UAV4D: Dynamic Neural Rendering of Human-Centric UAV Imagery Using Gaussian Splatting

Jaehoon Choi[1*], **Dongki Jung**[1], **Chris Maxey**[1], **Sungmin Eum**[2], **Yonghan Lee**[1]
Dinesh Manocha[1], **Heesung Kwon**[2]

[1]University of Maryland, College Park, USA
[2]DEVCOM Army Research Laboratory, USA

Abstract

Despite significant advancements in dynamic neural rendering, existing methods fail to address the unique challenges posed by UAV-captured scenarios, particularly those involving monocular camera setups, top-down perspective, and multiple small, moving humans, which are not adequately represented in existing datasets. In this work, we introduce UAV4D, a framework for enabling photorealistic rendering for dynamic real-world scenes captured by UAVs. Specifically, we address the challenge of reconstructing dynamic scenes with multiple moving pedestrians from monocular video data without the need for additional sensors. We use a combination of a 3D foundation model and a human mesh reconstruction model to reconstruct both the scene background and humans. We propose a novel approach to resolve the scene scale ambiguity and place both humans and the scene in world coordinates by identifying human-scene contact points. Additionally, we exploit the SMPL model and background mesh to initialize Gaussian splats, enabling holistic scene rendering. We evaluated our method on three complex UAV-captured datasets: VisDrone, Manipal-UAV, and Okutama-Action, each with distinct characteristics and $10-50$ humans. Our results demonstrate the benefits of our approach over existing methods in novel view synthesis, achieving a 1.5 dB PSNR improvement and superior visual sharpness.

Code — https://jh-choi.github.io/uav4d/
Appendices — https://www.arxiv.org/abs/2506.05011

Introduction

Unmanned Aerial Vehicles (UAVs) have become essential tools for capturing large-scale environments such as streets and urban areas, supporting applications like world mapping, human activity analysis, and security monitoring. Recent advances in neural rendering (Kerbl et al. 2023) have enabled photorealistic scene reconstructions from UAV data, driving progress in large-scale mapping. However, most of the existing works (Lin et al. 2024a; Maxey et al. 2024a,b) primarily focus on static scenes, overlooking the more challenging dynamic scenes, even though real-world environments are inherently dynamic, involving human activities and complex background changes. Although dynamic neural rendering methods are rapidly evolving, most existing approaches (Zhu and Tang 2025) focus on phone-captured video scenes (Yang et al. 2024b; Wu et al. 2024; Stearns et al. 2024; Wang et al. 2024a), multi-view synchronized camera setups (Luiten et al. 2024; Xu et al. 2024), or driving scenario (Chen et al. 2025), as seen in Fig. 1. In particular, OmniRe (Chen et al. 2025) is designed for multi-sensor system scenarios, such as multi-camera rigs, a LiDAR sensor, and 3D bounding boxes. However, these setups are not accessible in UAV-based scenarios. In summary, dynamic neural rendering methods tailored for UAV-based scenarios remain underexplored. To the best of our knowledge, only a few works (Maxey et al. 2024a,b) have attempted this, and they have some limitations. Some approaches use implicit neural representations (Fridovich-Keil et al. 2023) for volumetric rendering to generate novel-view images but face challenges in terms of maintaining high-fidelity outputs.

Applying dynamic neural rendering to UAV-based scenes presents two fundamental challenges. First, due to hardware constraints, UAVs typically lack access to a diverse range of sensors and cannot utilize multi-view synchronized camera systems or multi-sensor configurations. Consequently, scene reconstruction must rely solely on monocular dynamic video, which presents substantial challenges (Gao et al. 2022) for both accurate geometry reconstruction and photorealistic rendering. Furthermore, most dynamic objects in UAV-captured scenes are relatively small and multiple moving humans often appear simultaneously. These conditions pose significant challenges, making it difficult to apply the state-of-the-art approaches (Zhu et al. 2024; Wang et al. 2024a; Zheng et al. 2025; Stearns et al. 2024) that rely on depth estimation (Yang et al. 2024a), optical flow (Xu et al. 2022a) and point tracking models (Karaev et al. 2024).

Even recent 4D reconstruction methods (Wang et al. 2025b; Feng et al. 2025) powered by 3D foundation models (Wang et al. 2024b) face difficulties when reconstructing small dynamic humans. This challenge arises primarily because existing large-scale datasets for dynamic scenes rarely include scenarios with small and moving human instances. OmniRe (Chen et al. 2025), tailored for driving scenes, leverages LiDAR and 3D bounding box information to freely localize dynamic foreground objects in world coordinates. However, in UAV-based scenarios where LiDAR

*Corresponding Author (kevchoi@umd.edu)

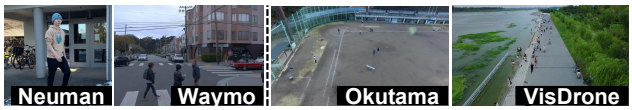


Figure 1: Comparison between existing benchmarks for dynamic neural rendering (columns 1–2) and UAV captured datasets (columns 3–4). UAVs typically cover wide areas from a top-down perspective.

is unavailable, human positions in world coordinates cannot be directly obtained. As a result, many of these previously mentioned methods are ineffective in UAV-based scenarios. Inspired by recent progress in human-scene reconstruction (Kocabas et al. 2024), we focus specifically on reconstructing and rendering moving humans, which are a key component of dynamic scenes captured by UAVs.

Main Results: We present UAV4D, a novel neural rendering framework that reconstructs dynamic human instances and static backgrounds from UAV-captured real-world data, enabling temporally consistent and photorealistic novel-view synthesis. Our methods begin with reconstructing the joint geometry of small dynamic humans and static backgrounds from monocular UAV-captured videos by integrating SMPL-based human meshes with dense background geometry. We leverage the 3D foundation model (Wang et al. 2025a) to estimate camera poses and perform dense background reconstruction, yielding consistent and robust background meshes across diverse scenarios. We observe that HMR2 (Goel et al. 2023) and SAM2 (Ravi et al. 2024) exhibit strong performance in UAV contexts, and we adopt them to recover 4D human motion trajectories as temporally aligned SMPL body meshes (Loper et al. 2023) from monocular videos. However, the human and background meshes suffer from inconsistent scales, and the absolute global positioning of the human meshes remains unknown. To resolve these discrepancies, we propose a scale optimization technique that aligns the background geometry from the 3D foundation model to the metric human mesh, estimating an optimal global scale parameter. Using the optimized scale parameter, we unify the background and human meshes within a consistent world coordinate system over time. Finally, we localize the human mesh in the world space by unprojecting the 2D ground contact point into a 3D location on the reconstructed background mesh.

We then initialize separate 3D Gaussian representations for humans and background using their respective geometric reconstructions. We maintain separate Gaussian splats for humans and background, which are jointly optimized and composited to render photorealistic full frames. Our approach explicitly decomposes dynamic human instances and static backgrounds by independently optimizing their Gaussian splats. By initializing with a strong Gaussian prior from the human SMPL mesh, we ensure that the optimization process properly learns the human Gaussian splats, rather than neglecting them due to their small pixel regions. The key contributions of this work include:

- We introduce UAV4D, the first neural rendering framework that reconstructs dynamic human motions and static backgrounds from UAV-captured monocular videos, enabling temporally consistent and photorealistic novel-view synthesis of full scenes.
- We propose a novel method for resolving the scale ambiguity issue in 3D foundation models and for aligning multiple human meshes with the background mesh. We also present a human placement method by identifying the ground contact point. By initializing both the human and background meshes in world coordinates, we decompose the human and background Gaussian splats to render the complete image.
- To the best of our knowledge, UAV4D is the first to enable dynamic Gaussian splatting for UAV-captured scenes with large number of small moving humans (e.g., up to 50). We evaluated its performance on three datasets with unique characteristics.

Related Work

3D Reconstruction for Static and Dynamic Scenes. Traditional 3D reconstruction typically consists of two components: Structure-from-Motion (SfM) and Multi-View Stereo (MVS) systems. SfM (Schonberger and Frahm 2016) extracts and matches features across multi-view images to estimate camera poses and sparse 3D points via triangulation. MVS (Furukawa and Hernández 2015) then estimates dense depth maps for each image by performing an accurate and efficient global search for the optimal 3D plane hypotheses. Recently, learning-based SfM (Teed and Deng 2021) and MVS (Yao et al. 2018; Yu and Gao 2020; Kim et al. 2021) approaches have leveraged large-scale datasets to enable end-to-end differentiable SfM and dense reconstruction. Notably, DUSt3R (Wang et al. 2024b) and MASt3R (Leroy, Cabon, and Revaud 2024) have emerged as breakthroughs by estimating globally consistent point maps from just two views without requiring any camera information. VGGT (Wang et al. 2025a) modifies the network architecture by incorporating alternative attention mechanisms for improved global alignment. However, these methods are primarily designed for static scenes and struggle to handle dynamic components such as humans. Some works (Zhang et al. 2022; Jung et al. 2021) train monocular depth estimation networks using pre-computed optical flow, or jointly optimize depth and camera poses (Kopf, Rong, and Huang 2021). More recently, concurrent studies (Leroy, Cabon, and Revaud 2024; Duisterhof et al. 2024; Feng et al. 2025) have achieved remarkable ability in dynamic scenes by fine-tuning MASt3R (Duisterhof et al. 2024) with large-scale dynamic datasets to estimate both camera poses and point maps. However, none of these previous methods perform well on UAV-captured datasets, as the humans in our data are relatively small compared to those in their training datasets. Therefore, we decompose the scene into static and dynamic components.

Dynamic Neural Rendering. Earlier neural rendering methods were limited to static scenes, but researchers quickly extended them to handle dynamic scenes. One popular direction (Yang et al. 2024b; Wu et al. 2024; Stearns et al. 2024; Lin et al. 2024b) is deformation field-based methods, which often use a time-conditioned deformation

network to warp 3D Gaussians into each time frame. Another line of work (Yang et al. 2023b; Duan et al. 2024) directly models 4D Gaussian primitives by integrating 3D Gaussians with the time dimension. Some approaches (Zhu et al. 2024; Wang et al. 2024a; Zheng et al. 2025; Stearns et al. 2024) estimate 3D scene motion and point correspondences using pretrained models (Xu et al. 2022a; Doersch et al. 2023; Yang et al. 2024a; Karaev et al. 2024; Yang et al. 2023a) to deform 3D Gaussians. HuGS (Kocabas et al. 2024) presents an approach similar to ours, using Gaussian splats to represent both humans and the scene. However, as their target scenes consist of phone-captured videos (Jiang et al. 2022), they focus exclusively on a single, large person within the scene. Most of these methods are not designed for UAV-captured datasets, as they struggle to handle scenes with multiple pedestrians (more than 10 people) and very small human figures. UAV-Sim (Maxey et al. 2024a) and TKP (Maxey et al. 2024b) are similar tasks to ours, as they use multi-resolution feature spaces in NeRF to represent small dynamic objects. However, we take a different approach by reconstructing the human and background meshes to initialize the Gaussian splats. We then render all the Gaussian splats to generate the full images.

Method

Our method takes as input monocular videos of human-centric scenes (Akshatha et al. 2023; Cao et al. 2021; Barekatain et al. 2017) captured by UAVs. Given this input, our goal is to enable dynamic neural rendering using 3D Gaussian splatting (Kerbl et al. 2023). To achieve this, our approach involves initializing the human-scene from 2D images, reconstructing the 3D geometry of both the human and the scene, and optimizing the Gaussian splats for dynamic neural rendering.

Preliminaries

3D Gaussian Splatting (3DGS) (Kerbl et al. 2023) represents scenes as an unordered set of 3D Gaussian primitives, rendered through a differentiable rasterization process (Zwicker et al. 2002). Each Gaussian component g_i is defined as its mean $\mu_i \in \mathbb{R}^3$ and 3D covariance $\Sigma_i \in \mathbb{R}^{3\text{x}3}$. The covariance matrix is decomposed into two learnable components, a scaling matrix $S_i \in \mathbb{R}^3$ and a rotation matrix $R_i \in \mathbb{R}^{3\text{x}3}$, as $\Sigma_i = R_i S_i {S_i}^T {R_i}^T$. Furthermore, each Gaussians store an opacity logit $o_i \in \mathbb{R}$ and color c_i defined by spherical harmonics (SH) coefficients. To render an image from a given view, the color of each pixel is computed by α-blending K ordered Gaussians using the following equation: $C = \sum_{i=1}^{K} c_i \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j)$ where density α_i is computed by the multiplication of 2D Gaussians with covariance $\Sigma_i' \in \mathbb{R}^{2\text{x}2}$ and a learnable point-wise opacity o_i.

Skinned Multi-Person Linear (SMPL) Model (Loper et al. 2023) is a commonly used parametric human body model. The SMPL model exploits a template human mesh in the canonical rest pose, defined as $\mathcal{M}_h \in (\mathcal{V}_c, \mathcal{F})$ with vertices $\mathcal{V}_c \in \mathbb{R}^{6890\text{x}3}$. It takes body shape parameters θ, β as input and outputs a posed 3D mesh with deformed vertices, where $\theta \in \mathbb{R}^{24\text{x}3\text{x}3}$ and $\beta \in \mathbb{R}^{10}$ represent the pose and shape parameters, respectively. The pose parameters θ consist of the global rotation of the root joint (pelvis), and the 23 local rotations of other articulated joints relative to their parents along the kinematic chain. SMPL uses n_b predefined joints and Linear Blend Skinning (LBS) weights W, defined as $W(v_c) = \{w_1, ..., w_{n_b}\}$. The vertex v_c in the canonical template mesh can be deformed to the articulated space via the LBS, as $v = (\Sigma_{k=1}^{n_b} W_k(v_c) B_k) v_c$, where $B = \{B_1, ..., B_{n_b}\}$ is the target joint transformation.

Initialization

At this stage, we first utilize off-the-shelf models (Wang et al. 2025a; Goel et al. 2023; Ravi et al. 2024) to obtain initial estimates, such as camera poses, point maps, human masks, human meshes, for reconstructing the geometry of both the background scene and the human.

Scene. Given a video sequence $\left\{I_t \in \mathbb{R}^{H\times W\times 3}\right\}_{t=1}^{N}$, we first employ a 3D foundation model (Wang et al. 2025a) to estimate camera poses $\{P_t\}_{t=1}^{N}$ including intrinsics K and extrinsics, rotation R_t and translation T_t, point map $X_t \in \mathbb{R}^{H\times W\times 3}$ and corresponding confidence $C_t \in \mathbb{R}^{H\times W\times 1}$.

Human. Due to the relatively small size of the human subjects, it is challenging to apply recent 3D tracking systems (Goel et al. 2023) trained on standard tracking tasks. Since our goal is precise reconstruction and rendering, we use bounding boxes $\left\{B_t^k\right\}_{k=1}^{K_t}$ as prompts for SAM2 (Ravi et al. 2024) to generate accurate instance masks for K_t humans. These human segmentation masks $\left\{M_t^k\right\}_{k=1}^{K_t}$ are then fed into HMR2.0 (Goel et al. 2023) to predict camera-frame human meshes from each image I_t for K_t detected people. Thus, we can obtain the human mesh parameters for each human, defined as $\left\{\theta_t^k, \beta_t^k\right\}_{k=1}^{K_t}$ at timestep t.

Human Mesh Refinement. However, due to the challenging nature of UAV-based imagery, HMR2.0 (Goel et al. 2023) often produces noisy initializations. We introduce a refinement method to remove human mesh parameters with poor quality. Since the human meshes reside in the camera coordinate system, we project each individual human mesh into the image space and compute the bounding box $\left\{\bar{B}_t^k\right\}_{k=1}^{K_t}$ for each projected human. We then calculate the ratio between B_t^k and $\bar{B}_t^k$, and remove any human meshes for which $\bar{B}_t^k / B_t^k$ exceeds a threshold η_{box}. Furthermore, we measure the overlap between dilated human masks M_{t-1}^k, M_{t+1}^k from the previous and next time frames. If the current bounding box with missing human mesh parameters overlaps in both adjacent frames, we interpolate the missing parameters θ_t^k, β_t^k using the human mesh parameters from the previous frame $(\theta_{t-1}^k, \beta_{t-1}^k)$ and the next frame $(\theta_{t+1}^k, \beta_{t+1}^k)$ applying linear interpolation to both the quaternion pose parameters and the shape vectors.

Joint Human-Scene Reconstruction

Although our initialization provides a reasonable estimate of human poses, all predictions remain in the camera coordinate space. As a result, when the camera moves, the estimated human motion becomes implausible. To align the human and scene in the world coordinate, we reconstruct

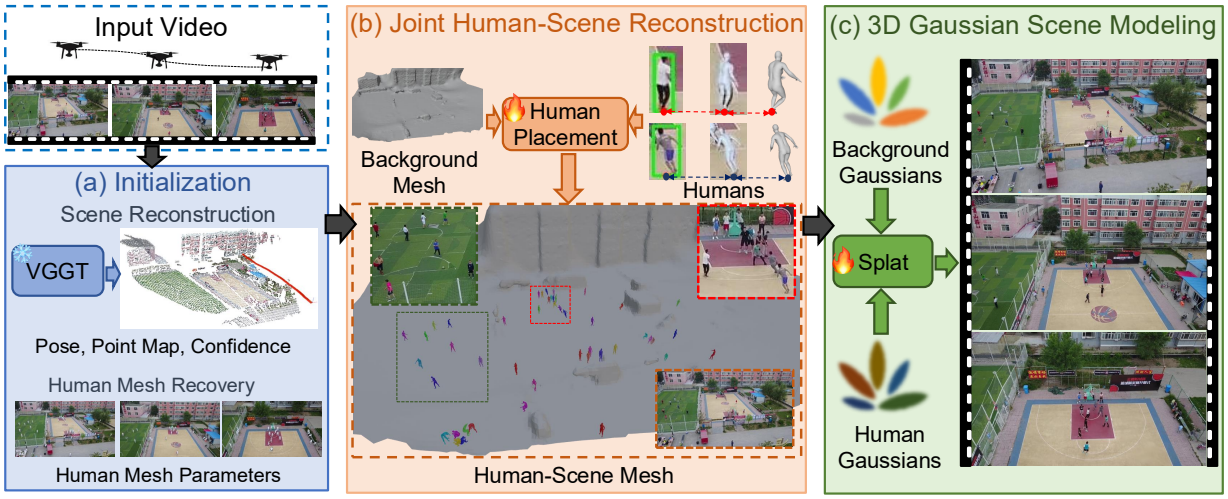


Figure 2: Overview of Our Approach: (a) We begin the initialization process for reconstructing the background scene and the human mesh. (b) We reconstruct the background mesh, determine the optimal scale for alignment, and position the human mesh by identifying the ground contact points. (c) We jointly optimize the human and background Gaussians for rendering.

the background region and identify human-ground contact points to position the human mesh in the scene accurately. We observe that UAVs typically capture wide-area footage with the full human body visible, allowing us to leverage 2D bounding boxes to estimate the human-ground contact points. However, aligning the human mesh with the background remains challenging due to two key issues: (1) in complex environments, the reconstructed background geometry is often noisy and structurally intricate, making it difficult to reliably detect human-scene contact points; and (2) the inherent scale ambiguity of the 3D foundation model (Wang et al. 2025a).

Background Geometry Reconstruction. To address the first issue, we first utilize all point maps $\{X_t\}_{t=1}^{N}$ and corresponding confidence map $\{C_t\}_{t=1}^{N}$ for background reconstruction. The point map represents a dense, pixel-aligned 3D location map for each corresponding image in the world coordinate frame. We stack all these point maps to represent the 3D geometry of the entire scene, although they include noisy points. Then, we compute the η_{conf} percentile of the pixel-aligned confidence values and use it to filter the point maps, removing noisy dynamic regions and intricate geometries with low confidence. The filtered point maps are then stacked, and Poisson surface reconstruction (Kazhdan, Bolitho, and Hoppe 2006) is applied to generate the background mesh $\mathcal{M}_{bg}$. Thanks to this filtering, the background mesh is generated as a clean, flat ground surface, free from noisy dynamic human regions or complex structures that would otherwise make it difficult to identify human-scene contact points.

Scale Optimization. Furthermore, to address the scale ambiguity issue of the 3D foundation model (Wang et al. 2025a), we propose an alignment method between the background and the human mesh. Since the human mesh generated by SMPL (Loper et al. 2023) is defined in metric scale, we can align the background mesh relative to the metric human mesh. We introduce a scale parameter, σ, which adjusts the point maps X_t, background mesh $\mathcal{M}_{bg}$, and camera poses accordingly.

Our idea is to optimize the scale parameter from a *bone length* perspective. Inspired by (Müller et al. 2024), we leverage the correlation between observed 2D keypoints and 3D joints for this alignment. We first use ViTPose (Xu et al. 2022b) to detect 2D keypoints j^k for each person k in the image. Then, we lift the 2D keypoints j_t^k to 3D joints using scaled point maps X_t, defined as $\hat{J}_t^k = \sigma X_t(i,j)$ where $(i,j) \in j_t^k$. Given the human mesh parameters θ_t^k, β_t^k from HMR2.0 (Goel et al. 2023), we can obtain the 3D location J_i of interest joint i from the mesh. Then we compute the 3D bone length $d_{(p,c)}^k$ for each joint pair (p,c) associated with the main body joints $\mathcal{J}_{body}$ of person k in the image at time t, as defined by:

$$\begin{aligned} \hat{d}_{(p,c),t}^k &= J_p(\theta_t^k, \beta_t^k) - J_c(\theta_t^k, \beta_t^k) \\ d_{(p,c),t}^k(\sigma) &= \sigma X_t(p) - \sigma X_t(c) \end{aligned} \tag{1}$$

The loss function $L(\sigma)$ is computed as the L2 norm loss between the predicted 3D joint-based bone lengths $\hat{d}_{(p,c),t}^k$ and the corresponding ground-truth values $d_{(p,c),t}^k(\sigma)$. Specifically, the loss is calculated across all time steps t, individuals k, and joints (p,c) in the main body, where minimizing this loss ensures that the predicted bone lengths align closely with the ground truth. The scale parameter σ is optimized using the L-BFGS algorithm (Nocedal 1980) applied to this loss. The loss function for optimizing the scale parameter σ is as follows:

$$L(\sigma) = \sum_t \sum_{k \in \mathcal{K}} \sum_{(p,c) \in \mathcal{J}_{body}} ||\hat{d}_{(p,c),t}^k - d_{(p,c),t}^k(\sigma)||_2. \tag{2}$$

The update rule is $\sigma_{n+1} = \sigma_n - \alpha \mathcal{H}_n \nabla L(\sigma_n)$ where α is the learning rate and, $\mathcal{H}_n$ is the approximation of the inverse

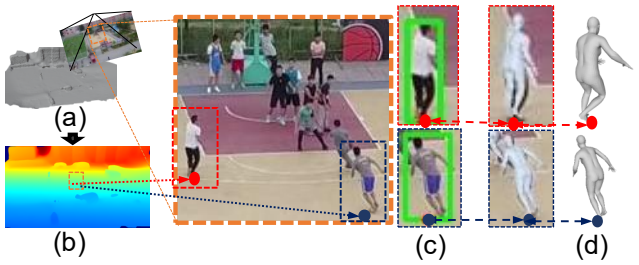


Figure 3: Human Placement. We identify the 2D contact point (c) from the bounding box and use the depth (b) from the mesh (a) to unproject the 3D ground contact point (d).

Hessian. The optimization proceeds for T_{opt} steps, and the objective is minimized through gradient-based optimization.

Human Placement. After determining the optimal scale σ, we apply this scale to the background mesh M_{bg} and camera poses P_t to align them with the human mesh. In conjunction with the camera pose, it is necessary to initialize the root translation in the world coordinate system. Our approach takes advantage of the fact that UAV-captured images provide top-down views, which facilitate the identification of the ground contact point. This contact point serves as a plausible initial estimate for the root translation. Given the background mesh M_{bg}, we render the depth map D_t from the specified viewpoint. Using the 2D bounding box $B_t^k = (x_{\text{min}}, y_{\text{min}}, x_{\text{max}}, y_{\text{max}})$ projected from the human mesh, we detect the 2D contact point as $(x_c, y_c) = ((x_{\text{min}} + x_{\text{max}})/2,\ y_{\text{max}})$. Subsequently, we use the depth value from D_t to unproject this 2D contact point for each individual human into the world coordinate system as follows:

$$\psi_t = \sigma R^T [K^{-1} D_{x_c,y_c} [x_c, y_c, 1]^T] - \sigma R^T T. \quad (3)$$

ψ_t is the 3D ground contact point in the world coordinate. Given this, we can accurately position all individual human meshes within the background mesh, with all meshes defined in the world coordinate system, as seen in Fig. 3.

3D Gaussian Scene Modeling

We model the scene using distinct Gaussian representations for dynamic humans and the static background, ensuring both temporal coherence and rendering clarity. We employ standard 3D Gaussians to model the background Gaussians g_{bg} and the human Gaussians g_h. Both Gaussians g_h and g_{bg} contain learnable attributes (μ, r, s, c, o), which represent the 3D mean, 3D rotation, scaling factor, opacity factor, and color value, respectively. These are then merged into a unified human-scene primitive $g_{all} = g_{bg} + g_h$, which is subsequently fed into the Gaussian rasterizer (Kerbl et al. 2023) for rendering.

Background Gaussians. The background is represented by static Gaussians g_{bg} , which are initialized using stacked multi-view point maps X_t. For UAV scenes captured at high altitudes, we use all point maps without filtering. In contrast, for UAV scenes captured at low altitudes, we employ a confidence map to filter the point maps and use human masks M_t^k to remove point maps corresponding to human regions, stacking only the static background regions.

Human Gaussians. For human Gaussians, inspired by GART (Lei et al. 2024), we utilize the canonical template mesh $\mathcal{M}_h$ in the rest pose to represent the 3D Gaussian splats. The skinning transformation for each Gaussian, $A(t) = (R_A(t), T_A(t))$, is derived from the LBS weight $LBS(g_h, \theta_t)$, which transform the canonical Gaussian positions μ and rotations r to the world frame.

$$g_h(t) = (R_A(t)\mu + T_A(t) + \psi_t,\ R_A(t) \cdot r,\ s, c, o). \quad (4)$$

The changes in θ_t over time result in updates to the transformations of the key joints and linearly interpolates Gaussians to obtain the deformed position $\mu(t)$ and rotations $r(t)$ at different time steps.

Optimization. We optimize all Gaussian attributes g_{bg}, g_h, the human poses of all SMPL parameters θ_t, ψ_t for each frame t and the corresponding skimming weights.

The overall loss function for optimization is defined as:

$$L_{total} = (1 - \lambda_{pho})L_1 + \lambda_{pho} L_{ssim} + \lambda_o L_o + L_{smpl}, \quad (5)$$

where L_1 and L_{ssim} represent the photometric loss between the rendered full image and the ground truth image. L_{smpl} consists of various loss terms related to human Gaussian splats and human regions. We set λ_{pho} to 0.2 and λ_o to 0.05. Details are provided in the supplementary material.

Experimental Results

To measure the effectiveness of UAV4D, we evaluate our method on novel view synthesis tasks, using every 8th frame as the held-out test set. We report PSNR, SSIM, and LPIPS for full images and human-related regions, to assess dynamic reconstruction capabilities. Details of the datasets are provided in the appendix. The human regions are obtained using a pretrained segmentation model which get a prompt from groundtruth bonding box.

Baselines: To evaluate rendering quality, we compare our method with four comparative methods: TKP (Maxey et al. 2024b), 3DGS (Kerbl et al. 2023), DfGS (Yang et al. 2024b), and 4DGS (Wu et al. 2024). To the best of our knowledge, TKP (Maxey et al. 2024b) is the only work that has designed dynamic NeRFs for UAV scenes, while 4DGS (Wu et al. 2024) and DfGS (Yang et al. 2024b) represent the state-of-the-art methods in dynamic Gaussian-splatting.

Quantitative Comparison

In Table 1, we present the rendering quality of novel-view synthesis across three datasets. We compute the average rendering metrics across 4 scenes for VisDrone (Cao et al. 2021), 4 scenes for Manipal-UAV (Akshatha et al. 2023), and 4 scenes for Okutama-Action (Barekatain et al. 2017). Due to the small size of the human subjects in UAV datasets, the rendering metrics for the full image do not reflect the dynamic rendering quality of the human regions directly. Moreover, since the Manipal-UAV dataset was captured at an exceptionally high altitude, the rendering metrics of the full image are less informative regarding the human regions and instead provide a better indication of the background region's rendering quality. In contrast, the VisDrone dataset

Method	Okutama-Action			Manipal-UAV			VisDrone		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
TKP (Maxey et al. 2024b)	28.05	0.796	0.417	27.98	0.718	0.396	23.98	0.644	0.450
3DGS (Kerbl et al. 2023)	29.90	0.867	0.228	29.46	0.856	0.150	24.59	0.782	0.220
4DGS (Wu et al. 2024)	25.90	0.780	0.410	30.87	0.843	0.253	22.58	0.628	0.440
DfGS (Yang et al. 2024b)	22.64	0.748	0.443	31.22	0.894	0.114	15.46	0.522	0.614
UAV4D (Ours)	30.36	0.875	0.184	30.94	0.897	0.084	26.03	0.800	0.147

Table 1: Quantitative Comparison of our method with recent comparative works on the three datasets. Please see the Appendix for detailed comparisons on the four scenes in each dataset. Red, orange, and yellow indicate the first, second, and third best performing algorithms for each metric, respectively.

Method	Okutama-Action		Manipal-UAV		VisDrone	
	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑	PSNR ↑	SSIM ↑
TKP (Maxey et al. 2024b)	19.14	0.631	18.38	0.752	17.65	0.544
3DGS (Kerbl et al. 2023)	19.07	0.656	17.99	0.748	16.51	0.548
4DGS (Wu et al. 2024)	18.18	0.602	19.23	0.770	16.25	0.510
DfGS (Yang et al. 2024b)	16.33	0.571	18.65	0.762	12.59	0.354
UAV4D (Ours)	19.49	0.664	19.38	0.784	17.98	0.604

Method	PSNR ↑	SSIM ↑	LPIPS ↓
Ours	26.03	0.800	0.147
wo SMPL	24.33	0.763	0.212
wo Scale	23.02	0.651	0.250
wo Refine	25.79	0.772	0.165

Table 2: Quantitative Results. (a) The left table shows PSNR and SSIM across datasets, focusing on human-only regions cropped using precise human masks. Due to the small size of the human regions, it is not possible to compute the LPIPS metric. Consequently, we report only the PSNR and SSIM metrics. (b) The right table presents ablation on VisDrone dataset.

contains multiple humans, with relatively larger pixel regions compared to other datasets, allowing the full image rendering quality to be more directly correlated with the holistic scene rendering quality. Furthermore, certain scenes in VisDrone feature a significant number of moving pedestrians (approximately 30–50 individuals), presenting a particularly challenging scenario for dynamic reconstruction. Our method demonstrates superior rendering quality compared to other recent approaches, underscoring its robustness. In the Okutama-Action dataset, our method also exhibits better rendering quality than alternative methods. The Okutama-Action dataset is characterized by substantial camera variation and distinct viewing angles, which can destabilize deformation-field-based methods. However, our method shows strong resilience in handling these dynamic scenes, which involve substantial pose variation and changes in viewing angle.

In the VisDrone and Okutama-Action datasets, we observe that the training of 4DGS (Wu et al. 2024) and DfGS (Yang et al. 2024b) is unstable, with the training of deformation-based fields often failing. We attribute this instability to the presence of multiple small moving objects within the scenes, which indicates that a single deformation field is insufficient to capture the required deformations accurately. As a result, the training process for these models becomes highly unstable. In contrast, TKP (Maxey et al. 2024b) demonstrates greater robustness across various dynamic scenes, yielding superior performance. For the Manipal-UAV dataset, our method outperforms others in most cases, except for the PSNR values for DfGS (Yang et al. 2024b). We observe that DfGS (Yang et al. 2024b) delivers superior rendering performance in regions with dense foliage and moving vegetation (e.g., swaying leaves). Our method currently struggles to represent such deformations, particularly in swaying trees, which cover large areas in UAV-captured datasets.

Table 2-(a) reports the rendering quality across three datasets, focusing on human-only regions cropped using precise human masks. Due to the small pixel size of the human regions, we are unable to use the LPIPS (Zhang et al. 2018) metric, which relies on pretrained neural networks to measure metrics. Across all datasets, our method demonstrates superior rendering quality for human regions compared to other approaches. This indicates that our strategy effectively reconstructs and renders dynamic humans.

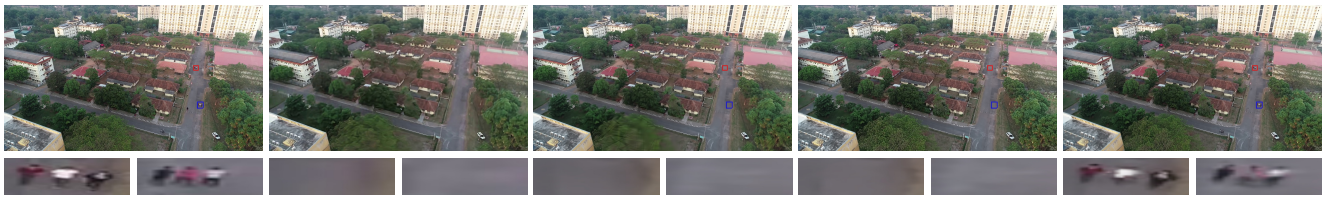
Ablation Study Table 2-(b) presents the results of an ablation study, averaged across four scenes in the Visdeon dataset. “wo SMPL” indicates the version where we did not use human Gaussian splats, thus reconstructing only the background splats, which limits the model to static humans and removes moving people. “wo Scale” refers to the version where the human mesh is placed without scale optimization. In this case, we set the initial scale σ to 40 and place the human mesh without any scale adjustment, leading to misalignment between the human regions and the human mesh. “wo Refine” follows the full pipeline but omits the human mesh refinement process described in the the Initialization section. This version shows a slight improvement in artifacts caused by inaccurate in the human mesh.

Qualitative Comparison

Figure 4 presents visual examples from the VisDrone dataset (Cao et al. 2021). Overall, our method achieves the highest rendering quality, producing a photorealistic background and sharp human regions, where the detailed shapes of the humans are more accurately recovered compared to other

(a) Scene (b) TKP (c) 3DGS (d) 4DGS (e) UAV4D (Ours)

Figure 4: Qualitative Comparison of VisDrone (Cao et al. 2021) (top row) and Okutama-Action Dataset (Barekatain et al. 2017) (bottom row). We visualize the zoomed-in blue and red regions, which emphasize the dynamic humans. Our method demonstrates superior capability in reconstructing small, moving humans compared to other existing approaches.



(a) Scene (b) TKP (c) 4DGS (d) DfGS (e) UAV4D (Ours)

Figure 5: Qualitative Comparison of Manipal-UAV Dataset (Akshatha et al. 2023). We visualize the zoomed-in blue and red regions, which emphasize the dynamic humans. Our method demonstrates superior capability in reconstructing small, moving humans compared to other existing approaches.

Figure 6: Scene Editing Samples: Due to the nature of scene decomposition, we can move or remove the humans within the black-dotted boxes (2-3 columns). The fourth image illustrates the decomposition of the background and humans.

methods. TKP (Maxey et al. 2024b) generates blurred renderings, while 3DGS (Kerbl et al. 2023) is unable to represent moving individuals, although it successfully reconstructs clear background regions. The training of 4DGS (Wu et al. 2024) occasionally fails. A similar trend is observed in the Okutama-Action dataset (Barekatain et al. 2017), as shown in Figure 4. Figure 5 displays examples from the Manipal-UAV dataset (Akshatha et al. 2023). Our method can reconstructs extremely small humans, occupying only around 10x10 pixels, outperforming all other methods. However, DfGS (Yang et al. 2024b) shows better performance in rendering swaying leaves and trees, which are frequently observed in this dataset.

Scene Editing Since our Gaussian scene is composed of separate human and background Gaussians, our system enables scene editing. As demonstrated in Fig. 6, we can easily remove, and translate humans, thanks to the decomposition of the foreground and background. This capability can be extended to downstream applications, such as simulating dynamic scenes captured by UAV.

Conclusion and Future Work

We have introduced UAV4D, a method for dynamic Gaussian splatting in UAV-captured environments. We propose a scale optimization and human placement technique to reconstruct human-scene geometry in the world coordinate system. This human-scene geometry is then used to initialize 3D Gaussian splats for dynamic scenes. It is important to note that the explicit representation of human geometry in our method is crucial for accurately modeling small, moving humans, ensuring they are not neglected during the optimization of Gaussian splats. Our joint optimization approach demonstrates improved performance in novel view synthesis, both for background and human regions. In future work, we plan to modify our human mesh refinement algorithm to better align with 2D images and incorporate segmentation information to represent other dynamic components using dynamic Gaussian splats. Our approach can be combined with downstream UAV applications such as person detection (Maxey et al. 2024a).

Acknowledgments

This work was supported in part by ARO Grant W911NF2310352 and Army Cooperative Agreement W911NF2120076

References

Akshatha, K.; Karunakar, A.; Satish Shenoy, B.; Phani Pavan, K.; Chinmay, V. D.; et al. 2023. Manipal-UAV person detection dataset: A step towards benchmarking dataset and algorithms for small object detection. *ISPRS Journal of Photogrammetry and Remote Sensing*, 195: 77–89.

Barekatain, M.; Martí, M.; Shih, H.-F.; Murray, S.; Nakayama, K.; Matsuo, Y.; and Prendinger, H. 2017. Okutama-action: An aerial view video dataset for concurrent human action detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 28–35.

Cao, Y.; He, Z.; Wang, L.; Wang, W.; Yuan, Y.; Zhang, D.; Zhang, J.; Zhu, P.; Van Gool, L.; Han, J.; et al. 2021. VisDrone-DET2021: The vision meets drone object detection challenge results. In *Proceedings of the IEEE/CVF International conference on computer vision*, 2847–2854.

Chen, Z.; Yang, J.; Huang, J.; de Lutio, R.; Esturo, J. M.; Ivanovic, B.; Litany, O.; Gojcic, Z.; Fidler, S.; Pavone, M.; Song, L.; and Wang, Y. 2025. OmniRe: Omni Urban Scene Reconstruction. In *The Thirteenth International Conference on Learning Representations*.

Doersch, C.; Yang, Y.; Vecerik, M.; Gokay, D.; Gupta, A.; Aytar, Y.; Carreira, J.; and Zisserman, A. 2023. Tapir: Tracking any point with per-frame initialization and temporal refinement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 10061–10072.

Duan, Y.; Wei, F.; Dai, Q.; He, Y.; Chen, W.; and Chen, B. 2024. 4d-rotor gaussian splatting: towards efficient novel view synthesis for dynamic scenes. In *ACM SIGGRAPH 2024 Conference Papers*, 1–11.

Duisterhof, B.; Zust, L.; Weinzaepfel, P.; Leroy, V.; Cabon, Y.; and Revaud, J. 2024. MASt3R-SfM: a Fully-Integrated Solution for Unconstrained Structure-from-Motion. *arXiv preprint arXiv:2409.19152*.

Feng, H.; Zhang, J.; Wang, Q.; Ye, Y.; Yu, P.; Black, M. J.; Darrell, T.; and Kanazawa, A. 2025. St4RTrack: Simultaneous 4D Reconstruction and Tracking in the World. *arXiv preprint arXiv:2504.13152*.

Fridovich-Keil, S.; Meanti, G.; Warburg, F. R.; Recht, B.; and Kanazawa, A. 2023. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12479–12488.

Furukawa, Y.; and Hernández, C. 2015. Multi-view stereo: A tutorial. *Foundations and Trends® in Computer Graphics and Vision*, 9(1-2): 1–148.

Gao, H.; Li, R.; Tulsiani, S.; Russell, B.; and Kanazawa, A. 2022. Monocular dynamic view synthesis: A reality check. *Advances in Neural Information Processing Systems*, 35: 33768–33780.

Goel, S.; Pavlakos, G.; Rajasegaran, J.; Kanazawa, A.; and Malik, J. 2023. Humans in 4D: Reconstructing and tracking humans with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14783–14794.

Jiang, W.; Yi, K. M.; Samei, G.; Tuzel, O.; and Ranjan, A. 2022. Neuman: Neural human radiance field from a single video. In *European Conference on Computer Vision*, 402–418. Springer.

Jung, D.; Choi, J.; Lee, Y.; Kim, D.; Kim, C.; Manocha, D.; and Lee, D. 2021. Dnd: Dense depth estimation in crowded dynamic indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 12797–12807.

Karaev, N.; Rocco, I.; Graham, B.; Neverova, N.; Vedaldi, A.; and Rupprecht, C. 2024. Cotracker: It is better to track together. In *European Conference on Computer Vision*, 18–35. Springer.

Kazhdan, M.; Bolitho, M.; and Hoppe, H. 2006. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7.

Kerbl, B.; Kopanas, G.; Leimkühler, T.; and Drettakis, G. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics*, 42(4).

Kim, T.; Choi, J.; Choi, S.; Jung, D.; and Kim, C. 2021. Just a few points are all you need for multi-view stereo: A novel semi-supervised learning method for multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6178–6186.

Kocabas, M.; Chang, J.-H. R.; Gabriel, J.; Tuzel, O.; and Ranjan, A. 2024. Hugs: Human gaussian splats. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 505–515.

Kopf, J.; Rong, X.; and Huang, J.-B. 2021. Robust consistent video depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1611–1621.

Lei, J.; Wang, Y.; Pavlakos, G.; Liu, L.; and Daniilidis, K. 2024. Gart: Gaussian articulated template models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 19876–19887.

Leroy, V.; Cabon, Y.; and Revaud, J. 2024. Grounding image matching in 3d with mast3r. In *European Conference on Computer Vision*, 71–91. Springer.

Lin, J.; Li, Z.; Tang, X.; Liu, J.; Liu, S.; Liu, J.; Lu, Y.; Wu, X.; Xu, S.; Yan, Y.; et al. 2024a. Vastgaussian: Vast 3d gaussians for large scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5166–5175.

Lin, Y.; Dai, Z.; Zhu, S.; and Yao, Y. 2024b. Gaussian-flow: 4d reconstruction with dynamic 3d gaussian particle. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21136–21145.

Loper, M.; Mahmood, N.; Romero, J.; Pons-Moll, G.; and Black, M. J. 2023. SMPL: A skinned multi-person linear model. In *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, 851–866.

Luiten, J.; Kopanas, G.; Leibe, B.; and Ramanan, D. 2024. Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In *2024 International Conference on 3D Vision (3DV)*, 800–809. IEEE.

Maxey, C.; Choi, J.; Lee, H.; Manocha, D.; and Kwon, H. 2024a. Uav-sim: Nerf-based synthetic data generation for uav-based perception. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, 5323–5329. IEEE.

Maxey, C.; Choi, J.; Lee, Y.; Lee, H.; Manocha, D.; and Kwon, H. 2024b. TK-planes: Tiered K-planes with high dimensional feature vectors for dynamic UAV-based scenes. *arXiv preprint arXiv:2405.02762*.

Müller, L.; Choi, H.; Zhang, A.; Yi, B.; Malik, J.; and Kanazawa, A. 2024. Reconstructing People, Places, and Cameras. *arXiv preprint arXiv:2412.17806*.

Nocedal, J. 1980. Updating quasi-Newton matrices with limited storage. *Mathematics of computation*, 35(151): 773–782.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; et al. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4104–4113.

Stearns, C.; Harley, A.; Uy, M.; Dubost, F.; Tombari, F.; Wetzstein, G.; and Guibas, L. 2024. Dynamic gaussian marbles for novel view synthesis of casual monocular videos. In *SIGGRAPH Asia 2024 Conference Papers*, 1–11.

Teed, Z.; and Deng, J. 2021. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34: 16558–16569.

Wang, J.; Chen, M.; Karaev, N.; Vedaldi, A.; Rupprecht, C.; and Novotny, D. 2025a. VGGT: Visual Geometry Grounded Transformer. *arXiv preprint arXiv:2503.11651*.

Wang, Q.; Ye, V.; Gao, H.; Zeng, W.; Austin, J.; Li, Z.; and Kanazawa, A. 2024a. Shape of Motion: 4D Reconstruction from a Single Video.

Wang, Q.; Zhang, Y.; Holynski, A.; Efros, A. A.; and Kanazawa, A. 2025b. Continuous 3D Perception Model with Persistent State. *arXiv preprint arXiv:2501.12387*.

Wang, S.; Leroy, V.; Cabon, Y.; Chidlovskii, B.; and Revaud, J. 2024b. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20697–20709.

Wu, G.; Yi, T.; Fang, J.; Xie, L.; Zhang, X.; Wei, W.; Liu, W.; Tian, Q.; and Wang, X. 2024. 4d gaussian splatting for real-time dynamic scene rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20310–20320.

Xu, H.; Zhang, J.; Cai, J.; Rezatofighi, H.; and Tao, D. 2022a. Gmflow: Learning optical flow via global matching. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8121–8130.

Xu, Y.; Zhang, J.; Zhang, Q.; and Tao, D. 2022b. Vitpose: Simple vision transformer baselines for human pose estimation. *Advances in neural information processing systems*, 35: 38571–38584.

Xu, Z.; Peng, S.; Lin, H.; He, G.; Sun, J.; Shen, Y.; Bao, H.; and Zhou, X. 2024. 4K4D: Real-Time 4D View Synthesis at 4K Resolution. In *CVPR*.

Yang, J.; Gao, M.; Li, Z.; Gao, S.; Wang, F.; and Zheng, F. 2023a. Track anything: Segment anything meets videos. *arXiv preprint arXiv:2304.11968*.

Yang, L.; Kang, B.; Huang, Z.; Xu, X.; Feng, J.; and Zhao, H. 2024a. Depth anything: Unleashing the power of large-scale unlabeled data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10371–10381.

Yang, Z.; Gao, X.; Zhou, W.; Jiao, S.; Zhang, Y.; and Jin, X. 2024b. Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 20331–20341.

Yang, Z.; Yang, H.; Pan, Z.; and Zhang, L. 2023b. Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. *arXiv preprint arXiv:2310.10642*.

Yao, Y.; Luo, Z.; Li, S.; Fang, T.; and Quan, L. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, 767–783.

Yu, Z.; and Gao, S. 2020. Fast-mvsnet: Sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1949–1958.

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zhang, Z.; Cole, F.; Li, Z.; Rubinstein, M.; Snavely, N.; and Freeman, W. T. 2022. Structure and motion from casual videos. In *European Conference on Computer Vision*, 20–37. Springer.

Zheng, C.; Xue, L.; Zarate, J.; and Song, J. 2025. GSTAR: Gaussian Surface Tracking and Reconstruction. *arXiv preprint arXiv:2501.10283*.

Zhu, J.; and Tang, H. 2025. Dynamic Scene Reconstruction: Recent Advance in Real-time Rendering and Streaming. *arXiv preprint arXiv:2503.08166*.

Zhu, R.; Liang, Y.; Chang, H.; Deng, J.; Lu, J.; Yang, W.; Zhang, T.; and Zhang, Y. 2024. Motiongs: Exploring explicit motion guidance for deformable 3d gaussian splatting. *Advances in Neural Information Processing Systems*, 37: 101790–101817.

Zwicker, M.; Pfister, H.; Van Baar, J.; and Gross, M. 2002. EWA splatting. *IEEE Transactions on Visualization and Computer Graphics*, 8(3): 223–238.