4DGS360: 360° Gaussian Reconstruction of Dynamic Objects from a Single Video

Jae Won Jang[1], Yeonjin Chang[1], Wonsik Shin[1], Juhwan Cho[1], and Nojun Kwak[1] *

Seoul National University, Seoul, South Korea
{pert0407, yjean8315, wonsikshin, hj99cho, nojunk}@snu.ac.kr

Abstract. We introduce 4DGS360, a diffusion-free framework for 360° dynamic object reconstruction from casual monocular video. Existing methods often fail to reconstruct consistent 360° geometry, as their heavy reliance on 2D-native priors causes initial points to overfit to visible surface in each training view. 4DGS360 addresses this challenge through an advanced 3D-native initialization that mitigates the geometric ambiguity of occluded regions. Our proposed 3D tracker, AnchorTAP3D, produces reinforced 3D point trajectories by leveraging confident 2D track points as anchors, suppressing drift and providing reliable initialization that preserves geometry in occluded regions. This initialization, combined with optimization, yields coherent 360° 4D reconstructions. We further present iPhone360, a new benchmark where test cameras are placed up to 135° apart from training views, enabling 360° evaluation that existing datasets cannot provide. Experiments show that 4DGS360 achieves state-of-the-art performance on the iPhone360, iPhone, and DAVIS datasets, both qualitatively and quantitatively. Project website at https://jaewon040.github.io/4dgs360/

Keywords: Monocular dynamic novel view synthesis · 360° 4D reconstruction · Monocular dynamic scene dataset

1 Introduction

We present 4DGS360, a diffusion-free framework for 360° dynamic object reconstruction from casual monocular video. By leveraging 3D-native occlusion-aware initialization, our method ensures faithful 3D reconstruction even at extreme novel viewpoints.

Dynamic 3D reconstruction has long been a central area of interest in computer vision. Beyond static 3D Gaussian Splatting [16], 4D (3D + time) reconstruction is increasingly demanded in practical applications such as video content creation, spatial computing for VR/MR/AR, and 3D holographic media. In particular, reconstructing from a casual monocular video, a setting that reflects real-world capture conditions, has garnered significant attention [23,55].

* Corresponding author.

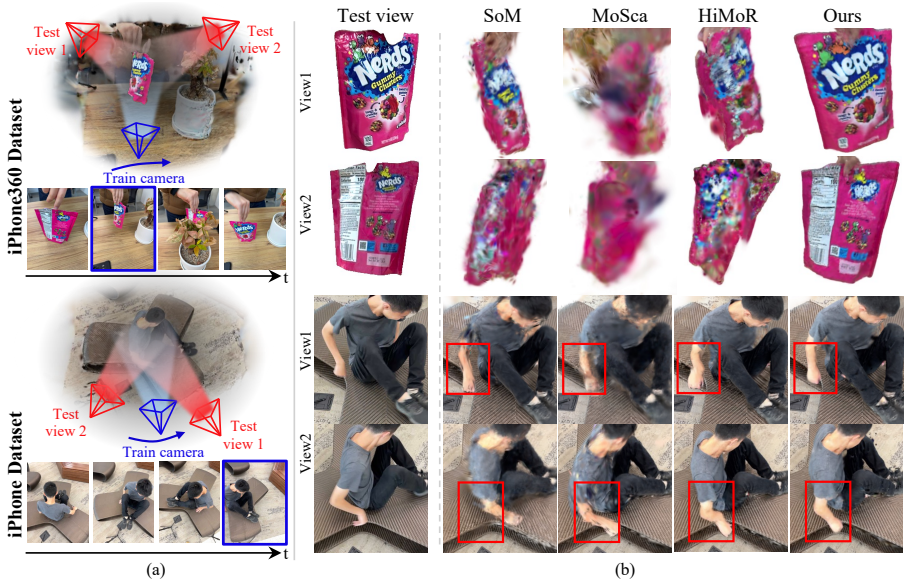

Fig. 1: Our 4DGS360 model reconstructs 360° dynamic object geometry from a monocular video and produces higher-fidelity renderings than prior methods under both ordinary and extreme novel-view conditions. We also introduce the iPhone360 dataset, which captures dynamic objects in a monocular setting and includes test cameras far outside the training trajectory for 360° evaluation.

4D reconstruction under a monocular setting is a highly ill-posed problem [11, 22, 28], as only a single viewpoint is available per frame, with no multi-view stereo cues available. Recent works [25, 49] have addressed this challenge by leveraging pretrained 2D point tracking models [8], which establish cross-frame correspondences in dynamic videos.

However, as shown in Fig. 1, existing methods still fail to reconstruct regions observed at extreme novel viewpoints (e.g., $> 90°$ from the current view), even when those regions are visible in other frames of the input video. We argue that this failure stems from a heavy reliance on 2D-native priors [8, 9] during initialization. Existing methods [23, 25] initialize 3D dynamic trajectories using 2D tracking results, which offer more reliable performance than 3D tracking models [58, 65]. These 2D tracks must be lifted into 3D to form initial gaussians. However, depth maps used in lifting only provide depth for surfaces visible in the current frame. As a result, the 3D positions of occluded track points remain ambiguous, causing the initialized geometry to overfit to the visible surfaces at each timestep. Since this incomplete geometry is not recovered during optimization, reconstruction of occluded regions fails, even when they are visible in other frames.

We propose 4DGS360 to address the longstanding challenge of complete 360° dynamic object reconstruction by introducing AnchorTAP3D, a novel 3D native

tracker that overcomes the limitations of both 2D and 3D tracking models effectively. AnchorTAP3D leverages confident 2D track points as anchors for 3D tracking, improving long-term reliability and resolving the depth ambiguity of occluded regions without additional training. AnchorTAP3D provides a reliable initialization that unlocks the full potential of the optimization strategy used in prior works. In particular, the As-Rigid-As-Possible (ARAP) regularization, which previously failed to operate correctly on corrupted initialized geometry in occluded regions, can now function as intended, enabling significantly improved 360° reconstructions. As shown in Fig. 1(b), our method establishes more stable and coherent 4D reconstructions of dynamic objects, even under large view extrapolation.

However, existing monocular datasets [11, 34, 35] for dynamic scene reconstruction are limited to evaluate reconstruction quality at truly novel viewpoints, as they lack sufficient view disparity between train and test cameras. Even the iPhone dataset [11], which provides the largest such disparity, remains insufficient to properly evaluate 360° reconstruction. To address this, we introduce iPhone360, a new dataset designed to evaluate 360° reconstruction of dynamic objects under realistic conditions with large disparity between train and test camera views. Our iPhone360 dataset includes training videos of real-world dynamic objects captured with casual handheld movements, exhibiting diverse motion ranging from object manipulation to human motion. Fig. 1(a) demonstrates how our iPhone360 dataset provides significantly larger view disparity compared to the iPhone dataset, enabling 360° evaluation.
Overall, we provide the following contributions:

- We propose **4DGS360**, which employs a novel 3D-native initialization method based on *AnchorTAP3D*, enabling consistent 360° dynamic reconstruction from monocular video.
- We present **iPhone360 dataset**, a new dataset that captures real-world dynamic objects with various scenarios. Test cameras are placed up to 135° apart from training views enabling evaluation of models under extreme novel-view conditions.
- Our approach achieves state-of-the-art performance both qualitatively and quantitatively under ordinary and extreme novel-view synthesis conditions.

2 Related Work

2.1 Dynamic Novel-View Synthesis

Novel-view synthesis has rapidly advanced with the advent of NeRF [32] and 3D Gaussian Splatting (3DGS) [16]. NeRF represents scenes implicitly using multi-layer perceptrons (MLPs), enabling view-consistent rendering across various reconstruction tasks [1–3, 10, 21, 41, 47, 48, 63]. In contrast, 3DGS models scenes explicitly with 3D Gaussian primitives, enabling real-time rasterization-based rendering through its efficient explicit formulation. Owing to these advantages, the 3DGS framework has been widely extended to diverse applications [4, 6, 7, 13, 27, 29, 31, 33, 42, 54, 64, 67].

Extending these ideas to dynamic scenes, models reconstruct objects and environments that change over time. Recent works [25, 49] model temporal deformation by estimating the trajectories of canonical-space Gaussians. Some methods [56, 62] encode motion implicitly via MLPs, while others represent motion explicitly by assigning trajectories to individual Gaussians [30, 44, 55] or by combining learned bases [49] and hierarchical motion fields [25].

Recent studies further explore the monocular setting, where a single camera captures dynamic scenes under realistic conditions. However, monocular reconstruction remains highly ill-posed and typically relies on pretrained models [12, 14, 15, 19, 46, 52, 59–61] or 2D tracking cues [8, 9, 26, 51] to recover motion on the image plane. While effective for visible regions, these approaches fail to reconstruct occluded geometry, leading to overfitting to the training views. Other works [17, 45, 57] employ large diffusion models [39] to synthesize unseen view. However, these methods require substantial computational cost for training generative model, and often fail to leverage information across video frames. Furthermore, as demonstrated in Fig. 5, they show limited performance gains even when combined with existing state-of-the-art 4D novel view synthesis methods at extreme novel viewpoints.

Our method addresses this limitation by introducing occluded geometry-aware tracking at initialization, enabling consistent reconstruction beyond observed views and serving as a stronger baseline even when combined with diffusion-based approaches.

2.2 Monocular setting datasets

A number of benchmarks have been proposed for dynamic scene reconstruction under the monocular setting. D-NeRF [37] uses synthetic scenes to evaluate temporal deformation modeling, while HyperNeRF [35], Nerfies [34], and the iPhone dataset [11] capture real-world dynamic scenes. These datasets mainly evaluate temporal interpolation or short-range novel view synthesis, where test cameras remain close to the train camera. In particular, Hypernerf and Nerfies datasets are not fully aligned with real-world capture conditions. They alternate between two cameras per frame to construct training sets. The iPhone dataset maintains a more faithful monocular setting, but the gaps between train and test cameras are limited. Therefore, we release a new dataset, **iPhone360**, for comprehensive 360° evaluation with extreme novel view test cameras under realistic monocular capture conditions.

2.3 2D and 3D Point Tracking

2D Tracking. Recent point tracking methods adopt deep learning approaches. Among them, TAPIR [9] improves accuracy through global matching and refinement, while transformer-based models such as CoTracker [14] iteratively infer point position and visibility, achieving robustness under occlusion. Self-supervised approaches like BootsTAP [8] further enhances performance on unlabeled real-world data using pseudo ground truth from pretrained trackers.

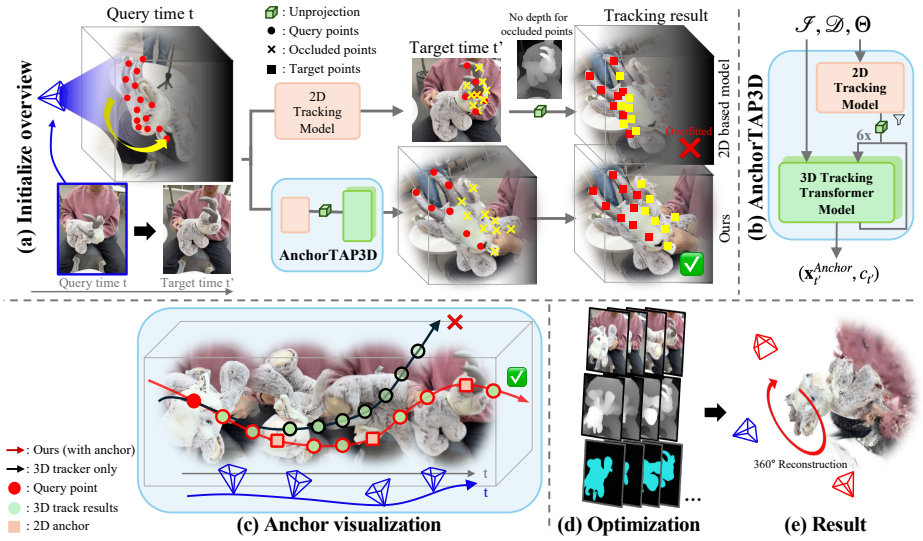

Fig. 2: Overview of 4DGS360 (a) illustrates the initialization stage using AnchorTAP3D, where reliable geometry can be obtained even when tracked points become occluded. This contrasts with 2D-based models, which cannot estimate depth for occluded points. (b) presents the overall architecture of AnchorTAP3D, and (c) highlights the difference between tracking with and without anchors, showing that anchors effectively prevent error drift. (d) depicts the training process that follows initialization, and (e) shows that our model finally enables full 360° reconstruction. The blue pyramids represent training cameras, while the red pyramids indicate test cameras.

3D Tracking. Among 3D tracking models [14, 15], TAPIP3D [65] achieves notable tracking results by employing a transformer-based architecture that performs tracking directly in XYZ space, leveraging unprojected image features to construct a spatio-temporal 3D feature cloud. It models temporal correspondence through 3D neighborhood-to-neighborhood attention, where local geometric features near query and target points are jointly attended to infer motion trajectories. This design enables TAPIP3D to capture smooth and continuous motion across frames and handle moderate occlusions without explicit depth supervision. While 3D tracking provides more geometric understanding than 2D tracking models, it is sensitive to errors in depth maps and camera calibration, and tracking failures tend to accumulate over time, reducing long-term stability. On the other hand, 2D tracking is generally more resilient to noise in real-world settings but provides limited spatial understanding. To address this, we propose **AnchorTAP3D**, which leverages high confidence 2D track points as anchors for 3D tracking, enabling consistent geometry-aware initialization for our model.

3 Method

Our goal is to reconstruct a 360° 4D representation over time, given a sequence of T training frames $\mathcal{I} = \{I_t\}_{t=1}^{T}$, depth maps $\mathcal{D} = \{D_t\}_{t=1}^{T}$, and camera parameters $\Theta = \{\theta_t\}_{t=1}^{T}$ obtained either from dataset sensors or pretrained estimators. Our method represents dynamic objects using a set of static 3D Gaussians defined in a canonical space, which are deformed over time according to a hierarchical motion structure. Sec. 3.1 describes the preliminaries and details our 4D scene representation. Sec. 3.2 introduces our occluded geometry–aware initialization using the proposed tracking model, AnchorTAP3D. Finally, Sec. 3.3 presents the optimization strategy that captures dynamic motion while enforcing local rigidity and visual consistency. Fig. 2 illustrates the overview of our method.

3.1 Preliminary: Dynamic Gaussian Splatting

A 3D scene is represented by a set of anisotropic Gaussian primitives $G(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \alpha, \mathbf{c})$, where each Gaussian is parameterized by its mean $\boldsymbol{\mu} \in \mathbb{R}^3$, covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times3}$, opacity $\alpha \in \mathbb{R}^+$, and view-dependent color coefficients $\mathbf{c} \in \mathbb{R}^{3(l+1)^2}$ from spherical harmonics (SH). This explicit and differentiable formulation enables photo-realistic scene representation and optimization in 3D space.

To handle dynamics, each static Gaussian in a canonical frame is deformed over time through $\mathbf{T}_t = [\mathbf{R}_t|\mathbf{t}_t] \in \mathrm{SE}(3)$, resulting in the time-dependent primitive

$$G_t = G(\mathbf{R}_t\boldsymbol{\mu}_0 + \mathbf{t}_t,\ \mathbf{R}_t\boldsymbol{\Sigma}_0\mathbf{R}_t^\top,\ \alpha,\ \mathbf{c}), \tag{1}$$

where $(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ denote the mean and covariance at canonical space.

We model $\mathbf{T}_t$ following the HiMoR [25], which encodes the deformation field as a tree of nodes. The deformation of a canonical Gaussian is guided by its nearby leaf nodes. Specifically, the transformation $\mathcal{T} = \{\,\mathbf{T}_t\,\}_{t=1}^{T}$ of a Gaussian G is obtained by interpolating the motions of its K nearest leaf nodes $\mathcal{V}$:

$$\mathbf{T}_t = \sum_{k\in\mathcal{N}(G,\mathcal{V})} w_k\,\mathcal{M}_k, \tag{2}$$

where $\mathcal{M} = \{\,\mathbf{M}_t\,\}_{t=1}^{T}$ for $\mathbf{M}_t \in \mathrm{SE}(3)$ denotes the motion of nodes, and w_k are Gaussian-specific interpolation weights.

Each node's motion $\mathbf{M}_t$ represents its transformation from the canonical frame to frame t, and is defined hierarchically with respect to its parent. To efficiently encode these motions, HiMoR represents each node's transformation as a weighted combination of its parent's shared motion bases:

$$\mathbf{M}_t = \sum_{m=1}^{M} v_m\,\mathcal{B}_{m,t}, \tag{3}$$

where $\{\mathcal{B}_{m,t}\}$ denote the motion bases shared among sibling nodes, and v_m are node-specific coefficients. This hierarchical formulation allows higher-level nodes

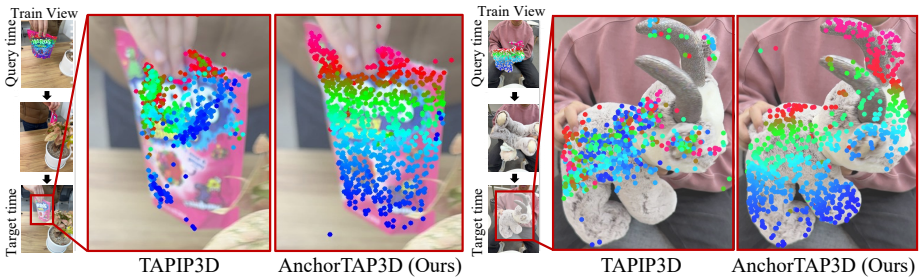

Fig. 3: Tracking Comparison. We compare the naive 3D tracking model TAPIP3D [65] with our anchor-based AnchorTAP3D (Ours). At the target time after undergoing rotation and occlusion, AnchorTAP3D produces noticeably more stable tracking results. This indicates that anchor-guided tracking is advantageous under challenging conditions such as occlusion.

to capture global, smooth motion patterns, while deeper nodes refine fine-grained local deformation across time and space.

Given camera parameters $\boldsymbol{\theta}_t$, the deformed canonical Gaussian G_t at time t is projected onto the image plane via

$$\boldsymbol{\mu}'_t = \Pi(\mathbf{R}_t\boldsymbol{\mu}_0 + \mathbf{t}_t; \boldsymbol{\theta}_t), \quad \boldsymbol{\Sigma}'_t = \Pi(\mathbf{R}_t\boldsymbol{\Sigma}_0\mathbf{R}_t^\top; \boldsymbol{\theta}_t), \tag{4}$$

where Π denotes the camera projection. After projection, each 3D Gaussian G_t becomes a 2D Gaussian on the image plane, parameterized by the mean $\boldsymbol{\mu}'_t \in \mathbb{R}^2$ and covariance $\boldsymbol{\Sigma}'_t \in \mathbb{R}^{2\times2}$.

The rendered color at pixel $\mathbf{p} \in \mathbb{R}^2$ is then computed by aggregating these 2D Gaussians using depth-sorted alpha blending:

$$\begin{aligned} C_t(\mathbf{p}) &= \sum_{i=1}^{N} \mathbf{c}_i\, \sigma_{i,t}(\mathbf{p}) \prod_{j=1}^{i-1} (1 - \sigma_{j,t}(\mathbf{p}))\,, \\ \sigma_{i,t}(\mathbf{p}) &= \alpha_i \exp\!\big(-\tfrac{1}{2}\,(\mathbf{p} - \boldsymbol{\mu}'_{i,t})^\top (\boldsymbol{\Sigma}'_{i,t})^{-1} (\mathbf{p} - \boldsymbol{\mu}'_{i,t})\big)\,, \end{aligned} \tag{5}$$

where $\sigma_{i,t}(\mathbf{p})$ denotes the 2D Gaussian opacity at pixel $\mathbf{p}$, and $\mathbf{c}_i$ is its view-dependent color. This dynamic Gaussian representation enables a differentiable rasterization framework.

3.2 Initialization Method

AnchorTAP3D. We introduce **AnchorTAP3D** (**Anchor**-guided **T**racking **A**ny **P**oint in **3D**), an advanced 3D tracking model designed to enhance dynamic 360° reconstruction under extreme novel views without additional training.

Given a sequence of frames $\mathcal{I}$, depth maps $\mathcal{D}$, and camera parameters Θ, AnchorTAP3D estimates the 3D location $\mathbf{x}_{t'}^{Anchor} \in \mathbb{R}^3$ of a 2D anchor point

$\mathbf{p}_t \in \mathbb{R}^2$ observed at time t, after it has moved to time t', along with its binary visibility $v_{t'}$:

$$(\mathbf{x}_{t'}^{Anchor}, v_{t'}) = \mathcal{P}_{t\to t'}^{Anchor3D}(\mathbf{p}_t, \mathcal{I}, \mathcal{D}, \Theta), \tag{6}$$

where $\mathcal{P}^{Anchor3D}$ denotes our unified anchor-guided model which integrates the strengths of a 2D tracking model and a 3D tracking model into a unified framework. The overall formulation can be decomposed into several components, which we describe below.

The 2D tracker, $\mathcal{P}^{2D}$, takes a query point $\mathbf{p}_t = (x_t, y_t)$ at time t and predicts its correspondence in the target frame t' as:

$$(\mathbf{p}_{t'}, c_{t'}) = \mathcal{P}_{t\to t'}^{2D}(\mathbf{p}_t, \mathcal{I}), \tag{7}$$

where $c_{t'}$ denotes the tracking confidence estimated by the 2D tracker.

The predicted 2D correspondence is then lifted to 3D space through inverse camera projection using the depth map and camera parameters at frame t':

$$\mathbf{x}_{t'}^{2D} = \mathit{\Pi}_{t'}^{-1}(\mathbf{p}_{t'}, D_{t'}, \boldsymbol{\theta}_{t'}). \tag{8}$$

This process generates candidate 3D points corresponding to each tracked 2D point. However, points with low 2D confidence, mostly occluded, cannot obtain reliable 3D positions from the depth map of the target frame.

To exclude unreliable correspondences, we define a binary mask based on the 2D confidence score:

$$\text{Mask}_{t'} = \begin{cases} 1, & c_{t'} > \tau, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Only points satisfying $c_{t'} > \tau$ are regarded as trustworthy and are used to form anchor points for subsequent 3D inference.

Within a sliding temporal window $w(t, t')$ of fixed length L, the transformer-based 3D tracker jointly processes all frames in the window (e.g., $L = 16$, with 8-frame overlap between successive windows). During each inference step, we collect reliable 3D anchor points obtained from high-confidence 2D tracks as:

$$\mathcal{X}_{w(t,t')}^{A} \;=\; \left\{ \mathbf{x}_{s'}^{2D} \;\middle|\; s' \in w(t,t'),\ \text{Mask}_{s'} = 1 \right\}. \tag{10}$$

Here, $\mathcal{X}_{w(t,t')}^{A}$ denotes the set of 3D anchor points that condition the transformer during the current temporal inference window, providing geometry-consistent supervision across frames (See Fig. 2(c)).

Finally, the anchor-guided 3D tracker predicts the target 3D point and its binary visibility score by conditioning on the anchor set:

$$(\mathbf{x}_{t'}^{Anchor}, v_{t'}) = \mathcal{P}_{t\to t'}^{3D}(\mathbf{p}_t, \mathcal{I}, \mathcal{D}, \Theta, \mathcal{X}_{w(t,t')}^{A}). \tag{11}$$

Unlike conventional 3D trackers that propagate a single query over time, AnchorTAP3D leverages multiple anchors as spatial-temporal constraints, suppressing error drift and enhancing geometric stability. The anchors are relatively unaffected by depth or calibration noise during tracking, providing stable initialization for our model. Fig. 3 illustrates that our model maintains long-term

tracking more effectively than a naive 3D tracking approach. Furthermore, Fig. 8 demonstrates that in the 4D reconstruction task, our method initialized with AnchorTAP3D better reconstructs the geometry of occluded regions compared to baselines using naive 2D or 3D tracking.

Initialization details. We initialize a set of dynamic Gaussians from the 3D tracks obtained at T query times. Specifically, N trajectories are randomly sampled to define Gaussian primitives whose positions and orientations vary over time.

The frame with the most visible Gaussians is set as the canonical frame. We then cluster the trajectories according to their temporal velocities using k-means into B groups. Within each cluster, frame-to-frame rigid transformations are estimated via Procrustes alignment [40], and the resulting temporal motions are stored as initial motion bases $\{\mathcal{B}_i\}_{i=1}^{B}$. Each Gaussian's motion is further weighted by its spatial distance to the corresponding cluster center. Previous methods [25, 49] discard invisible points during motion estimation since 2D tracking combined with depth-based unprojection cannot infer 3D positions for occluded points. Our approach leverages AnchorTAP3D to infer plausible 3D positions even for occluded regions. This capability enables motion bases that fully capture global object dynamics and establishes a foundation for complete 360° reconstruction.

For node initialization, we perform weighted random sampling of n_1 Gaussians from the canonical frame. Sampling weights are determined by both motion magnitude and spatial density, encouraging nodes to capture highly dynamic regions while maintaining even spatial coverage. The selected Gaussians initialize the first-level nodes with their positions and motion coefficients, and the higher-level nodes are initialized in the same manner with respect to their parent nodes.

3.3 Optimization Method

To optimize the initialized nodes and Gaussians, we employ rigidity regularization to preserve geometric structure and render-based regularization that compares rendered outputs with ground truth.

Rigidity regularization. To maintain long-term geometric consistency and locally rigid motion across non-adjacent frames, we employ a generalized As-Rigid-As-Possible (ARAP) [43] regularization. For each node pair (i, j) belonging to the same locally rigid cluster, we encourage both the pairwise distances and the relative local transformations to remain consistent between arbitrary frames t and t':

$$\mathcal{L}_{\text{arap}} = w_1 \sum_{(i,j)\in\mathcal{N}} \left| \|\mathbf{x}_i^t - \mathbf{x}_j^t\|_2 - \|\mathbf{x}_i^{t'} - \mathbf{x}_j^{t'}\|_2 \right| + w_2 \sum_{(i,j)\in\mathcal{N}} \left\| \mathbf{T}_j^{t^{-1}}(\mathbf{x}_i^t) - \mathbf{T}_j^{t'^{-1}}(\mathbf{x}_i^{t'}) \right\|_2, \quad (12)$$

where $\mathbf{x}_i^t$ denotes the position of node i at frame t, and $\mathbf{T}_j^t \in SE(3)$ represents the local transformation of node j. This rigidity constraint allows temporally coherent propagation of structural information, enabling 360° geometry even under large motion or occlusion.

Render-based regularization. For render-based regularization, we conduct RGB loss $\mathcal{L}_{\text{rgb}}$, which includes D-SSIM [53] loss and LPIPS [66] loss. We regularize the spatial compactness of the reconstructed object via a mask regularization loss $\mathcal{L}_{\text{mask}}$, enforce geometric alignment through a depth consistency loss $\mathcal{L}_{\text{depth}}$, and improve temporal correspondence using a 2D tracking loss $\mathcal{L}_{\text{2dtrack}}$. Together with the rigidity loss $\mathcal{L}_{\text{arap}}$ in Eq. (12), these objectives contribute to a temporally coherent and geometrically stable reconstruction:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{rgb}}\mathcal{L}_{\text{rgb}} + \lambda_{\text{mask}}\mathcal{L}_{\text{mask}} + \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{2dtrack}}\mathcal{L}_{\text{2dtrack}} + \lambda_{\text{arap}}\mathcal{L}_{\text{arap}}. \tag{13}$$

Further details are included in the supplementary.

4 Experiments

4.1 Datasets and Evaluation

Evaluation metrics. For evaluation, we adopt both pixel-level and perceptual metrics. While pixel-level metrics such as PSNR and SSIM have long been standard metrics in novel view synthesis, recent work [25] has shown that these pixel-level metrics are misaligned with perceptual quality in monocular dynamic 3D reconstruction with view disparity, due to the inherent difficulty of predicting 'exact' position in this ill-posed setting. Therefore, we report both metric types together. For perceptual evaluation, we use LPIPS [66] with AlexNet [20] features and the CLIP [38]-based metrics proposed in [25]: CLIP-I, which measures CLIP similarity between ground-truth and rendered images, and CLIP-T, which measures temporal consistency via CLIP similarity between rendered frames with 5 frames apart.

iPhone360 dataset. We introduce the iPhone360 dataset to address limitations of existing benchmarks that cannot evaluate 360° dynamic reconstruction performance. Unlike prior datasets [11,24,35,37], iPhone360 provides training images captured from a single handheld camera without camera teleportation, while synchronized test cameras are placed up to 70°–135° apart from the training view, enabling evaluation under extreme novel-view conditions. The large viewpoint disparity and realistic monocular capture setup make iPhone360 a suitable benchmark for evaluating the challenging 360° dynamic reconstruction.

The dataset consists of 6 dynamic scenes: *block2*, *goat*, *jacket*, *jelly*, *pull-up*, and *walk-around*. Each scene captures an object with realistic scenario. A compact comparison with existing dynamic datasets is shown in Tab. 1. All sequences are captured using multiple iPhone devices, and follow the same acquisition framework as iPhone dataset. We use LiDAR metric depth to perform scale and shift correction on the estimated depth [5,61] results, and use camera parameters from the sensor as ground truth.

In Tab. 2, we report perceptual metrics on our iPhone360 dataset. Since rendered backgrounds in extreme novel view settings tend to contain large empty regions, direct image comparison becomes less meaningful. To address this, we define a bounding box centered on the ground-truth dynamic object mask expanded

Fig. 4: iPhone360 Qualitative Comparison. In (a), Novel-view GTs are masked and model outputs are shown as renderings of dynamic Gaussians. Blue pyramids denote training cameras, and reds denote test camera locations in the *Zoom-out*. (b) shows bullet-time rendering results across 360° at a fixed time instant.

Table 1: Compact comparison across datasets. #**Tr.**: number of training cameras, **Wild**: in-the-wild scenario, **Max** $\Delta\theta$: maximum angular difference between training and test views, **360°**: ability to evaluate reconstruction in 360° manner.

Dataset	#Tr.	Wild	Max $\Delta\theta$	360°
D-NeRF [37]	∼150	✗	–	✗
HyperNeRF [35]	2	✗	–	✗
Nerfies [34]	2	✗	–	✗
NSFF [24]	24	✗	–	✗
iPhone [11]	1	✓	$< 45°$	✗
iPhone360	**1**	✓	**70°–135°**	✓

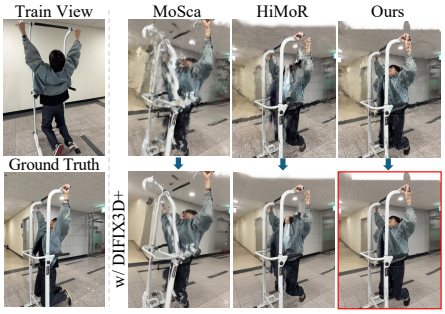

Fig. 5: Results with Diffusion Prior. Our method outperforms baselines when combined with DIFIX3D+ [57].

Table 2: Quantitative results on iPhone360 dataset.

	Block2			Goat			Jacket		
	CLIP-I↑	CLIP-T↑	LPIPS↓	CLIP-I↑	CLIP-T↑	LPIPS↓	CLIP-I↑	CLIP-T↑	LPIPS↓
MoSca [23]	0.7499	0.9513	0.5556	0.7379	**0.9522**	0.6332	0.7073	0.9440	0.5280
HiMoR [25]	0.8422	0.9549	0.3122	0.8357	0.9426	0.3260	0.7041	**0.9526**	0.2942
Ours	**0.9021**	**0.9633**	**0.2569**	**0.8706**	0.9345	**0.2244**	**0.9147**	0.9445	**0.2747**
	Jelly			Pull-up			Walk-around		
	CLIP-I↑	CLIP-T↑	LPIPS↓	CLIP-I↑	CLIP-T↑	LPIPS↓	CLIP-I↑	CLIP-T↑	LPIPS↓
MoSca [23]	0.6239	**0.9649**	0.7127	0.8081	0.9617	0.5241	0.7323	**0.9639**	0.7439
HiMoR [25]	0.6839	0.9357	0.3539	0.8862	0.9786	0.2240	0.8333	0.9590	0.4014
Ours	**0.8359**	0.9229	**0.3538**	**0.8978**	**0.9827**	**0.2163**	**0.8718**	0.9627	**0.3938**

by a spatial margin, and apply it to both the ground-truth and foreground-only rendered results. All metrics are then computed within this region. Our method achieves consistently higher scores on the iPhone360 dataset compared with baselines [23, 25]. As an additional reference alongside the CLIP-based metrics, we report pixel-level evaluation results in Tab. 4. The low absolute scores reflect the inherent difficulty of the task: achieving pixel-level accuracy in unseen regions under full extrapolation.

Furthermore, as shown Fig. 1 and Fig. 4, our reconstructions maintain coherent geometry even from extreme viewpoints, demonstrating the effectiveness of the proposed method. Fig. 4(b) demonstrates that our method achieves 360° object reconstruction from a single video.

iPhone dataset. The iPhone [11] dataset contains training videos captured while moving between two test cameras, designed to evaluate dynamic reconstruction models in a monocular setting. Following HiMoR [25], we report CLIP-I, CLIP-T, and LPIPS [66] scores on Tab. 3. Across these metrics, our method shows consistent improvements over previous models. We compare our model with NeRF-based [11,35] and 3DGS-based [25,44,49,62] approaches across five scenes. Our method outperforms on most scenes. The qualitative results on the *spin*

Table 3: Quantitative results of novel view synthesis on the iPhone dataset [11]. Highlighted cells denote best, second best, third best.

Method	Apple			Block			Paper-windmill		
	CLIP-I↑	CLIP-T↑	LPIPS↓	CLIP-I↑	CLIP-T↑	LPIPS↓	CLIP-I↑	CLIP-T↑	LPIPS↓
T-NeRF [11]	0.8275	0.9729	0.6695	0.8873	0.9749	0.4729	0.9304	0.9858	0.4837
HyperNeRF [35]	0.8314	0.9771	0.6626	0.8882	0.9756	0.4654	0.9218	0.9818	0.4319
Deformable 3DGS [62]	0.7822	0.9730	0.8558	0.8105	0.9720	0.7281	0.8652	0.9814	0.6411
Marbles [44]	0.8055	0.9653	0.7025	0.8492	0.9648	0.5303	0.8610	0.9791	0.6280
SoM [49]	0.8100	0.9721	0.6335	0.8658	0.9745	0.5083	0.9225	0.9833	0.3253
HiMoR [25]	0.8798	0.9747	0.5926	0.8707	0.9750	0.5059	0.9275	0.9853	0.3216
Ours	0.8775	0.9794	0.5414	0.8933	0.9770	0.4105	0.9446	0.9854	0.2055

Method	Spin			Teddy			**Mean**		
	CLIP-I↑	CLIP-T↑	LPIPS↓	CLIP-I↑	CLIP-T↑	LPIPS↓	CLIP-I↑	CLIP-T↑	LPIPS↓
T-NeRF [11]	0.8328	0.9565	0.5714	0.8242	0.9541	0.6337	0.8604	0.9688	0.5662
HyperNeRF [35]	0.8498	0.9594	0.4905	0.8836	0.9630	0.5801	0.8750	0.9714	0.5261
Deformable 3DGS [62]	0.7457	0.9712	0.5962	0.7791	0.9629	0.7601	0.7965	0.9721	0.7163
Marbles [44]	0.8272	0.9527	0.5761	0.8097	0.9606	0.6671	0.8305	0.9645	0.6208
SoM [49]	0.8510	0.9585	0.3832	0.8521	0.9676	0.5630	0.8603	0.9712	0.4827
HiMoR [25]	0.8853	0.9658	0.3696	0.8902	0.9744	0.5296	0.8907	0.9750	0.4639
Ours	0.9029	0.9626	0.2957	0.8891	0.9726	0.4856	0.9015	0.9754	0.3877

Fig. 6: iPhone dataset Qualitative result. Backpack scene shows that our model faithfully reconstructs unseen backside regions, clearly outperforming prior methods.

Table 4: Pixel-level Evaluation on iPhone and iPhone360 Datasets. All training and evaluations are conducted at 1x resolution.

	iPhone dataset		iPhone360 dataset	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
HiMoR	16.2441	**0.6329**	10.9883	0.6854
Ours	**16.3431**	0.6309	**11.3364**	**0.7012**

scene are included in Fig. 1. Additionally, Fig. 6 presents extreme-viewpoint renderings for the *backpack* scene, which lacks test ground truth, showing that our model reconstructs unseen regions with substantially higher fidelity than competing methods [25, 49].

DAVIS dataset. We evaluated our model on scenes from the DAVIS [36] dataset, which captures fast-moving objects but lacks ground-truth camera parameters and depth maps, making them closer to real-world conditions. As shown in Fig. 7, our method better preserves geometry and reconstructs occluded regions compared to previous models under in-the-wild scenarios.

4.2 Ablations

We conducted a qualitative ablation study on the *walk-around* and *jelly* scenes of the iPhone360 dataset to demonstrate the advantage of AnchorTAP3D over baselines using naive 2D and 3D tracking. As shown in Fig. 8, our method achieves better 360° reconstruction, reconstructing the geometry of occluded regions at extreme viewpoints, compared to 'w/o 3D init' and 'w/o Anchor'.

Fig. 7: DAVIS dataset Qualitative Comparison. Results show that our model maintains coherent geometry and recovers occluded parts more reliably than previous approaches.

Fig. 8: Ablations. Qualitative results for different initialization methods. 'w/o 3D init' directly unprojects 2D tracks, 'w/o Anchor' leverages naive 3D point tracking model [65], and Ours uses AnchorTAP3D at initialization. The left and right scenes correspond to the *walk-around* and *jelly* sequences from iPhone360, respectively.

The 'w/o 3D init' variant follows prior models [25, 49] in initializing with unprojected 2D tracks, and fails to preserve geometry in occluded regions. Using a naive 3D point tracking model [65] 'w/o Anchor' yields better performance than 'w/o 3D init' in *walk-around* by partially recovering geometry in occluded regions, but in the *jelly*, it completely failed to reconstruct occluded area. This indicates that 3D tracking becomes unreliable in long sequences without anchor guidance. Finally, Ours preserves the overall object shape across both scenes. These results demonstrate the effectiveness of our method in 4D reconstruction over both 2D track-based and naive 3D tracking approaches.

5 Limitations and Conclusion

Limitations. While our method demonstrates improved 360° object reconstruction compared to existing approaches, several limitations remain. While our AnchorTAP3D improves significantly upon naive applications of off-the-shelf 2D and 3D tracking models, the overall performance still depends on the capability of pretrained models. Furthermore, our method cannot synthesize extreme viewpoint's background regions if it is invisible in the input video. We believe this can be addressed in future work through integration with diffusion priors. As shown in Fig. 5, our method, with its better-preserved geometry, provides a more suitable starting point for diffusion model than existing approaches. Our method also still struggles with pixel-level metrics such as PSNR and SSIM under the inherently challenging large extrapolation settings of iPhone360, and we leave further improvement of pixel-level reconstruction accuracy to future work.

Conclusion. We present 4DGS360, a novel framework that enables extreme novel-view synthesis from monocular videos. Unlike prior methods that rely on pretrained 2D-based features and suffer from 3D ambiguity, we introduce AnchorTAP3D, an advanced 3D tracking model that provides reliable geometry-aware initialization. This design extends the effectiveness of optimization strategies, allowing consistent 360° reconstruction of dynamic objects. Furthermore, we identify limitations in current datasets and propose a new benchmark, iPhone360, which includes dynamic objects exhibiting diverse motion patterns with extreme viewpoint validations. This dataset enables more realistic evaluation of how well reconstruction models generalize to challenging real-world conditions.

Acknowledgements

This work was supported by the Korean Government through the grants from IITP (RS-2021-II211343 and RS-2025-25442338) and COMPA (RS-2026-25537086).

References

1. 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)
2. 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)
3. 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)
4. Charatan, D., Li, S., Tagliasacchi, A., Sitzmann, V.: pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In: CVPR (2024)
5. Chen, S., Guo, H., Zhu, S., Zhang, F., Huang, Z., Feng, J., Kang, B.: Video depth anything: Consistent depth estimation for super-long videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 22831–22840 (2025)
6. Chen, Y., Zheng, C., Xu, H., Zhuang, B., Vedaldi, A., Cham, T.J., Cai, J.: Mvsplat360: Feed-forward 360 scene synthesis from sparse views. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 107064–107086. Curran Associates, Inc. (2024). https://doi.org/10.52202/079017-3399, https://proceedings.neurips.cc/paper_files/paper/2024/file/c196239c5f9481e0db2755f31fe4585f-Paper-Conference.pdf
7. Chu, W.H., Ke, L., Fragkiadaki, K.: Dreamscene4d: Dynamic multi-object scene generation from monocular videos. Advances in Neural Information Processing Systems **37**, 96181–96206 (2024)
8. Doersch, C., Luc, P., Yang, Y., Gokay, D., Koppula, S., Gupta, A., Heyward, J., Rocco, I., Goroshin, R., Carreira, J.a., Zisserman, A.: Bootstap: Bootstrapped training for tracking-any-point. In: Proceedings of the Asian Conference on Computer Vision (ACCV). pp. 3257–3274 (December 2024)

9. Doersch, C., Yang, Y., Vecerik, M., Gokay, D., Gupta, A., Aytar, Y., Carreira, J., Zisserman, A.: Tapir: Tracking any point with per-frame initialization and temporal refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10061–10072 (2023)
10. Duckworth, D., Hedman, P., Reiser, C., Zhizhin, P., Thibert, J.F., Lučić, M., Szeliski, R., Barron, J.T.: Smerf: Streamable memory efficient radiance fields for real-time large-scene exploration. ACM Transactions on Graphics (TOG) **43**(4), 1–13 (2024)
11. Gao, H., Li, R., Tulsiani, S., Russell, B., Kanazawa, A.: Monocular dynamic view synthesis: A reality check. In: NeurIPS (2022)
12. Harley, A.W., Fang, Z., Fragkiadaki, K.: Particle video revisited: Tracking through occlusions using point trajectories. In: European Conference on Computer Vision (2022), https://api.semanticscholar.org/CorpusID:248069518
13. Hu, Y., Liu, Z., Shao, J., Lin, Z., Zhang, J.: Eva-gaussian: 3d gaussian-based real-time human novel view synthesis under diverse camera settings (2024), https://arxiv.org/abs/2410.01425
14. Karaev, N., Makarov, I., Wang, J., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. In: Proc. arXiv:2410.11831 (2024)
15. Karaev, N., Rocco, I., Graham, B., Neverova, N., Vedaldi, A., Rupprecht, C.: Cotracker: It is better to track together. In: Proc. ECCV (2024)
16. 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)
17. Kim, M., Lim, J., Han, B.: 4d gaussian splatting in the wild with uncertainty-aware regularization. Advances in Neural Information Processing Systems **37**, 129209–129226 (2024)
18. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. CoRR **abs/1412.6980** (2014), https://api.semanticscholar.org/CorpusID:6628106
19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 4015–4026 (October 2023)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. NeurIPS **25** (2012)
21. Lee, I., Jang, J.W., Seo, S., Kwak, N.: Divcon-nerf: Generating augmented rays with diversity and consistency for few-shot view synthesis. arXiv preprint arXiv:2503.12947 (2025)
22. 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. arXiv preprint arXiv:2312.02135 (2023)
23. 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)
24. 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)
25. 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)

26. 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)
27. Liu, T., Wang, G., Hu, S., Shen, L., Ye, X., Zang, Y., Cao, Z., Li, W., Liu, Z.: Mvsgaussian: Fast generalizable gaussian splatting reconstruction from multi-view stereo. In: European Conference on Computer Vision. pp. 37–53. Springer (2025)
28. 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)
29. 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)
30. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In: 2024 International Conference on 3D Vision (3DV). pp. 800–809. IEEE (2024)
31. Mallick, S.S., Goel, R., Kerbl, B., Steinberger, M., Carrasco, F.V., De La Torre, F.: Taming 3dgs: High-quality radiance fields with limited resources. In: SIGGRAPH Asia 2024 Conference Papers. pp. 1–11 (2024)
32. 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)
33. Paliwal, A., Ye, W., Xiong, J., Kotovenko, D., Ranjan, R., Chandra, V., Kalantari, N.K.: Coherentgs: Sparse novel view synthesis with coherent 3d gaussians. In: European Conference on Computer Vision. pp. 19–37. Springer (2024)
34. 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)
35. 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)
36. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Gool, L.V.: The 2017 davis challenge on video object segmentation (2018), https://arxiv.org/abs/1704.00675
37. 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)
38. 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)
39. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
40. Schönemann, P.H.: A generalized solution of the orthogonal procrustes problem. Psychometrika **31**(1), 1–10 (1966). https://doi.org/10.1007/BF02289451
41. Seo, S., Chang, Y., Kwak, N.: Flipnerf: Flipped reflection rays for few-shot novel view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22883–22893 (2023)

42. Shen, Y., Ceylan, D., Guerrero, P., Xu, Z., Mitra, N., Wang, S., Frühstück, A.: Supergaussian: Repurposing video models for 3d super resolution. In: European Conference on Computer Vision (ECCV) (2024)
43. Sorkine, O., Alexa, M.: As-rigid-as-possible surface modeling. In: Proceedings of the Fifth Eurographics Symposium on Geometry Processing. p. 109–116. SGP '07, Eurographics Association, Goslar, DEU (2007)
44. 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)
45. Team, S.D., Chen, X., Chu, F.J., Gleize, P., Liang, K.J., Sax, A., Tang, H., Wang, W., Guo, M., Hardin, T., Li, X., Lin, A., Liu, J., Ma, Z., Sagar, A., Song, B., Wang, X., Yang, J., Zhang, B., Dollár, P., Gkioxari, G., Feiszli, M., Malik, J.: Sam 3d: 3dfy anything in images (2025), https://arxiv.org/abs/2511.16624
46. Teed, Z., Deng, J.: Droid-slam: deep visual slam for monocular, stereo, and rgb-d cameras. In: Proceedings of the 35th International Conference on Neural Information Processing Systems. NIPS '21, Curran Associates Inc., Red Hook, NY, USA (2021)
47. Verbin, D., Hedman, P., Mildenhall, B., Zickler, T., Barron, J.T., Srinivasan, P.P.: Ref-nerf: Structured view-dependent appearance for neural radiance fields. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
48. Wang, C., Chai, M., He, M., Chen, D., Liao, J.: Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3835–3844 (2022)
49. 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)
50. Wang, R., Xu, S., Dong, Y., Deng, Y., Xiang, J., Lv, Z., Sun, G., Tong, X., Yang, J.: Moge-2: Accurate monocular geometry with metric scale and sharp details. Advances in Neural Information Processing Systems **38**, 35928–35959 (2026)
51. 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)
52. Wang, S., Leroy, V., Cabon, Y., Chidlovskii, B., Revaud, J.: Dust3r: Geometric 3d vision made easy. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 20697–20709 (June 2024)
53. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
54. Wenbo, C., Ligang, L.: Deblur-gs: 3d gaussian splatting from camera motion blurred images. Proc. ACM Comput. Graph. Interact. Tech. (Proceedings of I3D 2024) **7**(1) (2024). https://doi.org/10.1145/3651301, http://doi.acm.org/10.1145/3651301
55. Wu, D., Liu, F., Hung, Y.H., Qian, Y., Zhan, X., Duan, Y.: 4d-fly: Fast 4d reconstruction from a single monocular video. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 16663–16673 (2025)
56. 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)

57. Wu, J.Z., Zhang, Y., Turki, H., Ren, X., Gao, J., Shou, M.Z., Fidler, S., Gojcic, Z., Ling, H.: Difix3d+: Improving 3d reconstructions with single-step diffusion models. In: Conf. Comput. Vis. Pattern Recog. pp. 26024–26035 (2025)
58. Xiao, Y., Wang, J., Xue, N., Karaev, N., Makarov, Y., Kang, B., Zhu, X., Bao, H., Shen, Y., Zhou, X.: Spatialtrackerv2: 3d point tracking made easy. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025), https://arxiv.org/abs/2507.12462
59. Yang, J., Gao, M., Li, Z., Gao, S., Wang, F., Zheng, F.: Track anything: Segment anything meets videos (2023)
60. Yang, L., Kang, B., Huang, Z., Xu, X., Feng, J., Zhao, H.: Depth anything: Unleashing the power of large-scale unlabeled data. In: CVPR (2024)
61. 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)
62. 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)
63. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4578–4587 (2021)
64. Yu, Z., Chen, A., Huang, B., Sattler, T., Geiger, A.: Mip-splatting: Alias-free 3d gaussian splatting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19447–19456 (June 2024)
65. Zhang, B., Ke, L., Harley, A.W., Fragkiadaki, K.: Tapip3d: Tracking any point in persistent 3d geometry. arXiv preprint arXiv:2504.14717 (2025)
66. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Conf. Comput. Vis. Pattern Recog. pp. 586–595 (2018)
67. Zhao, L., Wang, P., Liu, P.: Bad-gaussians: Bundle adjusted deblur gaussian splatting. In: European Conference on Computer Vision (ECCV) (2024)