Completing Dynamic Human Reconstruction with I2V-Synthesized Views

Zhewen Zheng, Changwei Yao and Aman Goel
Carnegie Mellon University

Abstract

Reconstructing dynamic human avatars from monocular video remains a challenging problem due to severe depth ambiguity, self-occlusion, and inconsistent 2D point tracks under fast motion. We present a practical framework that combines generative view augmentation with human-aware 3D tracking to improve monocular 4D reconstruction quality. A commercial video diffusion model is used to generate a synthetic turnaround sequence conditioned on a pseudo-canonical frame, exposing geometry that is not visible in the original capture. We further incorporate 3D correspondences derived from the Momentum Human Rig (MHR), which provide stable, body-aware tracks that reduce drift and fragmentation compared to standard 2D point tracking. Using these improved priors within the Shape-of-Motion pipeline, our method produces more consistent reconstructions of articulated motion and reduces preprocessing time by an order of magnitude compared to TAPIR-based tracking. Experiments on DNA-Rendering sequences show measurable improvements in foreground PSNR and qualitative stability, demonstrating the effectiveness of combining generative augmentation with structured human priors for monocular 4D human reconstruction.

1 Introduction

Dynamic scenes are quite common in real scenarios, and recent works on 4D Gaussian Splatting [38] (GS) have enabled the representation and rendering of dynamic 3D scenes with persistent motions and geometry. With advances in dynamic reconstruction and novel view synthesis [48, 42, 43] (NVS), rendered dynamic scenes have become increasingly photorealistic and can handle more challenging situations, such as videos with heavy noise or drastic movement. However, because these methods still rely heavily on multi-view camera setups [6, 21] or auxiliary sensor inputs [19, 9], reconstructing a human avatar with plausible quality solely from a monocular video remains an open challenge. Although recent monocular 4D reconstruction methods can operate on regular dynamic scenes, either by modeling the scene as deformation fields between canonical and view space [28] or by using low-dimensional $SE(3)$ motion bases [39], they struggle to reconstruct human avatars exhibiting drastic motion in novel views due to the inherently ill-posed nature of the task.

The longstanding challenge for general in-the-wild videos lies in the poorly constrained nature of the 4D reconstruction problem. The task of inferring dynamic 3D human avatars from a monocular input video is particularly ill-posed and self-occluded. Recent advances in novel-view synthesis [4] and 4D Gaussian Splat Segmentation [12] show promising results toward alleviating these difficulties. Motivated by this, our initial plan was to build upon these methods and propose a framework for reconstructing dynamic human avatars from monocular video using 4D GS. The idea was to leverage the NVS capabilities of CogNVS [39] to produce multiple new viewpoints, pass these synthesized views to DiffuMan [13] for dense view generation, and then train a 4D GS model using the resulting dense set of frames. However, after several attempts, we encountered practical obstacles including non-open-source components, incompatible dataset requirements, and limited computational resources. Consequently, we shifted our focus to a more feasible objective: reconstructing human avatars with rapid motion from monocular video in novel views. Our approach is inspired by two insights. First, pretrained vision-language models (VLMs) can generate videos from static images using text instructions, providing additional views that help the model perceive full geometry and texture. Second, while data-driven priors are powerful, they introduce noise that becomes more severe with rapid movement. To address this, we incorporate 3D point tracks to enforce a globally coherent representation of scene geometry and motion.

To this end, based on these two insights, we propose a framework that augments the monocular video using generative priors and integrates 3D point tracking into a Shape-of-Motion (SoM) formulation. We conduct extensive evaluations comparing our method with SoM, demonstrating that our approach consistently outperforms SOM, especially in complex scenarios involving rapid, large-amplitude human motion. Our contributions can be summarized as follows:

- We introduce a generative augmentation strategy using off-the-shelf commercial large-scale video diffusion model to generate synthetic sequence to reveal geometry not visible in the original monocular capture.
- We integrate higher-quality human mesh priors from SAM-3D-Body, enabling more stable and body-aware correspondences while considerably reducing preprocessing time compared to TAPIR tracks.
- We propose a unified framework that combines generative augmentation and motion-aware tracking to reconstruct human avatars with drastic motion from monocular video.
- An empirical evaluation on several human performance sequences, demonstrating consistent improvements in PSNR and qualitative stability over SoM baseline in challenging scenarios with complex motion.

2 Related Work

Dynamic Scene Reconstruction from Monocular Video

Classic NeRF and 3D Gaussian Splatting variants[22, 14] extend to 4D by modeling per-frame deformation fields or time-conditioned radiance/opacity[41]. To ensure high-fidelity, most methods assume dense multi-view capture[33, 27] and do not extrapolate geometry/appearance in occluded regions. Recent monocular approaches [5, 37] inject additional structure and priors to overcome this limitation. An example of leveraging a geometric pipeline with diffusion priors is CogNVS [5], which first obtains an initial dynamic reconstruction using off-the-shelf 4D SLAM methods [17], renders novel views to expose unseen areas, and then trains a self-supervised conditional diffusion model with test-time finetuning to in-paint missing content. Another work, SoM [36] represents scene motion with a compact set of $\mathbf{SE}(3)$ bases to softly group rigid motion, and integrate noisy monocular depth and long range 2D tracks into a globally consistent dynamic representation. However, none of the above focuses on handling 4D human reconstruction and treats human pose dynamics as a part of the overall scene composition.

Human-Centric 4D Reconstruction

The abundance in derivative works around dynamic neural fields thus reinvigorated the world of human avatar representations, leading to works in sparse-view 4D which sought to enhance SMPL [20, 29] representations with neural primitives as appearance. One recent 3DGS-based human reconstruction method, GauHuman, [10] initializes canonical-space Gaussians with SMPL human priors, then applies Linear Blend Skinning (LBS) to refine pose and other fine details efficiently. This yields fast optimization and explicit rigging, making re-animation possible. Yet fine-grained clothing and overall appearance still lag behind the fidelity achieved by 4DGS with dense-view supervision, and background inpainting was not in scope [10]. Aside from neural primitives, numerous methods attempt to incorporate rich 2D priors; recent work DressRecon targets in-the-wild monocular videos of human motion and explicitly separates body and clothing dynamics within a hierarchical Gaussian motion field. Body Gaussians are initialized from a 3D pose prior to robustly handle large limb motions, while clothing is modeled by decomposed motion layers to capture non-rigid garments [34]. To stablize optimization, DressRecon uses image-based priors, employing [16] for accurate silhouettes and temporally consistent body mask, and deep features [26] for registering pixels to 3D model. The result is time-varying 3D humans in realistic loose clothing with strong monocular robustness. It is better in fidelity compared to GauHuman, though it does not provide the direct SMPL-driven re-pose-ability like GauHuman does. In our work, to ensure precise and temporally consistent limb reconstruction across consecutive frames, we use the MHR estimates provided by SAM-3D-Body as a conditioning constraint, which substantially mitigates the issue of missing limbs.

Novel View Synthesis

Novel view synthesis has progressed rapidly with the development of implicit neural representations such as NeRFs [23] and Gaussian-based primitives [15]. Considerable effort has been devoted to scaling these representations to larger and more complex environments [2, 35], accelerating optimization [24, 3], improving anti-aliasing [1], and extending them to dynamic scenes [31, 30]. Dynamic NeRF variants [31, 28] and deformable Gaussian primitives [38] have become dominant paradigms for modeling non-rigid motion, complemented by voxel-based methods [7] and tokenized representations [44]. While these approaches produce high-quality renderings, they typically require multi-view, fully-posed input videos, and only recently has monocular novel-view synthesis begun to gain traction [11, 39, 45]. Existing monocular NVS systems generally rely on test-time optimization for each input video, which is computationally expensive and still struggles to recover fine-grained dynamic detail [11]. Moreover, these methods focus primarily on reconstructing co-visible content, a limitation reinforced by evaluation protocols that only measure accuracy on overlapping pixels on training and inference views [25]. Our method is focused on casually-captured monocular videos, a more practical and challenging setup, using strong off-the-shelf priors.

3 Preliminaries

3D Gaussian Splatting. We leverage 3DGS [15] to represent a scene as a set of anisotropic 3D Gaussians $\mathcal{G} = \{G_i\}_{i=1}^{N}$, where each Gaussian G_i is defined by a mean position $\boldsymbol{\mu}_i \in \mathbb{R}^3$, a covariance matrix $\boldsymbol{\Sigma}_i \in \mathbb{R}^{3\times 3}$, an opacity coefficient α_i, and a color modeled using spherical harmonics (SH) coefficients $\mathbf{c}_i$. The Gaussian density at a 3D point $\mathbf{x}$ is given by:

$$\rho_i(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^\top \boldsymbol{\Sigma}_i^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right). \quad (1)$$

During rendering, each 3D Gaussian is projected onto the image plane as a 2D ellipse through the camera projection model. The covariance is transformed via the Jacobian $\mathbf{J}$ of the projection function:

$$\boldsymbol{\Sigma}_{i,\text{2D}} = \mathbf{J}\,\boldsymbol{\Sigma}_i\,\mathbf{J}^\top, \quad (2)$$

yielding a 2D Gaussian used for rasterization. A differentiable splatting process composites the Gaussians front-to-back using α-blending:

$$\mathbf{C}(\mathbf{p}) = \sum_{i=1}^{N} T_i(\mathbf{p})\,\alpha_i\,\mathbf{c}_i(\mathbf{d}), \quad (3)$$

where $\mathbf{p}$ is the pixel location, $\mathbf{d}$ is the viewing direction, and T_i denotes the transmittance from previously accumulated Gaussians.

Optimizing the scene involves updating Gaussian parameters $\{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i, \alpha_i, \mathbf{c}_i\}$ using gradient-based learning. This enables 3DGS to capture fine geometry and appearance while rendering at real-time frame rates due to its analytical rasterization pipeline. The flexibility of Gaussian primitives also facilitates extensions to dynamic or articulated scenes, where parameters evolve over time or are driven by a deformation model.

Dynamic Scene Representation. We follow SoM and adopt a canonical dynamic scene representation, where a global set of 3D Gaussians is defined in a canonical reference frame. Each Gaussian maintains time-invariant canonical parameters (μ_0, R_0, s, o, c), and its configuration at time t is obtained through a frame-dependent rigid transform in $SE(3)$:

$$T_{0\to t} = \begin{bmatrix} R_{0\to t} & t_{0\to t} \end{bmatrix}, \qquad \mu_t = R_{0\to t}\mu_0 + t_{0\to t}, \quad R_t = R_{0\to t}R_0. \tag{4}$$

Canonical Dynamic Representation. Instead of independently estimating the full motion trajectory for every Gaussian, the scene motion is parameterized by a compact set of shared $SE(3)$ motion bases $\{T_{0\to t}^{(b)}\}_{b=1}^{B}$. Each Gaussian is assigned motion coefficients $w^{(b)}$ with $\|w\|_1 = 1$, and its overall motion is expressed as

$$T_{0\to t} = \sum_{b=1}^{B} w^{(b)} T_{0\to t}^{(b)}. \tag{5}$$

This factorization captures low-dimensional motion structure, regularizes trajectories, and enforces global temporal coherence across the scene.

Rasterizing 3D Trajectories. Given the dynamic Gaussian representation, dense 3D trajectories can be queried between arbitrary frames. For a pixel p in frame t, let $H(p)$ denote intersecting Gaussians and α_i their transmittance. Its 3D coordinate transported to frame t' is computed as

$$\mathbf{X}_{t\to t'}(p) = \sum_{i\in H(p)} T_{0\to t'}\, \mu_{i,0}\, \alpha_i. \tag{6}$$

Using camera intrinsics $K_{t'}$ and extrinsics $E_{t'}$, the corresponding pixel and depth in frame t' are obtained via

$$\mathbf{U}_{t\to t'}(p) = \Pi(K_{t'} E_{t'} \mathbf{X}_{t\to t'}(p)), \qquad \mathbf{D}_{t\to t'}(p) = (E_{t'}\mathbf{X}_{t\to t'}(p))_z. \tag{7}$$

This rasterization formulation provides consistent long-range 3D motion queries and supports tracking through occlusions, forming the basis of dynamic reconstruction.

4 Method

We follow the overall formulation of SoM and retain its decomposition of monocular 4D reconstruction into appearance modeling, geometric supervision, and motion trajectory estimation. Let $V = \{I_t\}_{t=1}^{T}$ denote an input monocular video. SoM represents the dynamic scene using a set of 3D Gaussians and learns a deformation field that maps canonical points to their time-varying positions. While the original framework assumes that all relevant surface regions appear in the input sequence and that depth and long-term correspondences are sufficiently reliable, these assumptions break down for human videos with large articulated motion. We introduce several modifications to improve the stability of SoM in such settings, while keeping its core optimization unchanged.

4.1 Video Augmentation

In human-centric sequences, large portions of the subject (e.g., the back, hips, or underarms) may never be observed due to persistent self-occlusion. This lack of observability causes the appearance optimization in SoM to collapse on unobserved regions. To increase viewpoint diversity while preserving the original motion signals, we augment the input sequence using a video synthesized by the Gemini generative model [8]. Using the first high-quality frame I_1 as input, Gemini produces a synthetic rotation sequence V_{rot} in which the avatar rotates once around its vertical axis. To avoid introducing temporal discontinuities, the generated sequence is reversed and concatenated in a palindromic fashion:

$$V_{\text{aug}} = \big[\, \text{reverse}(V_{\text{rot}}) \,\|\, V \,\big]. \tag{8}$$

Figure 1: Failure case of the Shape-of-Motion (SOM) baseline on a fast-motion sequence. The reconstructed geometry exhibits limb tearing and temporal jitter, especially around the arms and legs, due to unstable 2D point tracks and missing observations in the monocular input.

The augmented frames are used only for the appearance and density components of the Gaussian representation, whereas the deformation field and motion trajectories are supervised exclusively on the original frames. In this way, the augmentation increases the visibility of surface regions while avoiding interference with the true dynamics of the scene.

4.2 Depth Estimation

Depth plays a central role in stabilizing SoM's geometry estimation through a depth reconstruction loss between rendered depth and externally predicted depth. The original method employs Depth Anything [40] (DA-2) for monocular depth estimation. However, we observe that DA-2 introduces significant high-frequency noise for fast-moving human limbs, producing depth distributions inconsistent with the rendered geometry and leading to noticeable artifacts in the learned deformation field. We replace this depth predictor with Depth Anything 3 [18] (DA-3), which yields substantially smoother and more stable depth estimates. For each frame I_t, we obtain a depth map

$$D_t^{\mathrm{da3}} = \mathrm{DepthAnything3}(I_t), \tag{9}$$

and retain SoM's depth loss formulation:

$$\mathcal{L}_{\mathrm{depth}} = \sum_t \|D_t^{\mathrm{render}} - D_t^{\mathrm{da3}}\|_1. \tag{10}$$

The improvement in depth quality reduces spurious gradients on articulated regions and ensures that geometric supervision remains consistent with the underlying surface during large motions.

4.3 Mesh-Based 3D Tracks

SoM relies on TAPIR-generated 2D tracks as long-range correspondences. In human videos containing rapid articulation, these 2D tracks frequently fail on limb endpoints due to self-occlusions and appearance changes, even when sampling density is increased. As a result, track-based motion supervision becomes unstable. We replace 2D tracks with mesh-based 3D tracks derived from SAM-3D-Body MHR predictions.

For each frame, SAM-3D provides an estimated Momentum Human Rig (MHR) output separable into shape + scale parameters and joint pose parameters. We observe that the predicted shape varies inconsistently across frames due to the ill-posed nature of monocular estimation, leading to frame-dependent mesh distortions. To avoid these fluctuations, we take the MHR model predicted in the first frame as a canonical template M_1 and keep its shape parameters fixed for the entire sequence. For all subsequent frames, we retain only the pose parameters θ_t predicted by SAM-3D-Body and retarget them onto the canonical mesh:

$$M_t = \mathrm{MHR}(\beta_1, \theta_t), \tag{11}$$

where β_1 denotes the fixed shape from the first frame. This retargeting step eliminates frame-to-frame shape noise while preserving the temporal variation of body articulation.

Even with this constraint, the estimated pose occasionally exhibits incorrect limb orientations. To further correct these inconsistencies, we enforce alignment between the depth rendered from the human mesh and the DA-3 depth:

$$\mathcal{L}_{\text{align}} = \|D_t^{\text{mhr}} - D_t^{\text{da3}}\|_1. \tag{12}$$

After refinement, we sample a set of mesh vertices as 3D track anchors $\{\mathbf{x}_t^k\}$. These tracks supervise the deformation field in place of SoM's 2D correspondences:

$$\mathcal{L}_{\text{track}} = \sum_{k,t} w_t \|T_t(\mathbf{x}_0^k) - \mathbf{x}_t^k\|_2, \tag{13}$$

where T_t is the learned deformation mapping from the canonical space, w_t denote the weight associated with each track. In our framework, these weights are further modulated based on the corresponding body part. Specifically, tracks corresponding to the limbs are generally assigned lower weights, as limbs typically exhibit faster motion, which results in lower detection confidence and consequently reduced reliability of the associated tracks. This modification provides more reliable trajectory supervision for articulated limbs and significantly improves stability in high-motion regions.

4.4 Optimization

The final training objective preserves the original structure of SoM, combining photometric reconstruction, depth consistency, motion smoothness, and trajectory alignment. The only changes lie in the sources of depth and track supervision:

$$\mathcal{L} = \mathcal{L}_{\text{recon}} + \lambda_{\text{track}}\mathcal{L}_{\text{track}} + \lambda_{\text{rigidity}}\mathcal{L}_{\text{rigidity}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}} \tag{14}$$

All rendering, deformation, and Gaussian parameter updates follow SoM without modification.

5 Experiments

Datasets & Baselines

All experiments were conducted on a single NVIDIA RTX 4090 GPU. We evaluate our method on a subset of the DNA-Rendering 4K4D dataset, which provides multi-view high-resolution recordings of human actors performing dynamic motions. To simulate monocular capture, we use only camera 25, a fixed viewpoint present in all sequences. Two sequences were selected to cover a range of human articulations and velocities: 0012_11, 0013_01

The original 4K frames are downsampled to 720 px width and symmetrically padded to a 720×1280 resolution to satisfy Veo 3.1 input requirements. After synthetic video generation (described below), all frames are symmetrically cropped to 720×860 to remove black bars before feeding them into

Table 1: Quantitative comparison with Shape-of-Motion on the DNA-Rendering sequences. We report PSNR (dB) on foreground (FG) and background (BG) regions separately. Higher is better.

Sequence	SOM (Baseline)		Ours	
	FG ↑	BG ↑	FG ↑	BG ↑
0012_11	24.76	37.04	25.81	39.86
0013_01	19.63	34.87	21.32	34.30
Mean	22.20	36.96	23.57	37.08

Shape-of-Motion and Depth Anything 3. We take every other frame which amounts to 120 frames per sequence for faster evaluation.

To compensate for missing back-facing geometry, we append 4 seconds of synthetic frames generated by Veo 3.1. The model is conditioned on the pseudo-canonical frame, which is used as both the start and end conditioning frame. For each sequence, Veo produces four candidate 720×1280 videos under seed 42, and we select the most physically plausible one without visible artifacts.

To ensure extrinsic consistency with the fixed camera in DNA-Rendering, we apply the following prompt:

"A single character stands in place and slowly rotates 360 degrees to show all sides of their body. The character turns in place; the camera remains perfectly static and centered. Neutral studio background, consistent lighting, no zoom, no camera movement."

The real and synthetic frames are concatenated to form the 12-second training sequence.

We replace TAPIR with SAM-3D-Body (dinov3 variant), which uses SAM2 for human segmentation masks and mask tracking. SAM-3D-Body provides body-aware correspondences and joint parameters for each frame. To improve temporal consistency, we apply a simple retargeting strategy: the shape and scale parameters are extracted from the pseudo-canonical frame, and per-frame joint parameters are used to track articulated motion across the sequence. This procedure is not part of the original SOM pipeline and is introduced as a lightweight human-aware correction mechanism.

Evaluation Metrics

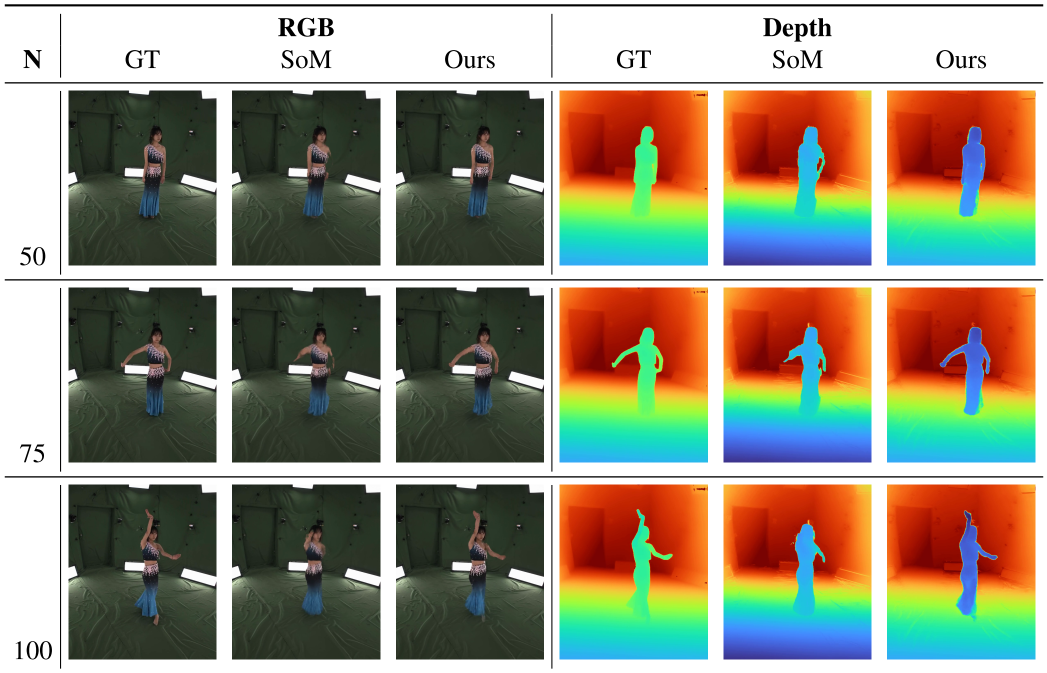
Evaluation: NVS quality (PSNR, SSIM [47], LPIPS [46]) against ground truth if available, re-pose performance (re-pose/pose-transfer across avatar and compare renders), segment/composting quality (IoU [32]), time cost, ablations.

Across all evaluated sequences, our method shows consistent improvements in foreground PSNR compared to the TAPIR-based SOM baseline. For example, on sequence 0013_01, the foreground PSNR increases from 19.63 dB to 21.32 dB, indicating a substantially more stable reconstruction of articulated body parts and fine-grained motion. This improvement is expected to be similar across other sequences, as the primary source of error in monocular human reconstruction arises from inconsistent or drifting point tracks on limbs and extremities.

TAPIR often produces track fragmentation or jitter under large, fast motions, and it requires over an hour of preprocessing per sequence. In contrast, our approach initializes 3D tracks using mesh-prior estimates from SAM-3D-Body, which provide coherent joint locations and body-aware correspondences. These priors offer significantly better temporal stability, especially on limbs undergoing rapid motion, and they reduce preprocessing time to under ten minutes. As a result, the reconstructed foreground exhibits more consistent geometry and smoother motion trajectories.

Background PSNR remains largely unchanged between the baseline and our method, which is expected because background regions are static across time and already well-handled by SOM. The

Figure 2: Qualitative results.



improvements primarily stem from better handling of articulated human motion, confirming that integrating a structured human prior directly benefits monocular 4D reconstruction.

6 Conclusion

We presented a practical approach for improving monocular 4D human reconstruction by combining generative view augmentation with human-aware 3D tracking. By introducing synthetic turnaround frames from a video diffusion model and substituting TAPIR with SAM-3D-Body correspondences, our method provides Shape-of-Motion with more complete geometric cues and temporally coherent tracks for articulated motion. The resulting reconstructions exhibit more stable limb dynamics, improved foreground fidelity, and significantly reduced preprocessing time. Although the improvements are modest in scale, they highlight the value of integrating strong human priors into general-purpose dynamic scene reconstruction pipelines.

There remain several promising directions for future work. First, tracking quality could be further improved by combining complementary sources of motion cues, such as merging SAM-3D-Body with TAPIR or MegaSaM to better handle fast or small-scale motions. Second, reconstruction efficiency may be improved by adaptively allocating computation to foreground regions once the background has converged. Finally, incorporating stronger geometric or generative priors, such as diffusion-based depth completion or multi-frame identity constraints, could further enhance reconstruction robustness in highly ambiguous monocular settings.

References

[1] Jonathan T Barron, Ben Mildenhall, et al. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *CVPR*, 2021.

[2] Jonathan T Barron, Ben Mildenhall, et al. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022.

[3] Anpei Chen and et al. Tensorf: Tensorial radiance fields. In *ECCV*, 2022.

[4] Kaihua Chen, Tarasha Khurana, and Deva Ramanan. Reconstruct, inpaint, finetune: Dynamic novel-view synthesis from monocular videos. *arXiv preprint arXiv:2507.12646*, 2025.

[5] Kaihua Chen, Tarasha Khurana, and Deva Ramanan. Reconstruct, Inpaint, Finetune: Dynamic Novel-view Synthesis from Monocular Videos, July 2025.

[6] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023.

[7] Sara Fridovich-Keil, Alex Yu, and et al. K-planes: Explicit radiance fields in space, time, and appearance. In *CVPR*, 2023.

[8] Google. Veo video generation in the gemini api. https://ai.google.dev/gemini-api/docs/video, 2025.

[9] Xiuye Gu, Yijie Wang, Chongruo Wu, Yong Jae Lee, and Panqu Wang. Hplflownet: Hierarchical permutohedral lattice flownet for scene flow estimation on large-scale point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3254–3263, 2019.

[10] Shoukang Hu and Ziwei Liu. GauHuman: Articulated gaussian splatting from monocular human videos. *arXiv preprint arXiv:*, 2023.

[11] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *ICCV*, 2021.

[12] Shengxiang Ji, Guanjun Wu, Jiemin Fang, Jiazhong Cen, Taoran Yi, Wenyu Liu, Qi Tian, and Xinggang Wang. Segment any 4d gaussians. *arXiv preprint arXiv:2407.04504*, 2024.

[13] Yudong Jin, Sida Peng, Xuan Wang, Tao Xie, Zhen Xu, Yifan Yang, Yujun Shen, Hujun Bao, and Xiaowei Zhou. Diffuman4d: 4d consistent human view synthesis from sparse-view videos with spatio-temporal diffusion models. *arXiv preprint arXiv:2507.13344*, 2025.

[14] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023.

[15] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *SIGGRAPH*, 2023.

[16] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv:2304.02643*, 2023.

[17] Zhengqi Li, Richard Tucker, Forrester Cole, Qianqian Wang, Linyi Jin, Vickie Ye, Angjoo Kanazawa, Aleksander Holynski, and Noah Snavely. MegaSaM: Accurate, Fast, and Robust Structure and Motion from Casual Dynamic Videos, December 2024.

[18] Haotong Lin, Sili Chen, Jun Hao Liew, Donny Y. Chen, Zhenyu Li, Guang Shi, Jiashi Feng, and Bingyi Kang. Depth anything 3: Recovering the visual space from any views. *arXiv preprint arXiv:2511.10647*, 2025.

[19] Xingyu Liu, Charles R Qi, and Leonidas J Guibas. Flownet3d: Learning scene flow in 3d point clouds. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 529–537, 2019.

[20] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Trans. Graphics (Proc. SIGGRAPH Asia)*, 34(6):248:1–248:16, October 2015.

[21] Jonathon Luiten, Georgios Kopanas, Bastian Leibe, and Deva Ramanan. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, pages 800–809. IEEE, 2024.

[22] 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. In *ECCV*, 2020.

[23] 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. In *ECCV*, 2020.

[24] Thomas Mueller, Oran Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. In *SIGGRAPH*, 2022.

[25] Michael Niemeyer, Ermano Lopes, Jonathan T Barron, and et al. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, 2022.

[26] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision, 2023.

[27] Linfei Pan, Daniel Barath, Marc Pollefeys, and Johannes Lutz Schönberger. Global Structure-from-Motion Revisited. In *European Conference on Computer Vision (ECCV)*, 2024.

[28] Keunhong Park et al. Nerfies: Deformable neural radiance fields. In *ICCV*, 2021.

[29] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3D hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 10975–10985, 2019.

[30] Sida Peng and et al. Neural sparse voxel fields. In *NeurIPS*, 2020.

[31] Albert Pumarola, Enric Corona, and et al. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 2021.

[32] Seyed Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian D. Reid, and Silvio Savarese. Generalized intersection over union: A metric and A loss for bounding box regression. *CoRR*, abs/1902.09630, 2019.

[33] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-Motion Revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[34] Jeff Tan, Donglai Xiang, Shubham Tulsiani, Deva Ramanan, and Gengshan Yang. DressRecon: Freeform 4D Human Reconstruction from Monocular Video, October 2024.

[35] Haithem Turki et al. Voxurf: Voxel-based efficient and accurate neural radiance fields. In *ICCV*, 2023.

[36] Qianqian Wang, Vickie Ye, Hang Gao, Jake Austin, Zhengqi Li, and Angjoo Kanazawa. Shape of Motion: 4D Reconstruction from a Single Video, July 2024.

[37] Zihan Wang, Jeff Tan, Tarasha Khurana, Neehar Peri, and Deva Ramanan. MonoFusion: Sparse-View 4D Reconstruction via Monocular Fusion, July 2025.

[38] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20310–20320, 2024.

[39] Yifan Xu and et al. Cognvs: A foundation model for novel view synthesis. In *CVPR*, 2024.

[40] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024.

[41] Zeyu Yang, Zijie Pan, Xiatian Zhu, Li Zhang, Jianfeng Feng, Yu-Gang Jiang, and Philip H. S. Torr. 4D Gaussian Splatting: Modeling Dynamic Scenes with Native 4D Primitives, August 2025.

[42] Tanveer Younis and Zhanglin Cheng. Sparse-view 3d reconstruction: Recent advances and open challenges. *arXiv preprint arXiv:2507.16406*, 2025.

[43] Yitian Yuan and Qianyue He. Efficient differentiable hardware rasterization for 3d gaussian splatting. *arXiv preprint arXiv:2505.18764*, 2025.

[44] Lunjun Zhang, Yuwen Xiong, Ze Yang, Sergio Casas, Rui Hu, and Raquel Urtasun. Copilot4d: Learning unsupervised world models for autonomous driving via discrete diffusion. *arXiv preprint arXiv:2311.01017*, 2023.

[45] Richard Zhang and et al. Mvimgnet: A large-scale dataset of multi-view images. *arXiv:2303.06163*, 2023.

[46] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*, abs/1801.03924, 2018.

[47] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. 13(4):600–612.

[48] Ziyue Zhu, Zhanqian Wu, Zhenxin Zhu, Lijun Zhou, Haiyang Sun, Bing Wan, Kun Ma, Guang Chen, Hangjun Ye, Jin Xie, et al. Worldsplat: Gaussian-centric feed-forward 4d scene generation for autonomous driving. *arXiv preprint arXiv:2509.23402*, 2025.