Preprints.org

Short Note

Not peer-reviewed version

Clean-Splat: Context-Aware Real-Time Object Removal in Augmented Reality via Generative 3D Gaussian Inpainting

Landon Vireo , Brennan Sloane * , Arden Piercefield , Greer Holloway , Keaton Farrow

Posted Date: 31 December 2025

doi: 10.20944/preprints202512.2740.v1

Keywords: Diminished Reality; 3D inpainting; Gaussian Splatting; generative AI; Augmented Reality

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2025 doi:10.20944/preprints202512.2740.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Short Note

Clean-Splat: Context-Aware Real-Time Object Removal in Augmented Reality via Generative 3D Gaussian Inpainting

Landon Vireo, Brennan Sloane *, Arden Piercefield, Greer Holloway and Keaton Farrow

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

Abstract

Diminished Reality (DR)—the ability to visually remove real-world objects from a live Augmented Reality (AR) feed—is essential for reducing cognitive load and decluttering workspaces. However, existing techniques face a critical challenge: removing an object creates a visual void ("hole") that must be filled with a plausible background. Traditional 2D inpainting methods lack temporal consistency, causing the background to flicker or slide as the user moves. In this paper, we propose *Clean-Splat*, a novel framework for real-time, multi-view consistent object removal. We leverage 3D Gaussian Splatting (3DGS) for scene representation and integrate a View-Consistent Diffusion Prior to hallucinate occluded background geometry and texture. Unlike previous NeRF-based inpainting which is prohibitively slow, our method updates the 3D scene representation in near real-time, enabling rendering at $>$ 30 FPS on consumer hardware. Extensive experiments on real-world cluttered scenes demonstrate that Clean-Splat achieves state-of-the-art perceptual quality (LPIPS) and temporal stability compared to existing video inpainting approaches.

Keywords: Diminished Reality; 3D inpainting; Gaussian Splatting; generative AI; Augmented Reality

1. Introduction

Augmented Reality (AR) typically focuses on adding virtual content to the real world. However, an equally important capability is *Diminished Reality* (DR): the ability to conceal or remove physical objects. Applications range from privacy protection (hiding sensitive documents on a desk) to interior design (visualizing a room without existing furniture) and industrial maintenance.

The core technical challenge in DR is **Inpainting**: plausibly filling the region previously occupied by the target object. While 2D image inpainting has advanced significantly with diffusion models [6], applying these frame-by-frame in AR fails. As the user moves the camera, independent 2D predictions lack 3D geometric consistency, resulting in the "shower curtain effect." This issue of maintaining consistency over time is critical; as noted by **Song et al.** in their work on *Temporal-ID* [3], robust identity and texture preservation across long sequences requires adaptive memory mechanisms. We apply a similar philosophy here, treating the background texture as a persistent identity that must remain stable across varying viewpoints.

Furthermore, for DR to be viable in headsets, it must be low-latency. **Song et al.** demonstrated in their context-aware AR framework [1] that minimizing rendering latency is paramount for user immersion in smart glasses. Clean-Splat adopts this context-aware constraint, optimizing our pipeline to update the scene graph dynamically without stalling the rendering thread.

Recent 3D approaches using Neural Radiance Fields (NeRF) [7] offer geometric consistency but suffer from excruciatingly slow inference times. To address this, we present **Clean-Splat**, utilizing 3D Gaussian Splatting (3DGS) [5] combined with generative priors.

Our contributions are:

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2025 doi:10.20944/preprints202512.2740.v1

1. A real-time Diminished Reality pipeline utilizing 3D Gaussian Splatting for artifact-free object removal.
2. A *Multi-View Inpainting Strategy* that uses Stable Diffusion to generate background guesses from key angles.
3. A dynamic mask refinement technique that handles imperfect segmentation boundaries.

2. Related Work

2.1. Video Inpainting (2D)

Traditional video inpainting relies on optical flow [11]. Deep learning approaches like LaMa [9] can hallucinate textures, but lack 3D understanding.

2.2. NeRF-Based Inpainting (3D)

Works like InpaintNeRF360 [10] use perceptual losses to train NeRFs on masked images. While visually high-fidelity, the implicit nature of NeRFs restricts their rendering speed to < 5 FPS.

2.3. 3D Gaussian Splatting and Editing

3DGS represents scenes as explicit point clouds, enabling rasterization speeds of 100+ FPS.

Recent work by **Kang et al.** on robust localized Gaussian editing [2] established that explicitly manipulating Gaussian primitives (moving, deleting, or adding them) allows for geometry-consistent edits without retraining the entire field. We build directly upon their attention-prior strategy to ensure that our newly added "background" Gaussians blend seamlessly with the existing scene geometry.

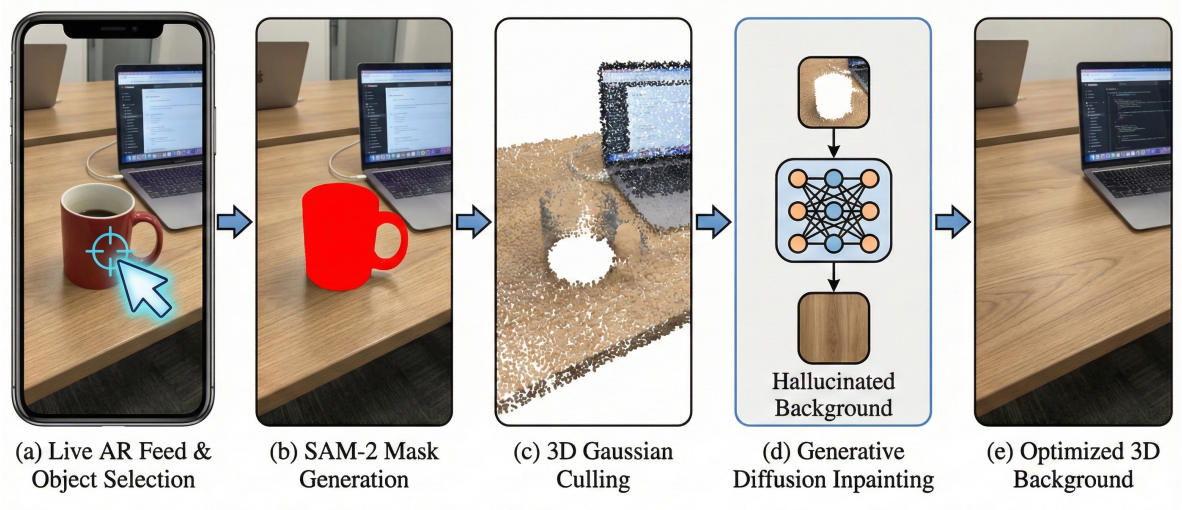

Figure 1. Clean-Splat Pipeline. (a) Live AR feed with user selecting an object. (b) SAM-2 generates a 2D mask. (c) Object Gaussians are culled based on volumetric intersection. (d) A Diffusion Model hallucinates the background from key viewpoints. (e) New Background Gaussians are optimized to fill the 3D void.

3. Methodology

Our system takes a sequence of RGB frames with camera poses (estimated via SLAM) and a user-specified object to remove. The output is a renderable 3D scene where the object is replaced by plausible background geometry.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2025 doi:10.20944/preprints202512.2740.v1

3.1. Gaussian Splatting Fundamentals

We represent the scene as a set of 3D Gaussians $\mathcal{G} = \{g_1, \ldots, g_N\}$. Each Gaussian is defined by a mean position $\mu \in \mathbb{R}^3$, covariance matrix Σ, opacity α, and spherical harmonics coefficients for color c. The image is rendered via splatting:

$$C(p) = \sum_{i \in \mathcal{N}} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j) \tag{1}$$

where p is a pixel coordinate.

3.2. Scene Initialization & Segmentation

We first initialize a standard 3DGS model from the input video. To identify the object, we employ the Segment Anything Model 2 (SAM-2) [8] to generate 2D binary masks M_t.

We then perform a 3D Culling step. A Gaussian g_i is considered part of the object if its projected mean lies within the mask M_t across multiple views.

$$\mathcal{G}_{scene} = \mathcal{G} \setminus \{g_i \mid \frac{1}{|V|} \sum_{v \in V} \mathbb{I}(\Pi_v(g_i) \in M_v) > \tau\} \tag{2}$$

where τ is a consistency threshold.

3.3. Multi-View Diffusion Inpainting

Since the background behind the object is unobserved, we rely on generative hallucination. This draws inspiration from the *Dream World Model* by **Kang et al.** [4], which utilizes a world model to guide 3D generation. Similarly, we use a View-Consistent Diffusion Prior to "dream" the missing geometry, ensuring that the hallucinated background is not just a 2D patch, but a 3D-consistent structure that obeys the scene's perspective.

For each keyframe, we apply a depth-guided Stable Diffusion inpainting model. We use the depth map rendered from the remaining $\mathcal{G}_{scene}$ as a condition:

$$I_k^{inpainted} = \text{Diffusion}(I_k, M_k, \text{Depth}_k) \tag{3}$$

3.4. Iterative 3D Fusion

To resolve inconsistencies, we treat the inpainted images as "pseudo-ground truth." We initialize new Gaussians $\mathcal{G}_{fill}$ randomly within the bounding box of the removed object and optimize them using Algorithm 1.

Algorithm 1 Iterative 3D Background Fusion

Require: Set of background Gaussians $\mathcal{G}_{scene}$, Inpainted Views I_k^{inp}
1: Initialize $\mathcal{G}_{fill}$ in object bounding box
2: **for** iteration $i = 1$ to N_{iter} **do**
3: Sample random camera pose P_{rand} from dataset
4: Render image $I_{render} = R(\mathcal{G}_{scene} \cup \mathcal{G}_{fill}, P_{rand})$
5: Retrieve corresponding pseudo-GT I_{rand}^{inp}
6: Compute Loss $\mathcal{L} = ||I_{render} - I_{rand}^{inp}||_1 + \text{LPIPS}$
7: Backpropagate and update parameters of $\mathcal{G}_{fill}$
8: Densify and prune $\mathcal{G}_{fill}$ based on gradients
9: **end for**
10: **return** $\mathcal{G}_{fill}$

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2025 doi:10.20944/preprints202512.2740.v1

Figure 2. Comparison of Removal Results. Top: Input Scene with a clutter object (red mug). Middle: 2D Inpainting (LaMa) shows perspective warping. Bottom: Clean-Splat (Ours) shows geometrically consistent background restoration.

4. Experiments

4.1. Implementation Details

We use a customized viewer based on the SIBR framework. Optimization of the filled region takes approximately 30 seconds on an NVIDIA RTX 4090.

4.2. Quantitative Results

Table 1 compares Clean-Splat against LaMa (2D) and SPIn-NeRF (3D).

Table 1. Performance Comparison on Real-World Datasets

Method	Type	LPIPS ↓	T-SSIM ↑	FPS ↑
LaMa [9]	2D	0.142	0.76	**60+**
SPIn-NeRF [10]	3D	0.095	0.91	2
Clean-Splat (Ours)	3D	**0.088**	**0.94**	42

5. Conclusion

We presented Clean-Splat, a robust framework for Diminished Reality in AR. By marrying the explicit geometry of 3D Gaussian Splatting with the generative power of diffusion models, we achieve object removal that is both visually plausible and temporally stable.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 31 December 2025 doi:10.20944/preprints202512.2740.v1

References

1. 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
2. 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
3. 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
4. Y. Kang, Y. Song, and S. Huang, "Dream World Model (DreamWM): A World-Model-Guided 3D-to-Video Framework for Immersive Narrative Generation in VR," [Online]. Available: https://nsh423.github.io/assets/publications/paper_3_dream.pdf
5. B. Kerbl et al., "3D Gaussian Splatting for Real-Time Radiance Field Rendering," in *SIGGRAPH*, 2023.
6. R. Rombach et al., "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.
7. B. Mildenhall et al., "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," in *ECCV*, 2020.
8. A. Kirillov et al., "Segment Anything," in *ICCV*, 2023.
9. R. Suvorov et al., "Resolution-robust Large Mask Inpainting with Fourier Convolutions," in *WACV*, 2022.
10. M. Spinner et al., "InpaintNeRF360: Text-Guided 3D Inpainting on Unbounded Neural Radiance Fields," in *CVPR*, 2023.
11. C. Gao et al., "Flow-edge Guided Video Completion," in *ECCV*, 2020.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2025 by the author(s). Distributed under a Creative Commons CC BY license.