Article

AI-Augmented Context-Aware Generative Pipelines for 3D Content

Sining Huang *, Yixiao Kang , Geyu Shen , Yukun Song

Posted Date: 4 August 2025

doi: 10.20944/preprints202508.0195.v1

Keywords: 3D scene generation; neural radiance fields; stylization; InstructPix2Pix; SIGNeRF

 doi:10.20944/preprints202508.0195.v1

Article

AI-Augmented Context-Aware Generative Pipelines for 3D Content

Sining Huang [1,*], Yixiao Kang [1], Geyu Shen [2] and Yukun Song [1]

[1] University of California, Berkeley, USA;
[2] Georgia Institute of Technology, Atlanta, USA
* Correspondence: sining_huang@berkeley.edu

Abstract

This work presents a system for transforming scenes composed of basic 3D primitives into high-fidelity 3D environments using advanced representations such as Neural Radiance Fields (NeRFs) and 3D Gaussian Splatting. The proposed approach leverages existing image stylization models and image-to-3D generative techniques to construct a pipeline that iteratively stylizes individual objects and integrates them into complex scenes. The effectiveness of the system is demonstrated through the progressive insertion of generated objects into a base scene, followed by an analysis of current limitations and potential directions for future development.

Keywords: 3D scene generation; neural radiance fields; stylization; InstructPix2Pix; SIGNeRF

1. Introduction

We present a pipeline that lets non-experts furnish and restyle scanned interior spaces through an intuitive interface. Users sketch simple 3D primitives and provide natural-language style prompts; the system converts these inputs into fully furnished, stylized room models with real-time visualization. The method integrates InstructPix2Pix for image-guided stylization and SIGNeRF for seamless object insertion into NeRF-based scenes, abstracting away technical complexity. We detail the architecture and implementation and show experiments that produce immersive, coherent results. This approach lowers the barrier to high-fidelity 3D content creation for VR/AR and broadens access to advanced scene generation.

2. Background

2.1. 3D Object Generation

Advances in neural representations and generative models have significantly accelerated 3D content generation, enabling the creation of high-quality and diverse assets. Representation methods include Neural Scene Representations, Explicit Representations, Point Clouds, Meshes, Multi-layer, and Implicit Representations. Among these, Neural Radiance Fields (NeRFs)[1] and Gaussian Splatting[2] are particularly impactful. NeRFs employ compact neural networks to reconstruct scenes by predicting light intensity and color from any direction.

Notable among recent methods is the Convolutional Reconstruction Model (CRM)[3], which generates six orthographic views from a single image to produce high-quality 3D models. Similarly, Triplane[4] and the Gaussian Reconstruction Model (GRM) [5] demonstrate strong performance in producing Gaussian Splatting-based representations. These models highlight the field's rapid progression toward more realistic and detailed 3D asset generation.

2.2. 3D Scene Generation and Editing

NeRFs have revolutionized 3D scene reconstruction and novel view synthesis but remain challenging to edit. Early methods such as NeRF-Editing [6] supported only basic deformations, while

 doi:10.20944/preprints202508.0195.v1

NeuMesh [7] introduced more advanced operations like texture and geometry editing. Despite these improvements, current NeRF editing tools still lack the flexibility and ease of traditional 3D modeling software.

Recent generative approaches combine text-to-3D generation with NeRF editing. Set-the-Scene [8] and Compositional 3D [9] enable structured scene creation using proxy objects. Instruct-NeRF2NeRF [10] further refines this with an Iterative Dataset Update (IDU) strategy that applies InstructPix2Pix [11] for controlled scene modifications.

3. Method

This section outlines the proposed pipeline (Figure 1) for transforming basic 3D primitives into stylized furniture guided by user-provided text prompts, and integrating the resulting objects into a target 3D scene. The system consists of three primary components:

1. **Primitives Stylizer**, which takes a single-view image of a primitive and produces a stylized version guided by a text prompt;
2. **Mesh Generator**, which converts the stylized image into a corresponding textured 3D mesh;
3. **Scene Integrator**, which incorporates the generated mesh into the target environment.

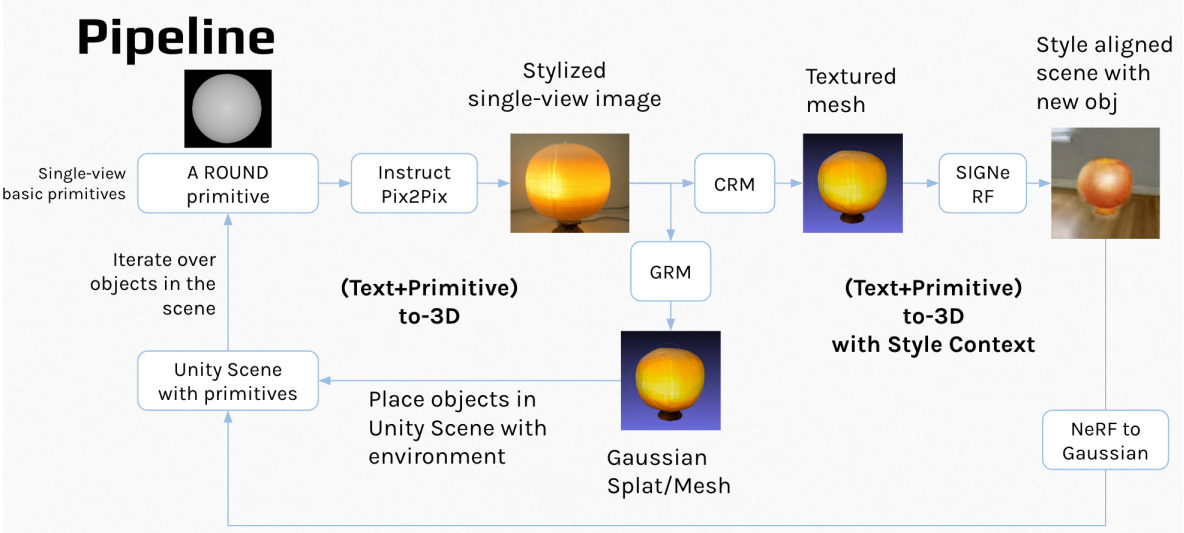

Figure 1. Stylized 3D Object Generation and Scene Integration Pipeline

We detail each of these components in the following subsections.

3.1. Primitives Stylizer

To stylize a primitive object based on textual input while preserving its underlying geometric structure, the system employs InstructPix2Pix [11], a state-of-the-art text-guided image editing model. InstructPix2Pix is particularly well-suited for this task due to its capability to selectively modify image regions according to natural language instructions, while preserving the integrity of unedited areas.

Trained on a large and diverse dataset of image editing tasks, InstructPix2Pix can produce high-quality, photorealistic stylizations that align with user intent. This allows us to effectively map the basic appearance of a primitive (e.g., a round shape) into a semantically meaningful object (e.g., a "Japanese paper lantern") with coherent lighting, texture, and visual style. An example output is shown in Figure 2.

The model's flexibility and generalization capabilities across a wide range of instructions make it an ideal choice for the Primitives Stylizer component. It enables interactive and intuitive customization of base primitives, which serves as the foundation for subsequent 3D generation.

 doi:10.20944/preprints202508.0195.v1

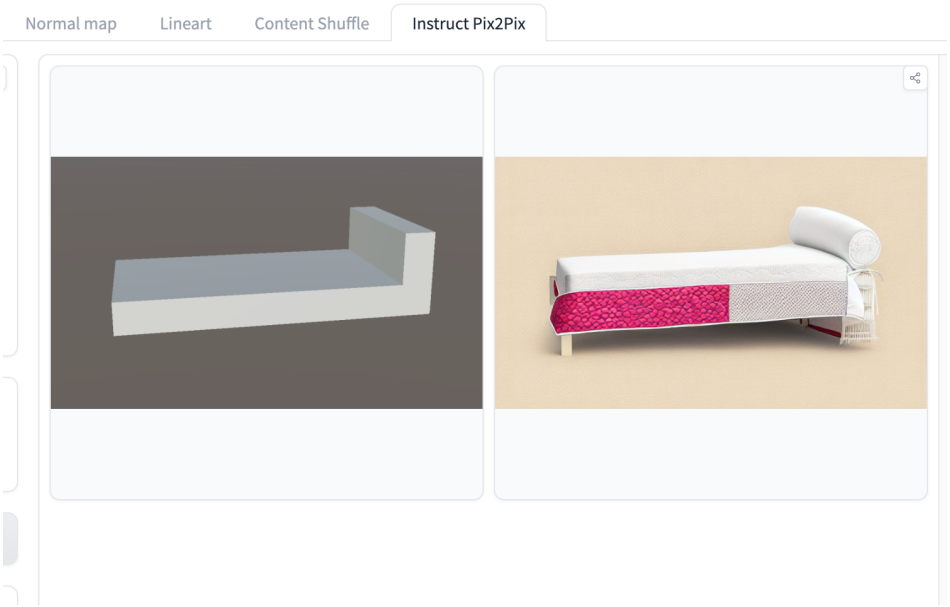

Figure 2. Result from InstructPix2Pix with the text prompt: "a modern bed in the apartment, clean background".

3.2. Mesh Generator

After obtaining a stylized single-view image from the Primitives Stylizer, the next step is to generate a textured 3D mesh that preserves view consistency, minimizes photometric loss with respect to the input view, and closely resembles real-world object geometry. The underlying 3D representation can be either surface-based (e.g., SDF, mesh) or volumetric (e.g., 3D Gaussians, NeRF). In this work, we explore two state-of-the-art models for mesh generation: the Convolutional Reconstruction Model (CRM) [3] and the Gaussian Reconstruction Model (GRM) [5].

3.3. Scene Integrator

To integrate the generated mesh into the target scene, the system employs SIGNeRF [12], a framework designed to support consistent multi-view scene editing using 2D image-based techniques within a 3D NeRF-based environment. SIGNeRF leverages the fact that NeRFs, trained from multi-view 2D images, implicitly establish a correspondence between 2D observations and the underlying 3D structure. This enables the application of powerful 2D editing tools—such as ControlNet [13]—to effect coherent and view-consistent modifications across the 3D scene.

 doi:10.20944/preprints202508.0195.v1

Figure 3. We combine the generated meshes from GRM in the Unity Scene

4. Experiments

4.1. Evaluation Setup

We evaluated our pipeline on an empty apartment scene with three objects of different shapes and sizes that are commonly found in apartments: (1) a sofa, (2) a lamp, and (3) a bed. These three objects are iteratively added into the apartment scene at different locations, simulating how users would like to iteratively add their objects to their target scene. After adding each object, SIGNeRF retrains the scene to stylize the object according to the scene background.

We capture the empty apartment scene with an iPhone and train an initial NeRF scene with Nerfstudio [14]. The collected NeRF dataset has a total of 303 images and is trained in 15 mins. This scene will be used as the base for our pipeline to add objects into. Then, we utilized our proposed pipeline to add the three objects iteratively into the scene in the order of: (1) a sofa, (2) a lamp, and (3) a bed. These objects are added using the first way to add objects in the scene mentioned in section 3.3, which is to directly add new images of the object into the original NeRF dataset.

4.2. Results

Table 1. Time taken for each step of the pipeline.

Object	Primitive -Stylization (s)	Mesh Generation (s)	SIGNeRF (min)	Total (min)
Sofa	16.7	30	28.3	29.1
Lamp	18.1	30	29.1	29.9
Bed	15.3	30	30.2	31.0

We show the before-after comparison of the apartment scene in Figure 4. In addition, we show the generated results for each of the three objects for each step of the entire pipeline in Figures 5 and 6. As seen in Figure 4, the sofa and lamp object appears translucent. This happens because they are the first two objects added into the scene. Each time we iteratively add an object into the scene, the conditioning signals such as depth and mask information of prior objects are lost. Moreover, since we utilized the first way to add objects into a scene with SIGNeRF, only the newly added images to the

 doi:10.20944/preprints202508.0195.v1

NeRF dataset contains the new object. The original images from the NeRF dataset are not updated, even at locations where the new object is supposed to be located. Thus, these factors contribute to the blurry and translucent results generated.

(a) Initial apartment scene.

(b) Apartment with sofa, bed and lamp added.

Figure 4. Comparison of apartment scene before and after adding objects using our pipeline.

From Figure 6 we can observe that the objects are not consistent across different view angles, especially for the bed and for certain angles of the sofa. Upon further investigation, we discover that this is due to the inconsistencies of the generated dataset by SIGNeRF across different view angles. This can be seen from Figure 7, where the bedsheets look white from certain angles but brown with a wood-like texture from other angles. As for the sofa, it is green from most view angles but the side, which looks gray. This shows that even by conditioning the Control-Net stylization through reference grids, it is not robust enough to produce view-consistent results.

 doi:10.20944/preprints202508.0195.v1

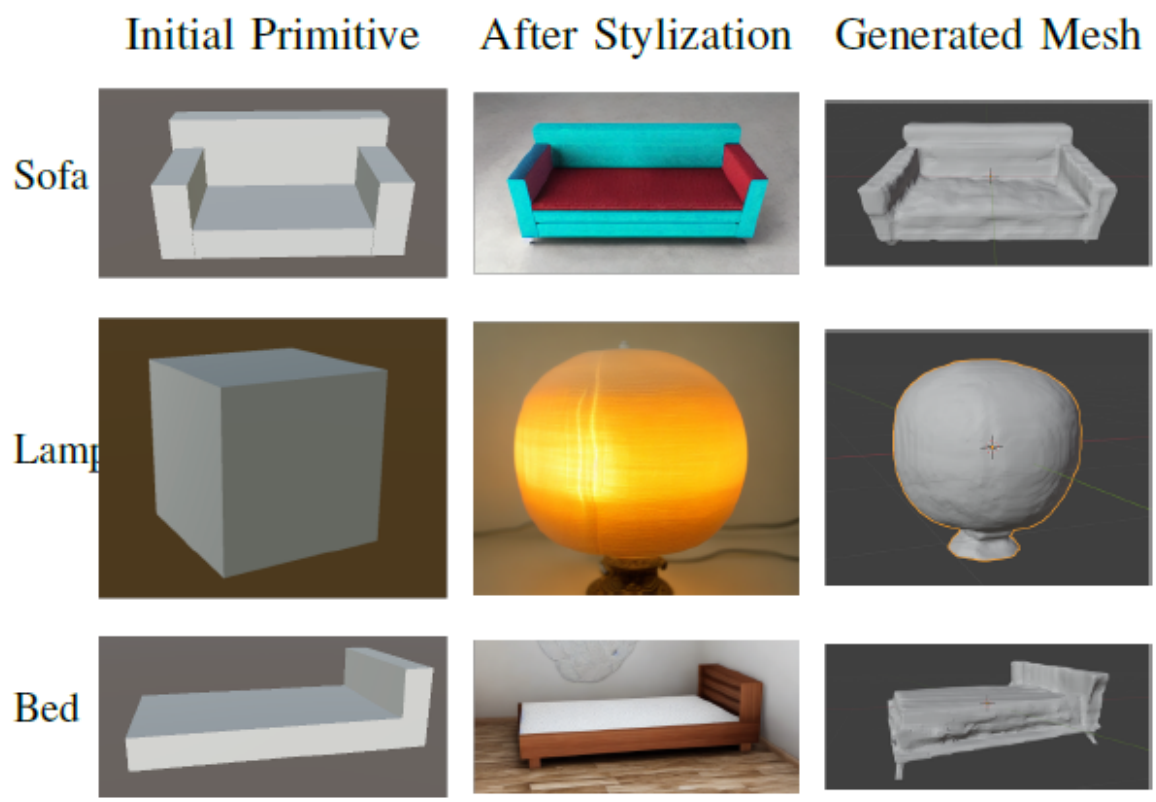

Figure 5. Results from each step in the pipeline before SIGNeRF.

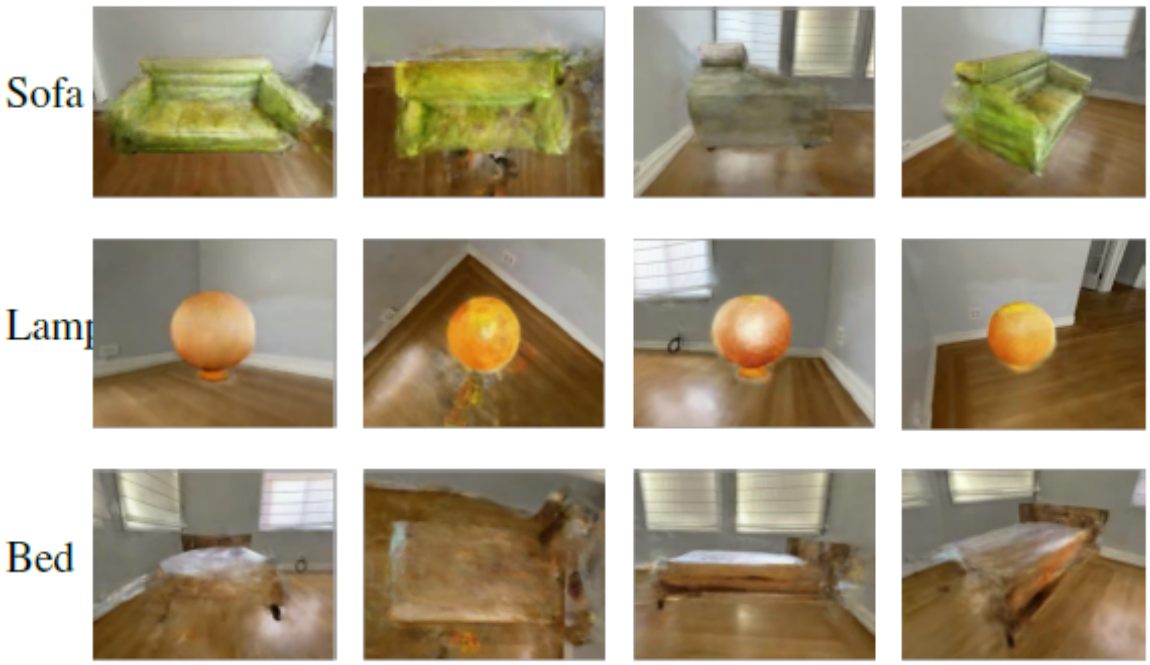

Figure 6. Three objects generated from our pipeline viewed at different view angles.

 doi:10.20944/preprints202508.0195.v1

Figure 7. Inconsistencies of the generated dataset from SIGNeRF across different view angles.

Given the unsatisfactory blurriness and view-inconsistencies observed in the outputs from the previous pipeline, an alternative approach was implemented. The model generated by the GRM was integrated into the Unity scene, replacing the primitive forms, as depicted in Figure 3. The resulting meshes are robust, allowing for user interactions such as moving and scaling. However, the material and color attributes of the objects, determined by the ControlNet results, do not correspond with their environmental context. This contrasts with our earlier pipeline, which stylized objects in accordance with the surrounding environment.

5. Conclusion

Overall, the system presents a practical and effective approach for generating high-fidelity 3D objects from basic primitives and integrating them into NeRF or 3D Gaussian Splatting (3DGS) scenes. By combining 2D image stylization models with image-to-3D reconstruction techniques, the proposed pipeline establishes a bridge between primitive-based design and realistic, stylized scene composition. Despite current challenges—particularly in view consistency and generation quality—the system provides a flexible foundation that can support and inform future research in interactive 3D scene design, object stylization, and accessible content creation workflows.

References

1. K. Gao, Y. Gao, H. He, D. Lu, L. Xu, and J. Li, "Nerf: Neural radiance field in 3d vision, a comprehensive review," *arXiv preprint arXiv:2210.00379*, 2022.
2. B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.
3. Z. Wang, Y. Wang, Y. Chen, C. Xiang, S. Chen, D. Yu, C. Li, H. Su, and J. Zhu, "Crm: Single image to 3d textured mesh with convolutional reconstruction model," *arXiv preprint arXiv:2403.05034*, 2024.
4. J. R. Shue, E. R. Chan, R. Po, Z. Ankner, J. Wu, and G. Wetzstein, "3d neural field generation using triplane diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 875–20 886.
5. Y. Xu, Z. Shi, W. Yifan, H. Chen, C. Yang, S. Peng, Y. Shen, and G. Wetzstein, "Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation," *arXiv preprint arXiv:2403.14621*, 2024.

 doi:10.20944/preprints202508.0195.v1

6. Y.-J. Yuan, Y.-T. Sun, Y.-K. Lai, Y. Ma, R. Jia, and L. Gao, "Nerf-editing: geometry editing of neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 353–18 364.
7. B. Yang, C. Bao, J. Zeng, H. Bao, Y. Zhang, Z. Cui, and G. Zhang, "Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing," in *European Conference on Computer Vision*. Springer, 2022, pp. 597–614.
8. D. Cohen-Bar, E. Richardson, G. Metzer, R. Giryes, and D. Cohen-Or, "Set-the-scene: Global-local training for generating controllable nerf scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 2920–2929.
9. R. Po and G. Wetzstein, "Compositional 3d scene generation using locally conditioned diffusion," *arXiv preprint arXiv:2303.12218*, 2023.
10. A. Haque, M. Tancik, A. A. Efros, A. Holynski, and A. Kanazawa, "Instruct-nerf2nerf: Editing 3d scenes with instructions," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 19 740–19 750.
11. T. Brooks, A. Holynski, and A. A. Efros, "Instructpix2pix: Learning to follow image editing instructions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 18 392–18 402.
12. J.-N. Dihlmann, A. Engelhardt, and H. Lensch, "Signerf: Scene integrated generation for neural radiance fields," *arXiv preprint arXiv:2401.01647*, 2024.
13. L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 3836–3847.
14. M. Tancik, E. Weber, E. Ng, R. Li, B. Yi, J. Kerr, T. Wang, A. Kristoffersen, J. Austin, K. Salahi, A. Ahuja, D. McAllister, and A. Kanazawa, "Nerfstudio: A modular framework for neural radiance field development," in *ACM SIGGRAPH 2023 Conference Proceedings*, ser. SIGGRAPH '23, 2023.