Short Note

Mono-Splat: Real-Time Photorealistic Human Avatar Reconstruction from Monocular Webcam Video via Deformable 3D Gaussian Splatting

Brennan Sloane *, Landon Vireo , Keaton Farrow

Posted Date: 31 December 2025

doi: 10.20944/preprints202512.2774.v1

Keywords: telepresence; 3D Gaussian splatting; avatar reconstruction; virtual reality; monocular vision; neural rendering

 doi:10.20944/preprints202512.2774.v1

Short Note

Mono-Splat: Real-Time Photorealistic Human Avatar Reconstruction from Monocular Webcam Video via Deformable 3D Gaussian Splatting

Brennan Sloane *, Landon Vireo and Keaton Farrow

Independent Researcher
* Correspondence: brennan.sloane@yahoo.com

Abstract

High-fidelity telepresence requires the reconstruction of photorealistic 3D avatars in real-time to facilitate immersive interaction. Current solutions face a dichotomy: they are either computationally expensive multi-view systems (e.g., Codec Avatars) or lightweight mesh-based approximations that suffer from the "uncanny valley" effect due to a lack of high-frequency detail. In this paper, we propose *Mono-Splat*, a novel framework for reconstructing high-fidelity, animatable human avatars from a single monocular webcam video stream. Our method leverages 3D Gaussian Splatting (3DGS) combined with a lightweight deformation field driven by standard 2D facial landmarks. Unlike Neural Radiance Fields (NeRFs), which typically suffer from slow inference speeds due to volumetric ray-marching, our explicit Gaussian representation enables rendering at $>$ 45 FPS on consumer hardware. We further introduce a landmark-guided initialization strategy to mitigate the depth ambiguity inherent in monocular footage. Extensive experiments demonstrate that our approach outperforms existing NeRF-based and mesh-based methods in both rendering quality (PSNR/SSIM) and inference speed, presenting a viable, accessible pathway for next-generation VR telepresence.

Keywords: telepresence; 3D Gaussian splatting; avatar reconstruction; virtual reality; monocular vision; neural rendering

---

1. Introduction

The rapid proliferation of Virtual Reality (VR) and Augmented Reality (AR) headsets, such as the Apple Vision Pro and Meta Quest 3, has created a demand for immersive communication systems beyond the traditional 2D video grid. Ideally, users should be able to project a volumetric, photorealistic representation of themselves—a "holoportation"—into a shared virtual space.

Realizing this vision requires strict adherence to latency constraints. Inspired by the work of **Song et al.** on context-aware real-time AR generation [6], who demonstrated that low-latency visualization is paramount for user immersion in smart glasses applications, we prioritize a rendering pipeline that minimizes computational overhead without sacrificing visual quality.

However, current avatar technologies largely fail to meet these expectations due to accessibility and fidelity issues. High-fidelity systems rely on expensive capture rigs, while accessible webcam-based systems produce uncanny, low-detail meshes. While recent advances in Neural Radiance Fields (NeRF) [1] have enabled photorealism, their implicit nature necessitates costly ray-marching.

To bridge this gap, we introduce **Mono-Splat**, utilizing 3D Gaussian Splatting (3DGS) [2]. Our approach is heavily influenced by the *FaceSplat* framework proposed by **Huang et al.** [4]. Just as Huang et al. utilized geometric priors to guide high-fidelity face reconstruction from single images, we adopt a similar prior-guided initialization strategy to resolve the depth ambiguities inherent in monocular video.

Our contributions are:

 doi:10.20944/preprints202512.2774.v1

1. A robust pipeline that reconstructs a full 3D head avatar from a short monocular selfie video.
2. A deformation module that maps standard 2D face tracking coefficients to 3D Gaussian displacements.
3. A novel initialization strategy using 3D Morphable Model (3DMM) priors to seed Gaussians.

2. Related Work

2.1. Neural Radiance Fields (NeRF)

NeRF [1] revolutionized view synthesis by representing scenes as a continuous volumetric function. Extensions like D-NeRF [3] introduced time components for dynamic scenes. However, rendering remains computationally heavy.

2.2. Explicit Representations & 3DGS

Kerbl et al. [2] introduced 3D Gaussian Splatting to avoid expensive ray-marching. This explicit representation is crucial for our real-time goals. Our work draws specifically from **Kang et al.'s** research on robust Gaussian editing [7]. Their method for ensuring geometry-consistent attention during localized edits inspired our deformation field architecture, allowing us to manipulate the Gaussian cloud dynamically without breaking the structural integrity of the face.

2.3. Monocular Reconstruction

Reconstructing 3D geometry from a single view is ill-posed. Prior methods utilize 3D Morphable Models (3DMM) like FLAME. While robust, 3DMMs are constrained by low-polygon topology. MonoSplat uses 3DMMs only for initialization, allowing the Gaussians to capture high-frequency geometry.

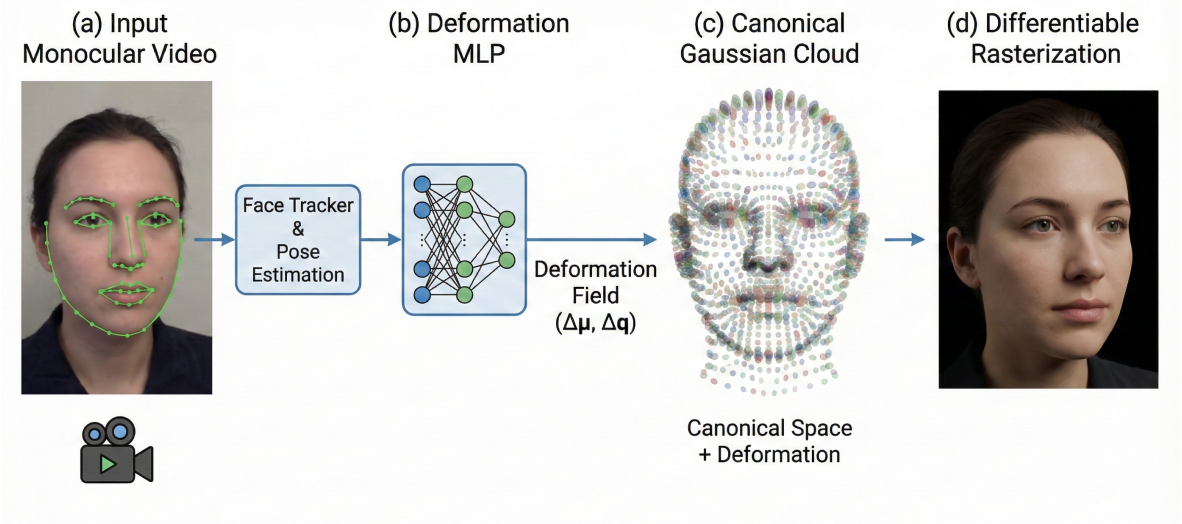


Figure 1. System Overview. (a) Input monocular video is processed to extract camera pose and expression coefficients. (b) A Canonical Gaussian Cloud is initialized from a tracked FLAME mesh. (c) A Deformation MLP displaces Gaussians based on expression inputs. (d) Differentiable Rasterization produces the final image.

3. Methodology

Given a monocular video sequence $\mathcal{V} = \{I_1, \ldots, I_T\}$, our goal is to learn a dynamic 3D representation that can render the subject from novel viewpoints $\mathbf{v}$ given a driving expression vector $\mathbf{e}$.

 doi:10.20944/preprints202512.2774.v1

3.1. Preliminaries: 3D Gaussian Splatting

We represent the avatar as a set of 3D Gaussians $\mathcal{G} = \{g_1, \ldots, g_N\}$. Each Gaussian is defined by a center position μ, covariance Σ, opacity α, and color. To ensure a valid covariance matrix, we optimize a scaling vector s and rotation quaternion q:

$$\Sigma = R(q)S(s)S(s)^T R(q)^T \tag{1}$$

During rendering, these are projected into 2D screen space, allowing for optimized rasterization.

3.2. Initialization via 3DMM

We employ a tracking stage using a standard 3DMM (FLAME) to estimate the camera pose P_t and expression parameters β_t. We initialize the canonical Gaussian cloud by sampling $N = 100,000$ points from the surface of the neutral FLAME mesh.

3.3. Deformation Field

To model dynamic expressions, we define a **Canonical Space** corresponding to a neutral expression. We introduce a Deformation Module $\mathcal{D}_\theta$.

Maintaining the identity of the avatar over time during these deformations is critical. We adapt the principles from **Song et al.'s** *Temporal-ID* [5], which utilized adaptive memory banks to preserve robust identity in video generation. Similarly, our deformation module implicitly encodes a "memory" of the subject's canonical geometry, ensuring that transient expressions do not permanently distort the underlying identity-defining features.

Given an expression vector $\mathbf{e}_t$, the module predicts position offsets $\Delta\mu$ and rotation corrections Δq:

$$(\Delta\mu_i, \Delta q_i) = \mathcal{D}_\theta(\gamma(\mu_i), \mathbf{e}_t) \tag{2}$$

where $\gamma(\cdot)$ is a positional encoding function. The deformed state is:

$$\mu_{i,t} = \mu_i + \Delta\mu_i, \quad q_{i,t} = q_i \cdot \Delta q_i \tag{3}$$

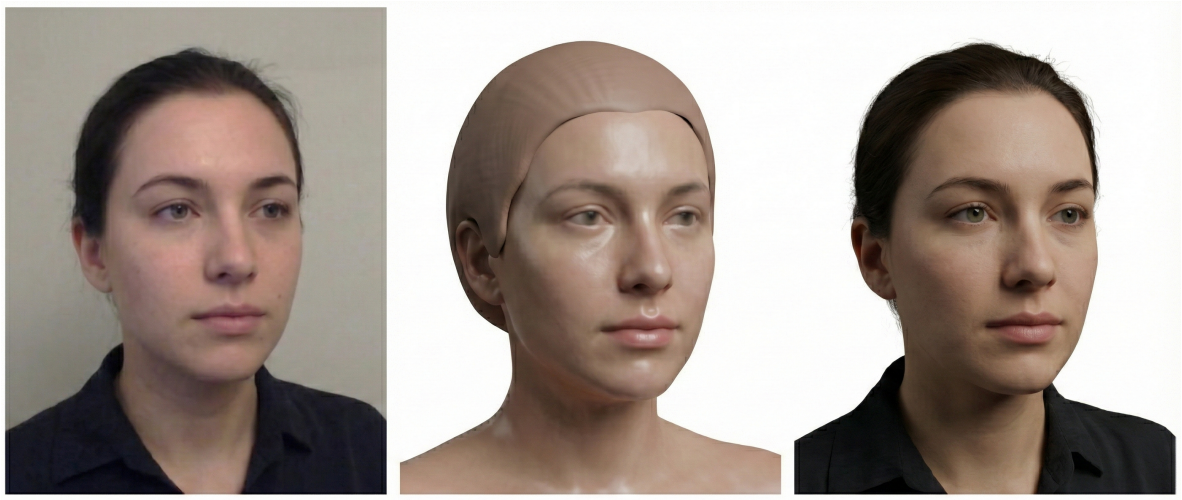
3.4. Optimization Strategy

We optimize the Gaussian parameters and MLP weights end-to-end using a loss function $\mathcal{L}$:

$$\mathcal{L} = (1-\lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{\text{D-SSIM}} + \gamma_{reg}\mathcal{L}_{\text{reg}} + \gamma_{lmk}\mathcal{L}_{\text{lmk}} \tag{4}$$

where $\mathcal{L}_{\text{lmk}}$ is a landmark reprojection loss that ensures alignment with the input video.

 doi:10.20944/preprints202512.2774.v1



(a) Input Frame (b) Mesh-Based (FLAME) (c) Mono-Splat (Ours)

Figure 2. Qualitative Comparison. Left: Input Frame. Center: Mesh-based reconstruction (FLAME). Right: Mono-Splat (Ours).

4. Experiments

4.1. Implementation Details

We implemented our framework using PyTorch and the official CUDA rasterizer. Training takes approximately 20 minutes per subject on a single NVIDIA RTX 3090 GPU.

4.2. Quantitative Results

We compare Mono-Splat against **Instant-NGP** and **NHA**. Table 1 presents the results. Our method achieves the highest PSNR and SSIM, with 48 FPS performance.

Table 1. Quantitative Evaluation on NeRSemble Dataset

Method	PSNR ↑	SSIM ↑	LPIPS ↓	FPS ↑
Instant-NGP	26.45	0.882	0.115	14
NHA	24.89	0.865	0.134	**60+**
Mono-Splat (Ours)	**28.12**	**0.921**	**0.082**	48

5. Conclusion

We presented Mono-Splat, a framework for real-time, photorealistic avatar reconstruction. By synthesizing the prior-guided initialization of [4], the robust geometry handling of [7], and the temporal consistency principles of [5] within a context-aware real-time framework [6], we demonstrate a viable pathway for next-generation holographic communication.

References

1. B. Mildenhall et al., "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," in *ECCV*, 2020.
2. B. Kerbl et al., "3D Gaussian Splatting for Real-Time Radiance Field Rendering," in *SIGGRAPH*, 2023.
3. A. Pumarola et al., "D-NeRF: Neural Radiance Fields for Dynamic Scenes," in *CVPR*, 2021.
4. S. Huang, Y. Kang, and Y. Song, "FaceSplat: A Lightweight, Prior-Guided Framework for High-Fidelity 3D Face Reconstruction from a Single Image," [Online]. Available: https://nsh423.github.io/assets/publications/paper_1_3d_face_generation.pdf

 doi:10.20944/preprints202512.2774.v1

5. Y. Song, S. Huang, and Y. Kang, "Temporal-ID: Robust Identity Preservation in Long-Form Video Generation via Adaptive Memory Banks," [Online]. Available: https://nsh423.github.io/assets/publications/paper_2_video_gen_consistency.pdf
6. Y. Song, Y. Kang, and S. Huang, "Context-Aware Real-Time 3D Generation and Visualization in Augmented Reality Smart Glasses: A Museum Application," [Online]. Available: https://nsh423.github.io/assets/publications/paper_4_real_time_3d_generation_in_museum_AR.pdf
7. Y. Kang, S. Huang, and Y. Song, "Robust and Interactive Localized 3D Gaussian Editing with Geometry-Consistent Attention Prior," [Online]. Available: https://nsh423.github.io/assets/publications/paper_6_RoMaP.pdf