Ablation Study

We conduct ablations to evaluate the impact of each component in our pipeline. The quantitative experiments (Table 4) are conducted on 4 Dycheck scenes (Apple, Spin, Space-out, Wheel). “Base” denotes a vanilla dynamic Gaussian Splatting [63]. “SMG-only” denotes the vanilla SMG without C-ARAP and LRM.

Compared to vanilla dynamic Gaussian Splatting, SMG-only already improves novel-view synthesis by regularizing Gaussian deformation with semantic motion structure. This is also reflected in qualitative ablation in Fig. 8. Without motion regularization, the vanilla 4DGS collapses under novel-view rendering, whereas SMG produces more plausible dynamic reconstructions. However, SMG alone does not reliably preserve topology during motion propagation, which limits its performance. Adding C-ARAP further improves the results. As shown in Table 4, removing C-ARAP leads to a clear performance drop. The qualitative results explain this degradation. Without C-ARAP, the motion graph is more likely to suffer from topology drift, resulting in unstable Gaussian motion and degraded reconstructions. By enforcing locally consistent motion propagation under uncertainty, C-ARAP substantially improves novel-view stability. LRM alone provides only limited improvement when directly applied to SMG. This is because LRM cannot correct topology drift by itself introduced by the uncertainty during optimization. Instead, LRM is most effective when combined with

C-ARAP. Without LRM, C-ARAP may still be affected by the optimization uncertainty during the optimization, which can occasionally produce floating Gaussian artifacts, as shown in the right-hand region of Fig 8.

The full model combines the semantic motion graph with both uncertainty modeling components C-ARAP and LRM. It achieves the best quantitative performance and the most consistent qualitative reconstructions across frames, demonstrating the complementary benefits of semantic motion modeling and uncertainty-aware regularization.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Base [63]	12.92	0.451	0.598
SMG-only	19.07	0.700	0.280
w/o C-ARAP	19.09	0.703	0.281
w/o LRM	19.96	0.731	0.248
Full SMG	**20.11**	**0.738**	**0.242**

Table 4: Ablations for evaluating the contribution of each pipeline component.

Fig. 8: Qualitative ablation on the effectiveness of each component in the pipeline.

5 Discussion

Limitations. Though SMG achieves state-of-the-art performance on novel-view synthesis benchmarks, a few challenges remain unresolved. Our method relies on off-the-shelf priors for initialization. Inaccuracies in these components may affect the performance of our model. SMG naturally improves as these upstream predictions become more accurate. Our method relies on per-scene optimization, which limits its scalability to large-scale or real-time applications. Recent advances [65] for feed-forward 4D Gaussian Splatting offer a promising direction toward more efficient and flexible dynamic reconstruction.

In summary, we present SMG, a novel monocular dynamic Gaussian framework that models Gaussian motion as structured semantic motion. To address uncertainty during Gaussian optimization, we introduce dynamic uncertainty modeling components, C-ARAP and LRM, which regularize motion in unconstrained regions using reliable graph nodes and preserve the dynamic object geometry. We further provide a multiview dataset with complex camera and scene motion that evaluates dynamic Gaussian Splatting under challenging real-world conditions. Our approach outperforms state-of-the-art methods for dynamic novel-view synthesis on real-world benchmarks and highlights the potential of semantic motion modeling for dynamic scene reconstruction.

References

1. Achlioptas, P., Abdelreheem, A., Xia, F., Elhoseiny, M., Guibas, L.: Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16. pp. 422–440. Springer (2020)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 5855–5864 (2021)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5470–5479 (2022)
4. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Zip-nerf: Anti-aliased grid-based neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19697–19705 (2023)
5. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 130–141 (2023)
6. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
7. Cen, J., Fang, J., Yang, C., Xie, L., Zhang, X., Shen, W., Tian, Q.: Segment any 3d gaussians. arXiv preprint arXiv:2312.00860 (2023)
8. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: European conference on computer vision. pp. 333–350. Springer (2022)
9. Chen, D.Z., Chang, A.X., Nießner, M.: Scanrefer: 3d object localization in rgb-d scans using natural language. In: European conference on computer vision. pp. 202–221. Springer (2020)
10. Chen, K., Zhong, Y., Li, Z., Lin, J., Chen, Y., Qin, M., Wang, H.: Quantifying and alleviating co-adaptation in sparse-view 3d gaussian splatting. arXiv preprint arXiv:2508.12720 (2025)
11. Chen, S., Guhur, P.L., Tapaswi, M., Schmid, C., Laptev, I.: Language conditioned spatial relation reasoning for 3d object grounding. Advances in neural information processing systems **35**, 20522–20535 (2022)
12. Chen, S., Zhu, H., Chen, X., Lei, Y., Yu, G., Chen, T.: End-to-end 3d dense captioning with vote2cap-detr. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11124–11133 (2023)
13. Chen, Z., Gholami, A., Nießner, M., Chang, A.X.: Scan2cap: Context-aware dense captioning in rgb-d scans. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3193–3203 (2021)
14. Du, Y., Zhang, Y., Yu, H.X., Tenenbaum, J.B., Wu, J.: Neural radiance flow for 4d view synthesis and video processing. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV). pp. 14304–14314 (2021)
15. 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)
16. Fiebelman, G., Cohen, T., Morgenstern, A., Hedman, P., Averbuch-Elor, H.: 4-legs: 4d language embedded gaussian splatting. In: Computer Graphics Forum. p. e70085. Wiley Online Library (2025)

17. 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)
18. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5501–5510 (2022)
19. 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)
20. 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)
21. Huang, S., Chen, Y., Jia, J., Wang, L.: Multi-view transformer for 3d visual grounding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15524–15533 (2022)
22. Kappel, M., Hahlbohm, F., Scholz, T., Castillo, S., Theobalt, C., Eisemann, M., Golyanik, V., Magnor, M.: D-npc: Dynamic neural point clouds for non-rigid view synthesis from monocular video. In: Computer Graphics Forum. p. e70038. Wiley Online Library (2025)
23. Kavan, L., Collins, S., Žára, J., O'Sullivan, C.: Skinning with dual quaternions. In: Proceedings of the 2007 symposium on Interactive 3D graphics and games. pp. 39–46 (2007)
24. 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)
25. Kerr, J., Kim, C.M., Goldberg, K., Kanazawa, A., Tancik, M.: Lerf: Language embedded radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 19729–19739 (2023)
26. Kim, C.M., Wu, M., Kerr, J., Goldberg, K., Tancik, M., Kanazawa, A.: Garfield: Group anything with radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21530–21539 (2024)
27. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 4015–4026 (2023)
28. Labe, I., Issachar, N., Lang, I., Benaim, S.: Dgd: Dynamic 3d gaussians distillation. In: European Conference on Computer Vision. pp. 361–378. Springer (2024)
29. Lee, Y.C., Zhang, Z., Blackburn-Matzen, K., Niklaus, S., Zhang, J., Huang, J.B., Liu, F.: Fast view synthesis of casual videos with soup-of-planes. In: European Conference on Computer Vision. pp. 278–296. Springer (2024)
30. 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)
31. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. arXiv preprint arXiv:2201.03546 (2022)
32. Li, J., Zhang, J., Bai, X., Zheng, J., Ning, X., Zhou, J., Gu, L.: Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 20775–20785 (2024)
33. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Newcombe, R., et al.: Neural 3d video synthesis

from multi-view video. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5521–5531 (2022)
34. Li, W., Zhou, R., Zhou, J., Song, Y., Herter, J., Qin, M., Huang, G., Pfister, H.: 4d langsplat: 4d language gaussian splatting via multimodal large language models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22001–22011 (2025)
35. Li, Z., Niklaus, S., Snavely, N., Wang, O.: 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. pp. 6498–6508 (2021)
36. Lin, H., Chen, S., Liew, J., Chen, D.Y., Li, Z., Shi, G., Feng, J., Kang, B.: Depth anything 3: Recovering the visual space from any views. arXiv preprint arXiv:2511.10647 (2025)
37. Liu, K., Zhan, F., Zhang, J., Xu, M., Yu, Y., El Saddik, A., Theobalt, C., Xing, E., Lu, S.: Weakly supervised 3d open-vocabulary segmentation. Advances in Neural Information Processing Systems **36**, 53433–53456 (2023)
38. Liu, Q., Liu, Y., Wang, J., Lyu, X., Wang, P., Wang, W., Hou, J.: Modgs: Dynamic gaussian splatting from casually-captured monocular videos with depth priors. In: International Conference on Learning Representations. vol. 2025, pp. 97048–97074 (2025)
39. Liu, Y.L., Gao, C., Meuleman, A., Tseng, H.Y., Saraf, A., Kim, C., Chuang, Y.Y., Kopf, J., Huang, J.B.: Robust dynamic radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13–23 (2023)
40. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. arXiv preprint arXiv:2308.09713 (2023)
41. Ma, X., Yong, S., Zheng, Z., Li, Q., Liang, Y., Zhu, S.C., Huang, S.: Sqa3d: Situated question answering in 3d scenes. arXiv preprint arXiv:2210.07474 (2022)
42. Miao, X., Bai, Y., Duan, H., Wan, F., Huang, Y., Long, Y., Zheng, Y.: Ctnerf: Cross-time transformer for dynamic neural radiance field from monocular video. Pattern Recognition **156**, 110729 (2024)
43. 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)
44. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG) **41**(4), 1–15 (2022)
45. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: Dinov2: Learning robust visual features without supervision. arXiv preprint arXiv:2304.07193 (2023)
46. Park, J., Bui, M.Q.V., Bello, J.L.G., Moon, J., Oh, J., Kim, M.: Splinegs: Robust motion-adaptive spline for real-time dynamic 3d gaussians from monocular video. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 26866–26875 (2025)
47. 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)
48. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. ACM Trans. Graph. **40**(6) (dec 2021)

49. 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)
50. Qin, M., Li, W., Zhou, J., Wang, H., Pfister, H.: Langsplat: 3d language gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20051–20060 (2024)
51. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
52. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: Sam 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
53. Seidenschwarz, J., Zhou, Q., Duisterhof, B.P., Ramanan, D., Leal-Taixé, L.: Dynomo: Online point tracking by dynamic online monocular gaussian reconstruction. In: 2025 International Conference on 3D Vision (3DV). pp. 1012–1021. IEEE (2025)
54. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., et al.: Dinov3. arXiv preprint arXiv:2508.10104 (2025)
55. Song, R., Liang, C., Xia, Y., Zimmer, W., Cao, H., Caesar, H., Festag, A., Knoll, A.: Coda-4dgs: Dynamic gaussian splatting with context and deformation awareness for autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 28031–28041 (2025)
56. 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)
57. Sun, H., Li, X., Shen, L., Ye, X., Xian, K., Cao, Z.: Dyblurf: Dynamic neural radiance fields from blurry monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7517–7527 (2024)
58. 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)
59. Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Lassner, C., Theobalt, C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12959–12970 (2021)
60. Wang, C., Zhuang, P., Siarohin, A., Cao, J., Qian, G., Lee, H.Y., Tulyakov, S.: Diffusion priors for dynamic view synthesis from monocular videos. arXiv preprint arXiv:2401.05583 (2024)
61. 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)
62. 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)
63. 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 (CVPR) (2024)

64. Wu, J., Tao, J., Wang, H., Liu, G., Kompella, R.R., Yan, Y.: Orientation-anchored hyper-gaussian for 4d reconstruction from casual videos. arXiv preprint arXiv:2509.23492 (2025)
65. Xu, Z., Li, Z., Dong, Z., Zhou, X., Newcombe, R., Lv, Z.: 4dgt: Learning a 4d gaussian transformer using real-world monocular videos. arXiv preprint arXiv:2506.08015 (2025)
66. Yang, Z., Yang, H., Pan, Z., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In: International Conference on Learning Representations (ICLR) (2024)
67. Yang, Z., Zhang, S., Wang, L., Luo, J.: Sat: 2d semantics assisted training for 3d visual grounding. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1856–1866 (2021)
68. 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)
69. Yoon, J.S., Kim, K., Gallo, O., Park, H.S., Kautz, J.: Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5336–5345 (2020)
70. Zhang, J., Li, J., Yu, X., Huang, L., Gu, L., Zheng, J., Bai, X.: Cor-gs: sparse-view 3d gaussian splatting via co-regularization. In: European Conference on Computer Vision. pp. 335–352. Springer (2024)
71. Zhang, K., Riegler, G., Snavely, N., Koltun, V.: Nerf++: Analyzing and improving neural radiance fields. arXiv preprint arXiv:2010.07492 (2020)
72. 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)
73. Zhao, X., Colburn, A., Ma, F., Bautista, M.A., Susskind, J.M., Schwing, A.G.: Pseudo-generalized dynamic view synthesis from a video. arXiv preprint arXiv:2310.08587 (2023)
74. Zheng, D., Huang, S., Wang, L.: Video-3d llm: Learning position-aware video representation for 3d scene understanding. arXiv preprint arXiv:2412.00493 (2024)
75. Zhi, S., Laidlow, T., Leutenegger, S., Davison, A.J.: In-place scene labelling and understanding with implicit scene representation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15838–15847 (2021)
76. Zhou, K., Zhong, J.X., Shin, S., Lu, K., Yang, Y., Markham, A., Trigoni, N.: Dynpoint: Dynamic neural point for view synthesis. Advances in Neural Information Processing Systems **36**, 69532–69545 (2023)
77. Zhou, S., Chang, H., Jiang, S., Fan, Z., Zhu, Z., Xu, D., Chari, P., You, S., Wang, Z., Kadambi, A.: Feature 3dgs: Supercharging 3d gaussian splatting to enable distilled feature fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21676–21685 (2024)
78. Zhou, S., Ren, H., Weng, Y., Zhang, S., Wang, Z., Xu, D., Fan, Z., You, S., Wang, Z., Guibas, L., et al.: Feature4x: Bridging any monocular video to 4d agentic ai with versatile gaussian feature fields. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 14179–14190 (2025)
79. Zhu, Z., Fan, Z., Jiang, Y., Wang, Z.: Fsgs: Real-time few-shot view synthesis using gaussian splatting. In: European conference on computer vision. pp. 145–163. Springer (2024)