HoliGS: Holistic Gaussian Splatting for Embodied View Synthesis

Anonymous ICCV submission

Paper ID

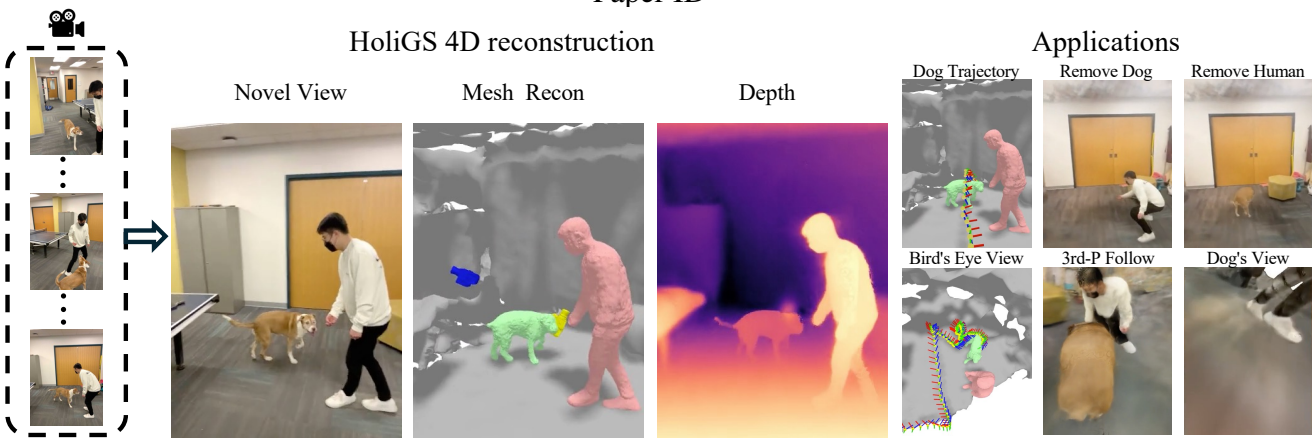

Figure 1. **Overview.** From a phone capture of humans and animals in motion, HoliGS reconstructs temporally consistent geometry, appearance, and depth, enabling novel-view synthesis, deformable mesh recovery, and dense depth estimation. These reconstructions support a range of embodied applications, including actor-specific view synthesis (*e.g.*, third-person and egocentric perspectives), object-specific removal, and actor-centric visualization (*e.g.*, dog's-eye view). HoliGS also enables spatiotemporal behavior analysis such as trajectory visualization.

Abstract

*We propose HoliGS, a novel deformable Gaussian splatting framework that addresses embodied view synthesis from long monocular RGB videos. Unlike prior 4D Gaussian splatting and dynamic NeRF pipelines, which struggle with training overhead in minute-long captures, our method leverages invertible Gaussian Splatting deformation networks to reconstruct large-scale, dynamic environments accurately. Specifically, we decompose each scene into a static background plus time-varying objects, each represented by learned Gaussian primitives undergoing global rigid transformations, skeleton-driven articulation, and subtle non-rigid deformations via an invertible neural flow. This hierarchical warping strategy enables robust free-viewpoint novel-view rendering from various embodied camera trajectories by attaching Gaussians to a complete canonical foreground shape (*e.g.*, egocentric or third-person follow), which may involve substantial viewpoint changes and interactions between multiple actors. Our experiments demonstrate that HoliGS achieves superior reconstruction quality on challenging datasets while significantly reducing both training and rendering time compared to state-of-the-art monocular deformable NeRFs. These results highlight a practical and scalable solution for EVS in real-world scenarios. The source code will be released.*

1. Introduction

Understanding and reconstructing dynamic 3D scenes from monocular video remains a fundamental challenge in computer vision, particularly in the context of Embodied View Synthesis (EVS), where camera trajectories dynamically follow actor motions. EVS tasks are crucial for immersive AR/VR experiences, interactive gaming, and robotics, demanding representations capable of handling complex non-rigid deformations, extreme viewpoint changes, and extended temporal sequences.

Despite recent advances in neural rendering for static scenes [20, 36], extending these techniques to dynamic and non-rigid scenarios reveals significant computational and representational challenges. Existing neural radiance fields (NeRF)-based methods [48] face high computational costs during both training and inference, particularly when scaling to minute-long sequences and involving multiple interacting objects. This significantly restricts their practical ap-

Figure 2. Performance of SOTA methods.

Method	Entire Scenes	Deform. Objects	Global 6-DOF Traj.	Long Videos	Extreme Views	Fast Rendering
BANMo	✗	✓	✗	✓	✓	✗
RAC	✗	✓	✗	✓	✓	✗
Vidu4D	✗	✓	✗	✓	✓	✓
MoSca	✓	✓	✗	✗	✗	✓
SoM	✓	✓	✗	✗	✗	✓
SC-GS	✓	✓	✗	✗	✗	✓
Dyn.Guss	✓	✓	✗	✓	✗	✓
G.Marbles	✓	✓	✗	✓	✗	✓
Total-Recon	✓	✓	✓	✓	✓	✗
Ours	✓	✓	✓	✓	✓	✓

Table 1. **Comparison to Related Work.** HoliGS targets embodied view synthesis of dynamic scenes and process *minute-long videos* of dynamic scenes, and render *extreme views*.

plicability in real-time environments.

Gaussian Splatting (GS) approaches [20], known for efficient rendering in static scenes through compact anisotropic Gaussian primitives, also encounter limitations in dynamic contexts. Current deformable Gaussian Splatting techniques [17, 30] are typically constrained to short-duration captures or scenarios with minimal non-rigid motion. When applied to EVS tasks involving intricate interactions, these methods yield inconsistent reconstructions with noticeable artifacts(see Figure 2).

Furthermore, several existing methods [23, 50, 58] rely heavily on off-the-shelf point-tracking models [17], introducing significant computational overhead and exhibiting fragility under severe occlusions. These methods also fail to generalize effectively to arbitrary viewpoint trajectories essential for comprehensive EVS scenarios, severely limiting their utility in real-world conditions marked by frequent occlusions and the need for viewpoint flexibility.

To overcome these critical limitations, we propose HoliGS, a holistic Gaussian Splatting method explicitly designed for EVS applications. Unlike previous methods, our framework introduces a Gaussian-based deformation model that directly manages articulated non-rigid transformations without relying on traditional tracking pipelines. This innovation ensures consistent and artifact-free reconstructions across complex sequences involving human and animal interactions.

Specifically, our approach includes a novel deformable Gaussian Splatting pipeline and an optimized strategy to maintain high-quality rendering under extreme viewpoint variations, such as egocentric, third-person follow, and overhead perspectives. Additionally, we integrate an invertible deformation model, enabling stable reconstructions over prolonged durations without sacrificing efficiency.

Extensive experimental evaluation demonstrates that HoliGS significantly outperforms state-of-the-art methods in terms of both rendering quality and computational speed, achieving real-time rendering capabilities on consumer hardware. Our results confirm robust performance across diverse, challenging, dynamic sequences featuring multiple interacting entities and complex articulated motions, scenarios where prior techniques either fail or produce substantial visual artifacts. The main contributions of this work are:

- We introduce a holistic Gaussian Splatting method for EVS tailored to 6-DOF embodied camera paths, outperforming existing state-of-the-art approaches [48, 67].
- We propose an invertible deformation model that ensures stable reconstruction over extended periods without compromising computational efficiency.
- We evaluate our model on diverse challenging dynamic scenes against existing methods and show that our approach achieves robust view synthesis and scalable to minute-long videos.

2. Related Work

Dynamic Scene Reconstruction. Reconstructing dynamic scenes from videos has been an active research area, traditionally relying on multi-view stereo systems [1, 3, 4, 6, 12, 24, 28, 34, 42, 51, 73]. Recently, another series of works focusing on monocular scene reconstruction methods [5, 13, 25, 27, 31, 35, 49, 54–56, 63, 64, 68, 70, 71]. Dynamic methods often utilize either temporal conditioning as an additional input dimension[39] or canonical-space representations with deformation fields [25, 38]. Grid-based representations [7, 45] have further accelerated these methods, enabling efficient optimization for dynamic scene reconstruction [6, 11, 47]. Despite significant progress, these approaches still suffer from high computational costs, especially in real-time and long video scenarios with complex motion patterns or prolonged video sequences.

Embodied View Synthesis (EVS). EVS introduces additional complexity, requiring representations capable of han-

dling camera trajectories that closely follow or interact with dynamic subjects. Existing methods like DyCheck [14] highlight the inadequacies of current benchmarks, which often do not accurately reflect realistic everyday scenarios involving limited viewpoints and complex dynamics. Methods designed specifically for monocular EVS [26, 48] aim to mitigate these issues through hybrid representations or generative methods. Nevertheless, these methods typically rely heavily on domain-specific priors or computationally intensive tracking modules, restricting their robustness under occlusions and generalization across diverse view trajectories.

Articulated Object Reconstruction. Articulated object reconstruction, especially for humans and animals, often utilze parametric templates [2, 33, 44, 74], which impose strong geometric priors and facilitate reconstruction from sparse views or monocular videos [15, 16, 21, 59]. However, these models typically struggle with capturing personalized or detailed appearance variations. More recent non-parametric neural methods have combined neural radiance fields with articulated models [8, 9, 29, 40, 60, 61, 65, 66], capturing richer detail but at a significant computational cost. Our method diverges by directly modeling articulated motion without relying on predefined parametric templates, instead employing a flexible Gaussian-based deformation model optimized for dynamic reconstruction.

Non-Rigid Structure from Motion. Non-rigid Structure from Motion (NRSfM) aims to reconstruct the 3D shape and deformation of objects from monocular videos, handling scenarios where scene points undergo complex, articulated, or continuous deformation. Traditional SfM and visual SLAM methods [37, 46, 53] typically assume static environments, enforcing strict epipolar constraints unsuitable for dynamic scenes. Recent methods address this limitation by jointly estimating camera poses, scene geometry, and deformation fields [22, 72]. These approaches, however, often rely on time-intensive test-time optimization or explicit motion segmentation, limiting their scalability and efficiency. Differently, our method leverages a Gaussian-based deformation model to explicitly encode articulated non-rigid transformations, enabling efficient reconstruction without the need for computationally costly per-video fine-tuning or explicit motion segmentation. This approach facilitates robust reconstruction of dynamic interactions in everyday monocular videos, effectively overcoming challenges posed by occlusions and extensive deformation.

The proposed framework, HoliGS, combines the advantages of articulated object reconstruction and static Gaussian Splatting to enable efficient, high-quality embodied view synthesis for dynamic scenes captured from monocular videos, overcoming limitations associated with existing methods.

In this section, we introduce HoliGS, a hierarchical 4D representation that models dynamic scenes as the union of a static background and time-varying deformable objects. Our framework leverages Gaussian Splatting to represent both the static and dynamic components and employs a series of invertible warping operations to capture articulated and non-rigid deformations. The final scene at time t is given by $\mathcal{S}(t)=\mathcal{G}(t) \cup \mathcal{H}$, where $\mathcal{H}$ is the set of static background Gaussians and $\mathcal{G}(t)$ contains the dynamic, time-varying Gaussians splitting articulated foreground objects.

2.1. Hierarchical Dynamic Warping

To robustly model motion ranging from whole-body translations to fabric flutter, we use a two-stage warping strategy. At a glance, large articulated displacements are first explained by a skeleton-driven transform, after which a soft, flow-based deformation field refines any residual non-rigid detail. All derivations and exact matrix expressions are deferred to the supplementary material.

Global movements. Every video frame is aligned to the camera via two rigid $\mathrm{SE}(3)$ transforms: the *background-to-camera* map G_b and the *object-root-to-camera* map G_o. Both transforms are regressed by lightweight Fourier MLPs that output six twist parameters per frame, giving us frame-specific poses without needing an external tracker.

Skeleton-driven warping. The core articulated motion is handled by a bone hierarchy with B bones. Each bone b has a static reference pose $\left(c_b^*, V_b^*, \Lambda_b^*\right)$ encoding center, rotation, and scale, respectively. At time t, a learned twist vector $\hat{\eta}_b(t) \in \mathrm{SE}(3)$ is exponentiated to produce the bone pose $J_b(t)$. We measure how much a 3-D point P_k^* belongs to each bone by a Mahalanobis distance in the bone's scaled–rotated frame; a softmax over these distances yields skinning weights $w(t)$. Dual-quaternion blend skinning (DQB) [18] fuses the individual bone transforms into a single $\mathrm{SE}(3)$ map $J(t)$, which is then applied to every Gaussian center, rotation, and scale. Conceptually, this step captures all "rigid-but-articulated" effects such as limbs, torsos, or tails.

Soft deformation field. After skeletal warping, many objects still exhibit subtle surface changes—loose clothing, hair swaying, muscle bulges—that cannot be explained by rigid bones. We address this with a *soft deformation field* $S(\cdot, \omega_d)$ implemented as an invertible RealNVP flow [10]. Given a canonical point X and a per-frame latent code ω_d, the field outputs a refined position $X' = S(X, \omega_d)$. Invertibility guarantees that S^{-1} exists; we therefore impose a 3-D cycle-consistency loss: $\mathcal{L}_{\text{cyc}} = \|S^{-1}(S(X, \omega_d), \omega_d) - X\|_2^2$, which forces the forward and reverse mappings to cancel out and stabilizes training. Because the flow operates in a *fixed* canonical space, it never has to chase a moving target, allowing it to converge quickly even when the deformations are highly nonlinear.

Why hierarchy matters. Articulated bones give the model an inductive bias toward plausible large-scale motion, while

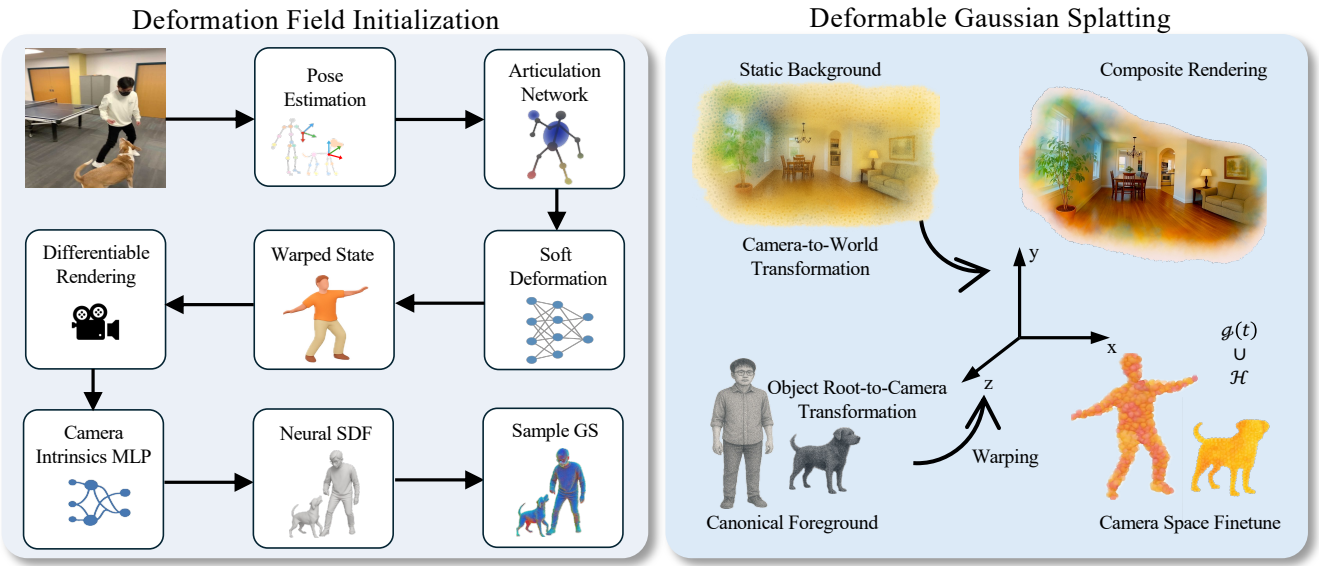

Figure 3. **HoliGS Pipeline.** *Left*—Warping network initialization: We jointly optimize poses, articulation, soft deformation, and in a neural SDF proxy to obtain a fast converging deformation field that provides a strong starting point for Gaussian splitting. *Right*—after initialization, the objective is switched to dynamic Gaussian splatting, and the deformed foreground is composited with the static background to yield the final 4D scene.

the soft field soaks up the remaining fine detail. Each module solves a simpler task and therefore converges faster than a single, monolithic deformation network. Empirically, the skeletal stage explains $\approx 90\%$ of visible motion energy, leaving only low-amplitude corrections to the RealNVP field. Full mathematical details—the Lie-algebra twist representation, the exact Mahalanobis weighting, and the DQB formulation—are provided in the supplementary materials.

Combined warping pipeline. Integrating the above components, a point X^* in canonical space is warped to its dynamic position at time t according to:

$$X^t = G_{\mathrm{o}}^{t\,-1} \cdot J^{t-1} \cdot S^{-1}\left(X^*, \omega_d^t\right). \tag{1}$$

Inspired by Omnimotion [57], HoliGS also enables a forward warp

$$X^* = S \cdot J^t \cdot G_{\mathrm{o}}^t\left(X^t, \omega_d^t\right). \tag{2}$$

This unified warping function seamlessly integrates global, skeletal articulation, and fine-scale deformations, enabling our framework to render high-quality 4D scenes with complex dynamics.

2.2. Deformation Network Initialization

For our dynamic scene representation, we establish initial transformation parameters by pre-training a neural SDF that warps sampled points on camera rays from the static state to the warped states, similar to [65]. We apply Posenet [19] to obtain the rigid-body transformations T^d and time-dependent skeletons for each deformable object in the scene. This network provides robust pose estimates even under challenging viewing conditions. Concurrently, we initialize the background component transformations T^s using camera pose information extracted from the capture device's motion sensors. This hybrid initialization strategy ensures stable convergence during subsequent optimization stages while accommodating both foreground dynamic objects and static background elements within our unified representation. Then, we initialize the foreground Gaussian point cloud from the pre-trained neural SDF by sampling points on its surface, with objective function:

$$\begin{aligned}\mathcal{L} = \underbrace{\mathcal{L}_{\text{photo}}}_{\substack{\text{photometric}\\ \text{consistency}}} + \underbrace{\lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{SDF}}\mathcal{L}_{\text{SDF}}}_{\substack{\text{geometric}\\ \text{constraints}}} \\ + \underbrace{\lambda_{\text{flow}}\mathcal{L}_{\text{flow}} + \lambda_{\text{cycle}}\mathcal{L}_{\text{cycle}}}_{\substack{\text{motion}\\ \text{consistency}}} + \underbrace{\mathcal{L}_{\text{seg}}}_{\substack{\text{mask}\\ \text{supervision}}}.\end{aligned} \tag{3}$$

Here, the photometric loss $\mathcal{L}_{\text{photo}}$ enforces appearance consistency. For geometry constraints: the depth term $\mathcal{L}_{\text{depth}} = \sum_{p^t} \|D(p^t) - \hat{D}(p^t)\|_2^2$ aligns our predicted depth $\hat{D}$ with an off-the-shelf monocular depth estimator D [41], promoting correct scene scale, and the SDF term $\mathcal{L}_{\text{SDF}} = \sum_{X_i^t} (\|\nabla_{X_i^t} \Phi_{\text{SDF}}(X_i^t)\|_2 - 1)^2$ enforces the signed distance field Φ_{SDF} to behave like a true distance function by constraining its gradient norm to one. Motion consistency is imposed by flow loss

$\mathcal{L}_{\text{flow}} = \sum_{p^t} \|V(p^t) - \hat{V}(p^t)\|_2^2$ and cycle loss where

$$\mathcal{L}_{\text{cycle}} = \sum_{i,j} \lambda_j \, \beta_{i,j} \, \|\mathcal{F}^{t'}_{\text{fwd},j}(\mathcal{F}^{t}_{\text{bwd},j}(X_i^t)) - X_i^t\|_2^2 \quad (4)$$

weighted by importance factors λ_j and $\beta_{i,j}$, aligning RAFT optical flow [52] and satisfying forward–backward cycle consistency. Finally, segmentation supervision is given by $\mathcal{L}_{\text{seg}} = \sum_{p^t} \|M_{\text{pred}}(p^t) - M_{\text{gt}}(p^t)\|_2^2$, with M_{gt} obtained from SAM [43]. $p^t \in \mathbb{R}^2$ represents pixel coordinates at time t, $X_i^t \in \mathbb{R}^3$ denotes the i-th sample point in world space corresponding to $X_i^t \in \mathbb{R}^3$ in camera space. Weights $\{\lambda_{\text{depth}}, \lambda_{\text{SDF}}, \lambda_{\text{flow}}, \lambda_{\text{cycle}}\}$ are tuned to balance these complementary constraints.

2.3. Deformable Gaussian Splatting Optimization Objectives

Our composite Gaussian Splatting representation incorporates N scene elements, global transformation matrices T_t^i, and bidirectional deformation fields F^i_{forward} and F^i_{backward}. The optimization process integrates multiple objectives to ensure high-quality reconstruction and temporal consistency:

$$\mathcal{L} = \mathcal{L}_{\text{photo}} + \mathcal{L}_{\text{depth}} + \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{normal}}. \quad (5)$$

Besides the loss terms we explained in initialization, $\mathcal{L}_{\text{photo}}$, $\mathcal{L}_{\text{depth}}$, and $\mathcal{L}_{\text{seg}}$, we incorporate additional normal supervision to align the estimated entire scene surface normals with observed ones $\mathcal{L}_{\text{normal}} = \sum_{p^t} \|N(p^t) - \hat{N}(p^t)\|^2$. This comprehensive optimization framework ensures geometric accuracy, appearance fidelity, and temporal consistency in our dynamic scene representation.

2.4. Embodied View Synthesis

To effectively perform EVS, HoliGS transforms dynamic 3D Gaussian primitives into consistent, egocentric viewpoints that naturally follow the motion of articulated objects, such as humans and animals. Specifically, for each Gaussian primitive, we apply a forward warping function $W_{t \to j} : X^* \to X_t$, which maps points from a canonical space X^* to the deformed configuration at time t. This deformation accounts explicitly for non-rigid articulated transformations, ensuring accurate representation of complex motions such as limb articulations or interactions among multiple entities.

Subsequently, to achieve embodied viewpoints, we employ a rigid-body transformation G_t^0, positioning the virtual egocentric camera within the world coordinate system. It aligns the viewer's perspective with the foreground, enabling realistic rendering of scenarios such as first-person views or third-person perspectives following actors in motion (illustrated in Figure **??**).

By integrating the deformation network, our method reliably synthesizes novel embodied viewpoints that remain coherent across complex motions. Our unified Gaussian-based deformation and viewpoint adjustment strategy significantly simplifies optimization and achieves near real-time performance. This enables practical usage in interactive AR/VR applications, immersive gaming experiences, and robotics, where rapid viewpoint changes and accurate motion tracking are essential.

2.5. Training and Optimization

We adopt a two-phase procedure to optimize our dynamic Gaussian representation: *Component pre-training* and *joint refinement*. During pre-training, each component (e.g., a deformable object or the static background) is optimized separately. Once pre-training is completed, all components are combined for joint refinement using color, depth, normal, and mask objectives. Training follows standard Gaussian Splatting protocols [20]. The synergy between our deformation-centric design and the parametric Gaussian framework accelerates convergence considerably. On NVIDIA H20 GPUs, each pre-training or refinement stage completes in about 30 minutes, enabling full scenes (including multiple deformable objects) to converge in two hours, significantly faster than other approaches.

Component pre-training. We initialize the deformation network by minimizing the overall loss (3), with default weights set as: $\lambda_{\text{depth}} = 5$ (or 1.5 for the HUMAN 1 sequence), $\lambda_{\text{color}} = 0.1$, $\lambda_{\text{flow}} = 1$, $\lambda_{\text{cycle}} = 1$, and $\lambda_{\text{segment}} = 1$. This eikonal term is weighted by $\lambda_{\text{SDF}} = 0.001$ to ensure proper geometric properties. For this computation, we sample 17 uniformly distributed points X_i^t along each camera ray r^t centered at the surface point derived from back-projecting the ground-truth depth.

Joint fine-tuning. During the joint optimization phase, we simultaneously refine all object representations by minimizing loss (5) for an additional 6,000 iterations. The default weights for these objectives are $\lambda_{\text{photo}} = 1$, $\lambda_{\text{normal}} = 1$, $\lambda_{\text{depth}} = 5$, and $\lambda_{\text{seg},j} = 1$. By default, we freeze the background's appearance and geometry parameters while allowing optimization of its global transformation T_0^b, the foreground objects' transformations T_t^f, and the foreground appearance and geometry parameters (for HUMAN 1, we use $\lambda_{\text{depth}} = 1.5$), we allow background appearance and geometry optimization during joint fine-tuning). This joint fine-tuning phase significantly enhances the visual coherence of foreground elements and improves the modeling of inter-object interactions.

2.6. Qualitative and Quantitative Results

Figure 5 shows representative visualizations comparing the photometric and depth reconstruction quality of HoliGS against Total-Recon [48], Deformable GS [69], and 4DGS [62]. These results demonstrate the superior performance of our method under various challenging conditions.

Quantitative results for novel view synthesis are reported

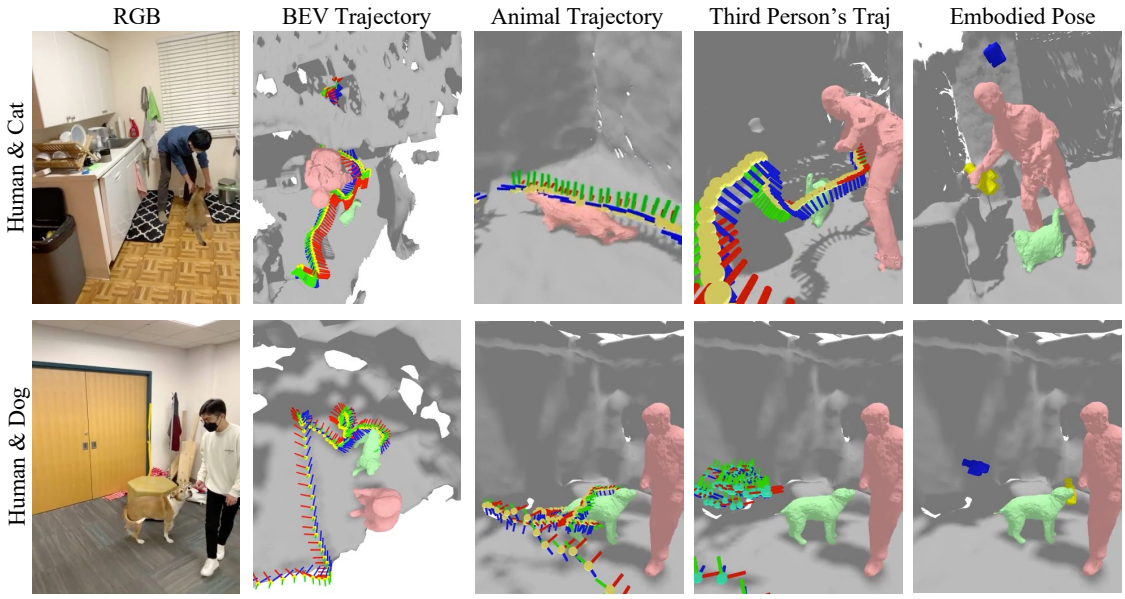

Figure 4. **Foreground Embodied Trajectory.** For two challenging sequences, HumanCat and HumanDog, we show: (i) the joint bird's-eye-view (BEV) trajectory of a foreground actor, (ii) the articulated animal trajectory, (iii) the articulated human trajectory, and (iv) both objects' embodied camera pose. Our method recovers smooth, collision-free paths that faithfully follow each actor while remaining mutually consistent, enabling stable first-person or over-the-shoulder replays for complex multi-agent interactions.

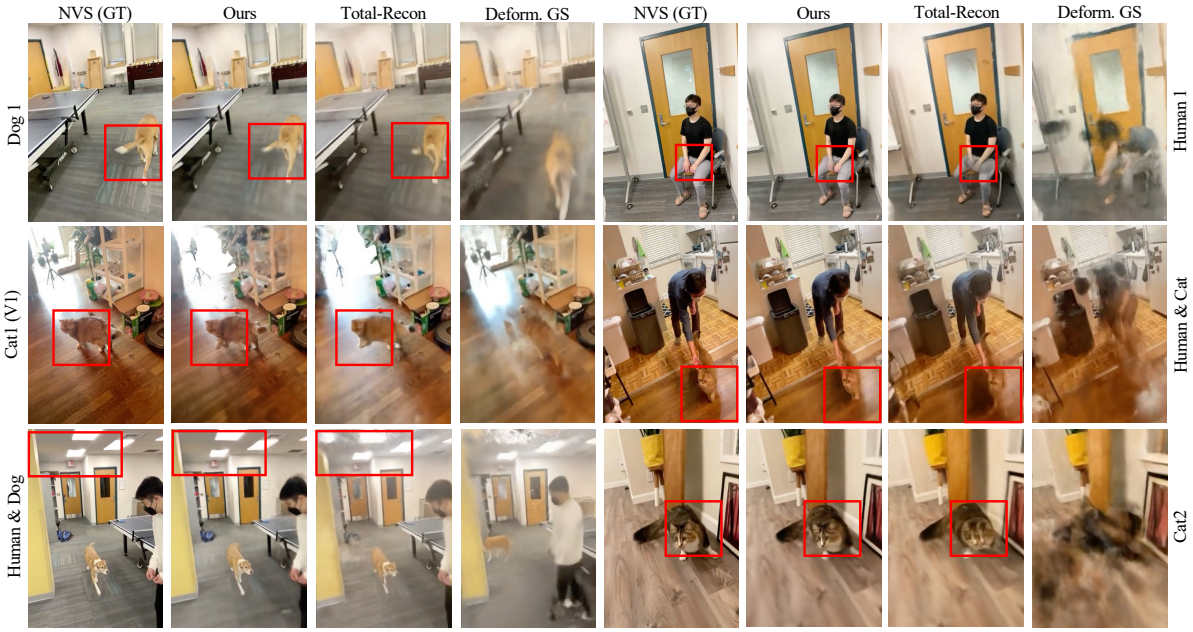

Figure 5. **Baseline Comparison.** We qualitatively compare HoliGS against four SOTA baselines and a direct NVS ground-truth reference across *Dog 1*, *Cat 1*, *Human 1*, and the challenging multi-actor *Human 2 & Cat* sequences. Each column shows photometric renderings (top) and corresponding depth reconstructions (bottom). Red inset boxes highlight the most error-prone regions for articulated motion and occlusion (e.g. tail swing, paw lift, garment folds, and human–animal interaction). Compared with baselines, HoliGS better preserves fine-grained appearance and yields geometrically consistent depth maps with fewer tearing or bleeding artifacts—especially under large viewpoint changes and prolonged, highly non-rigid deformations.

in Tables 2 and 3. Table 2 presents visual metrics across the Total-Recon dataset, while Table 3 reports depth accuracy

	Dog 1 (v1) (626 images)			Dog 1 (v2) (531 images)			Cat 1 (v1) (641 images)			Cat 1 (v2) (632 images)			Cat 2 (v1) (834 images)			Cat 2 (v2) (901 images)		
	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑
HyperNeRF	.634	12.84	.673	.432	14.27	.721	.521	14.86	.632	.438	14.87	.597	.641	12.32	.632	.397	15.68	.657
D^2NeRF	.540	13.37	.694	.546	11.74	.685	.687	10.92	.545	.588	11.88	.548	.556	12.55	.664	.595	12.71	.604
HyperNeRF (w/ depth)	.373	16.86	.730	.425	16.95	.740	.532	14.37	.621	.371	15.65	.617	.330	18.47	.728	.376	16.56	.670
D^2NeRF (w/ depth)	.507	13.44	.698	.532	11.88	.690	.685	10.81	.534	.580	12.00	.563	.561	12.59	.656	.553	12.76	.629
Total-Recon (w/ depth)	.271	17.60	.745	.313	17.78	.768	.382	15.77	.657	.333	16.44	.652	.237	21.22	.793	.281	18.52	.713
Deformable-gs (w/ depth)	.520	12.35	.432	.490	12.78	.450	.565	11.92	.398	.530	12.30	.410	.600	11.50	.380	.510	12.60	.420
4DGS (w/ depth)	.525	12.40	.425	.495	12.65	.445	.570	11.85	.390	.535	12.25	.415	.605	11.45	.375	.515	12.55	.430
GS-marble	−−	OOM	−−	.530	12.45	.430	−−	OOM	−−	−−	OOM	−−	−−	OOM	−−	−−	OOM	−−
MoSca	−−	OOM	−−	.312	19.95	.695	−−	OOM	−−	−−	OOM	−−	−−	OOM	−−	−−	OOM	−−
Shape-of-Motion	−−	OOM	−−	.282	20.85	.785	−−	OOM	−−	−−	OOM	−−	−−	OOM	−−	−−	OOM	−−
Ours	**.251**	**20.12**	**.825**	**.285**	**21.37**	**.791**	**.319**	**20.52**	**.711**	**.285**	**21.74**	**.693**	**.203**	**22.94**	**.693**	**.262**	**22.07**	**.763**

	Cat 3 (767 images)			Human 1 (550 images)			Human 2 (483 images)			Human - Dog (392 images)			Human - Cat (431 images)			Mean		
	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑
HyperNeRF	.592	13.74	.624	.632	11.94	.603	.585	14.97	.620	.487	15.04	.699	.462	13.52	.512	.531	14.00	.635
D^2NeRF	.759	11.03	.578	.588	11.88	.638	.630	12.13	.599	.576	12.41	.652	.628	10.41	.453	.611	11.97	.608
HyperNeRF (w/ depth)	.514	14.86	.635	.501	13.25	.664	.445	15.58	.665	.450	15.01	.704	.456	14.40	.535	.428	15.80	.667
D^2NeRF (w/ depth)	.730	11.08	.582	.585	12.14	.638	.609	12.11	.612	.608	12.30	.633	.645	10.51	.451	.599	12.02	.611
Total-Recon (w/ depth)	.261	19.89	.734	.213	18.39	.778	.264	16.73	.712	.256	16.69	.756	.233	17.67	.630	.278	18.11	.724
Deformable-gs (w/ depth)	.550	12.45	.410	.505	12.80	.430	.560	11.95	.400	.540	12.10	.420	.590	11.70	.390	.542	12.22	.413
4DGS (w/ depth)	.545	12.50	.415	.510	12.75	.435	.565	11.90	.405	.535	12.15	.425	.595	11.65	.385	.545	12.19	.413
GS-marble	−−	OOM	−−	.548	12.50	.415	.555	12.08	.405	.538	12.32	.418	.580	11.85	.399	−−	NA	−−
MoSca	−−	OOM	−−	−−	OOM	−−	.263	18.15	.711	**.241**	**21.10**	**.781**	.243	19.05	.730	−−	NA	−−
Shape-of-Motion	−−	OOM	−−	.214	18.45	.776	.262	16.78	.715	.253	16.75	.758	.235	17.55	.635	−−	NA	−−
Ours	**.247**	**20.50**	**.744**	**.211**	**20.19**	**.782**	**.251**	**18.78**	**.725**	.247	20.56	.776	**.229**	**21.34**	**.688**	**.263**	**21.31**	**.747**

Table 2. **Quantitative Comparisons on Novel View Synthesis (Visual Metrics)**. We compare our method to previous dynamic NVS works and their depth-supervised variants on the 11 sequences of our stereo RGB dataset in terms of LPIPS, PSNR, and SSIM. Our method significantly outperforms all baselines for all sequences.

	Dog 1		Dog 1 (v2)		Cat 1		Cat 1 (v2)		Cat 2		Cat 2 (v2)		Cat 3		Human 1		Human 2		Human - Dog		Human - Cat		Mean	
	Acc↑	ϵ_{depth}↓	Acc↑	ϵ_{depth}↓	Acc↑	ϵ_{depth}↓	Acc↑	ϵ_{depth}↓	Acc↑	ϵ_{depth}↓	Acc↑	ϵ_{depth}↓	Acc↑	ϵ_{depth}↓	Acc↑	ϵ_{depth}↓	Acc↑	ϵ_{depth}↓	Acc↑	ϵ_{depth}↓	Acc↑	ϵ_{depth}↓	Acc↑	ϵ_{depth}↓
HyperNeRF	.107	.687	.176	.870	.316	.476	.314	.564	.277	.765	.252	.811	.213	.800	.053	.821	.067	1.665	.072	.894	.162	.862	.198	.855
D^2NeRF	.219	.463	.220	.456	.346	.334	.403	.314	.333	.371	.339	.361	.231	.523	.066	1.063	.128	.890	.078	.847	.126	.880	.247	.739
HyperNeRF	.352	.331	.357	.338	.552	.206	.596	.209	.605	.154	.612	.170	.451	.285	.211	.591	.249	.611	.283	.565	.214	.613	.439	.374
D^2NeRF	.338	.423	.270	.445	.510	.325	.362	.313	.438	.298	.376	.318	.243	.496	.086	.984	.131	.813	.154	.789	.176	.757	.302	.549
Total-Recon	.841	.165	.790	.167	**.889**	**.184**	.894	.124	.967	.050	.925	.081	.949	.066	.909	.142	.849	.142	.827	.204	.914	.104	.895	.131
Def.GS	.172	.599	.183	.612	.320	.415	.328	.432	.295	.485	.271	.494	.225	.598	.070	.912	.109	.940	.085	.862	.145	.795	.215	.632
4DGS	.175	.603	.178	.620	.315	.423	.325	.436	.292	.481	.268	.499	.232	.592	.073	.908	.113	.936	.089	.859	.142	.802	.200	.651
GS-marble	−−	OOM	.180	.615	−−	OOM	−−	OOM	−−	OOM	−−	OOM	−−	OOM	.175	.710	.210	.838	.187	.801	.143	.799	−−	NA
MoSca	−−	OOM	.792	.165	−−	OOM	−−	OOM	−−	OOM	−−	OOM	−−	OOM	−−	OOM	.850	.141	.826	.205	.912	.106	−−	NA
S.o.M	−−	OOM	.788	.168	−−	OOM	−−	OOM	−−	OOM	−−	OOM	−−	OOM	.908	.144	.845	.145	.825	.206	.911	.108	−−	NA
Ours	**.845**	**.160**	**.795**	**.163**	.880	.190	**.898**	**.122**	**.970**	**.048**	**.928**	**.079**	**.955**	**.064**	**.915**	**.138**	**.855**	**.139**	**.830**	**.202**	**.920**	**.102**	**.901**	**.127**

Table 3. **Quantitative Comparisons on Novel View Synthesis (Depth Metrics)**. We compare HoliGS to previous works on the Total-Recon dataset in terms of the average accuracy at 0.1m (Acc@0.1m) and the RMS depth error ϵ_{depth} (units: meters). Our method significantly outperforms all baselines for all sequences.

metrics (Acc@0.1m and RMS depth error). Our method consistently outperforms the baselines in both sets of metrics.

Table 4 evaluates the contribution of each deform component systematically removing key elements: the depth supervision, the normal supervision, the deformation field F^t, soft deformation S, pose initialization from external estimators, and the rigid transformation T_j^t, where j identifies a deformable object. For all ablations, we maintain the same core optimization objectives used in our full method while initializing camera parameters T_b^t from device sensors. For configurations without rigid body modeling, we initialize each object's pose with predictions from PoseNet and optimize them during reconstruction; for row 6, we replace these predictions with identity transformations.

Geometric supervision. Table 4 demonstrates that removing depth supervision (row 2) significantly reduces average accuracy. Ablation visualization (in Appendix) reveals

Methods	Depth Loss	Normal Loss	Deform. Obj.	Root Init.	Root Motion	Deform. Soft	LPIPS↓	Acc@0.1m↑
(1) **Full model**	✓	✓	✓	✓	✓	✓	**.263**	**.896**
(2) w/o loss $\mathcal{L}_{\text{depth}}$	✗	✓	✓	✓	✓	✓	.385	.847
(3) w/o loss $\mathcal{L}_{\text{normal}}$	✓	✗	✓	✓	✓	✓	.288	.832
(4) w/o deform. $\mathbf{J}_j$	✓	✓	✗	✓	✓	✓	.305	.853
(5) w/o Soft deform $\mathbf{G}_j$	✓	✓	✓	✗	✓	✓	.293	.870
(6) w/o root-body init.	✓	✓	✓	✗	✓	✗	.301	.862
(7) w/o root-body $\mathbf{G}_j$	✓	✓	✓	✗	✗	✓	N/A	N/A

Table 4. **Ablation Study.** Removing depth supervision (2) significantly hurts performance, while removing the deformation field (3) and PoseNet-initialization of root-body poses (4) hurts moderately. Most importantly, removing root-body poses entirely (5) prevents convergence (N/A) as the deformation field alone has to explain *global* object motion (see Figure 4). These experiments justify our hierarchical modeling of motion, as even root-bodies without a deformation field (3) or poorly initialized root-bodies (4) are better than no root-bodies (5). We visualize these ablations in Appendix.

that this stems from scale inconsistency between objects - while removing depth supervision does not severely impact training-view RGB renderings, it introduces critical failure modes in novel-view reconstructions: (a) floating foreground objects, evidenced by misaligned shadows, and (b) incorrect occlusion relationships between subjects. Without depth supervision, our method overfits to training perspectives and produces a degenerate scene representation where objects fail to maintain consistent scale relationships.

Similarly, our results show that normal supervision (row 3) provides crucial geometric guidance. Without normal constraints, the model struggles to capture fine surface details and produces less coherent object boundaries, particularly in regions with complex geometry. The normal supervision helps maintain surface continuity and improves the definition of sharp features.

Deformation modeling. Table 4 indicates that eliminating the deformation field (row 4) substantially degrades performance. Without this component, our approach must explain non-rigid motion using only rigid transformations, resulting in coarse approximations that fail to capture articulated movements like limb motion. The MLP-based soft deformation component (row 5) further enhances our model's ability to represent complex non-rigid movements through the transformation (1).

Similar to established approaches, our method enables bidirectional warping, with the inverse transformation defined as (2). This hierarchical structure allows our model to handle both global positioning and local deformations effectively. Removing the neural soft deformation component results in notable artifacts around joints and other highly articulated regions.

Removing pose initialization from external networks (row 6) produces similarly detrimental effects, leading to noisy appearance and geometry artifacts. Most significantly, Table 4 shows that eliminating object-specific rigid transformations entirely (row 7) causes optimization failure (N/A), even though the deformation field and soft deformation components can theoretically represent all continuous motion. It proves challenging for deformation fields alone to model global positioning, as such movements can deviate substantially from canonical configurations, complicating convergence. These findings justify our hierarchical motion representation, which explicitly models object positioning through rigid transformations while capturing non-rigid deformations through a combination of MLPs. Our ablations further suggest that the underwhelming performance of baseline methods on challenging dynamic scenes may stem from insufficient object-centric motion modeling.

In this work, we have presented a novel approach for embodied view synthesis from monocular RGB videos, with a particular focus on dynamic scenes featuring humans interacting with animals. Our primary technical contribution is a deformable Gaussian splatting framework that hierarchically decomposes scene dynamics into object-level motions, which are further decomposed into rigid transformations and localized deformations. This hierarchical structure enables effective initialization of object poses and facilitates optimization over long sequences with significant motion.

Future Work. We aim to integrate event-aware sensors (e.g., event cameras or high-frame-rate IMUs) to better capture motion discontinuities. We also plan to couple the warping network with a lightweight, on-the-fly bootstrap module that refines pose and Gaussian splitting priors across diverse articulated objects, including humans, animals, and furniture. To support real-time embodied view synthesis on mobile platforms, we will improve our splitting kernels and memory layout for deployment on AR glasses and edge devices.

Limitations. Despite the demonstrated effectiveness of our approach, our generic pose estimation sometimes mis-match the anatomical accuracy of specialized parametric models such as SMPL[32] for humans, which offer more robust initializations and appropriate joint constraints.

References

[1] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O'Toole, and Changil Kim. Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. In *CVPR*, 2023. 2

[2] Marc Badger, Yufu Wang, Adarsh Modh, Ammon Perkes, Nikos Kolotouros, Bernd G Pfrommer, Marc F Schmidt, and Kostas Daniilidis. 3d bird reconstruction: a dataset, model, and shape recovery from a single view. In *ECCV*, 2020. 3

[3] Aayush Bansal, Minh Vo, Yaser Sheikh, Deva Ramanan, and Srinivasa Narasimhan. 4d visualization of dynamic events from unconstrained multi-view videos. In *CVPR*, 2020. 2

[4] Mojtaba Bemana, Karol Myszkowski, Hans-Peter Seidel,

and Tobias Ritschel. X-fields: Implicit neural view-, light- and time-image interpolation. In *SIGGRAPH Asia*, 2020. 2

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

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

[7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *ECCV*, 2022. 2

[8] Xu Chen, Yufeng Zheng, Michael J Black, Otmar Hilliges, and Andreas Geiger. Snarf: Differentiable forward skinning for animating non-rigid neural implicit shapes. In *ICCV*, 2021. 3

[9] Xu Chen, Tianjian Jiang, Jie Song, Max Rietmann, Andreas Geiger, Michael J Black, and Otmar Hilliges. Fast-snarf: A fast deformer for articulated neural fields. *IEEE TPAMI*, 2023. 3

[10] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 3

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

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

[13] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *ICCV*, 2021. 2

[14] Hang Gao, Ruilong Li, Shubham Tulsiani, Bryan Russell, and Angjoo Kanazawa. Monocular dynamic view synthesis: A reality check. In *NeurIPS*, 2022. 3

[15] Shubham Goel, Georgios Pavlakos, Jathushan Rajasegaran, Angjoo Kanazawa, and Jitendra Malik. Humans in 4d: Reconstructing and tracking humans with transformers. *arXiv preprint arXiv:2305.20091*, 2023. 3

[16] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *CVPR*, 2018. 3

[17] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Cotracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024. 2

[18] Ladislav Kavan, Steven Collins, Jiří Žára, and Carol O'Sullivan. Skinning with dual quaternions. *SI3D*, 2007. 3

[19] Alex Kendall, Matthew Grimes, and Roberto Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *ICCV*, 2015. 4

[20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM TOG*, 2023. 1, 2, 5

[21] Nikos Kolotouros, Georgios Pavlakos, Michael J Black, and Kostas Daniilidis. Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In *ICCV*, 2019. 3

[22] Johannes Kopf, Xuejian Rong, and Jia-Bin Huang. Robust consistent video depth estimation. In *CVPR*, 2021. 3

[23] Jiahui Lei, Yijia Weng, Adam Harley, Leonidas Guibas, and Kostas Daniilidis. Mosca: Dynamic gaussian fusion from casual videos via 4d motion scaffolds. *arXiv preprint arXiv:2405.17421*, 2024. 2

[24] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *CVPR*, 2022. 2

[25] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, 2021. 2

[26] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based rendering. In *CVPR*, 2023. 3

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

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

[29] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM TOG*, 2021. 3

[30] Qingming Liu, Yuan Liu, Jiepeng Wang, Xianqiang Lyv, Peng Wang, Wenping Wang, and Junhui Hou. Modgs: Dynamic gaussian splatting from casually-captured monocular videos. In *ICLR*, 2025. 2

[31] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *CVPR*, 2023. 2

[32] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*. 2023. 8

[33] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: A skinned multi-person linear model. *ACM TOG*, 2023. 3

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

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

[36] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021. 1

[37] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Transactions on Robotics*, 2015. 3

[38] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021. 2

[39] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM TOG*, 2021. 2

[40] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. In *CVPR*, 2021. 3

[41] Luigi Piccinelli, Yung-Hsu Yang, Christos Sakaridis, Mattia Segu, Siyuan Li, Luc Van Gool, and Fisher Yu. Unidepth: Universal monocular metric depth estimation. In *CVPR*, 2024. 4

[42] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 2021. 2

[43] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 5

[44] Nadine Rüegg, Shashank Tripathi, Konrad Schindler, Michael J Black, and Silvia Zuffi. Bite: Beyond priors for improved three-d dog pose estimation. In *CVPR*, 2023. 3

[45] Sara Fridovich-Keil and Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, 2022. 2

[46] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016. 3

[47] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *CVPR*, 2023. 2

[48] Chonghyuk Song, Gengshan Yang, Kangle Deng, Jun-Yan Zhu, and Deva Ramanan. Total-recon: Deformable scene reconstruction for embodied view synthesis. In *ICCV*, 2023. 1, 2, 3, 5

[49] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE TVCG*, 2023. 2

[50] Colton Stearns, Adam Harley, Mikaela Uy, Florian Dubost, Federico Tombari, Gordon Wetzstein, and Leonidas Guibas. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In *SIGGRAPH Asia*, 2024. 2

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

[52] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *ECCV*, 2020. 5

[53] Zachary Teed and Jia Deng. DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. In *NeurIPS*, 2021. 3

[54] Fengrui Tian, Shaoyi Du, and Yueqi Duan. Mononerf: Learning a generalizable dynamic radiance field from monocular videos. In *ICCV*, 2023. 2

[55] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Nonrigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *ICCV*, 2021.

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

[57] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking everything everywhere all at once. In *ICCV*, 2023. 4

[58] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of motion: 4d reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024. 2

[59] Yufu Wang and Kostas Daniilidis. Refit: Recurrent fitting network for 3d human recovery. In *ICCV*, 2023. 3

[60] Yikai Wang, Xinzhou Wang, Zilong Chen, Zhengyi Wang, Fuchun Sun, and Jun Zhu. Vidu4d: Single generated video to high-fidelity 4d reconstruction with dynamic gaussian surfels. In *NeurIPS*, 2024. 3

[61] Chung-Yi Weng, Brian Curless, Pratul P Srinivasan, Jonathan T Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. In *CVPR*, 2022. 3

[62] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *CVPR*, 2024. 5

[63] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. D2 nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. *arXiv preprint arXiv:2205.15838*, 2022. 2

[64] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *CVPR*, 2021. 2

[65] Gengshan Yang, Minh Vo, Natalia Neverova, Deva Ramanan, Andrea Vedaldi, and Hanbyul Joo. Banmo: Building animatable 3d neural models from many casual videos. In *CVPR*, 2022. 3, 4

[66] Gengshan Yang, Chaoyang Wang, N Dinesh Reddy, and Deva Ramanan. Reconstructing animatable categories from videos. In *CVPR*, 2023. 3

[67] Gengshan Yang, Shuo Yang, John Z Zhang, Zachary Manchester, and Deva Ramanan. Ppr: Physically plausible reconstruction from monocular videos. In *ICCV*, 2023. 2

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

[69] Ziyi Yang, Xinyu Gao, Wen Zhou, Shaohui Jiao, Yuqing Zhang, and Xiaogang Jin. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *CVPR*, 2024. 5

[70] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. In *CVPR*, 2020. 2

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

[72] Zhoutong Zhang, Forrester Cole, Zhengqi Li, Michael Rubinstein, Noah Snavely, and William T. Freeman. Structure and motion from casual videos. In *ECCV*, 2022. 3

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

[74] Silvia Zuffi, Angjoo Kanazawa, David Jacobs, and Michael J. Black. 3D menagerie: Modeling the 3D shape and pose of animals. In *CVPR*, 2017. 3