Investigating Geometric Conditioning for Neural Rendering Artifact Correction

Alvaro Lopez, and Mohammadreza Tavasoli Naeini, and Oleksii Nakhod

Abstract—High-quality reconstruction and artifact removal are critical challenges in computational imaging pipelines, particularly for 3D Gaussian Splatting (3DGS) and Neural Radiance Fields (NeRF), where imperfect optimization often leads to distortions in novel-view synthesis (NVS). Recent image-repair models, such as Difix3D+, demonstrate strong priors for correcting view-dependent artifacts by leveraging large-scale diffusion models. However, these methods typically operate exclusively in the RGB domain, remaining blind to the underlying scene geometry.

In this work, we conduct a systematic investigation into whether augmenting Difix3D+ with explicit geometric and structural cues can improve its repair capabilities. We introduce depth- and edge-aware variants of the pipeline by integrating lightweight T2I-Adapters that condition the single-step diffusion model on depth maps and Canny edge maps. Our study encompasses multiple settings: depth estimation on DL3DV, controlled experiments on NeRF synthetic scenes with ground-truth geometry, and evaluations on real-world ScanNet++ data.

While it is commonly assumed in generative AI that increased control enhances performance, leading one to expect that geometric conditioning would offer clearer guidance, our results reveal a more nuanced reality: naive geometric conditioning actively degrades 3D reconstruction quality through a phenomenon we term *Artifact Propagation*, where errors in rendered geometry contaminate the conditioning signal and force the repair model to solidify rather than eliminate artifacts. Crucially, these findings offer valuable diagnostic insights. By identifying the limitations of straightforward conditioning, our work establishes a foundation for future research to develop more principled and robust strategies for integrating geometry into diffusion-based view repair.

Index Terms—Neural Rendering, 3D Gaussian Splatting, Diffusion Models, Image Restoration, Geometric Conditioning, T2I-Adapter.

1 Introduction

The rapid evolution of neural rendering techniques, most notably Neural Radiance Fields (NeRF) [1] and 3D Gaussian Splatting (3DGS) [2], has enabled the synthesis of photorealistic novel views from sparse image collections. Recent advances in the field, such as pixelSplat [3] and MVSplat [4] have shown that it is possible to reconstruct 3D scenes via 3D Gaussian Splatting from sparse input images, allowing for fast novel-view synthesis (NVS). However, these methods are not without flaws. In practical scenarios—characterized by sparse views, unconstrained illumination, or textureless regions—reconstructions often exhibit characteristic defects. These include "floaters" (density accumulated in empty space), "foggy" geometry, and high-frequency aliasing or blurring in novel viewpoints.

Addressing these artifacts typically involves either constraining the optimization process (via regularization terms) or applying post-hoc image repair. The latter approach has gained traction with the advent of generative diffusion models. Methods like Difix3D+ [5] utilize distilled Stable Diffusion models (specifically SD-Turbo) to act as a restoration prior, mapping degraded renders to cleaner counterparts in a single forward pass. These repaired images can then be used to supervise the underlying 3D representation, closing the loop.

However, existing repair pipelines, including the baseline Difix3D+, operate primarily in the RGB domain. They treat the rendered image as a generic 2D degradation problem, ignoring the rich 3D structural information available in the rendering pipeline. This theoretical limitation becomes apparent in geometrically complex regions—such as occlusions, thin structures, or depth discontinuities—where the RGB model may hallucinate textures that are inconsistent with the scene's perspective.

In this work, we examine the commonly held assumption that geometric conditioning can enhance neural rendering repair. Depth or edge cues are often believed to provide useful structural constraints, potentially guiding the diffusion model toward geometrically consistent corrections. To test this assumption, we modify Difix to incorporate T2I-Adapter-based [6] structural guidance.

Contributions. The central question addressed by this project is whether naively injecting geometric cues into a neural rendering repair module actually helps, especially when the underlying 3D geometry is imperfect. Our contributions are threefold:

- **Geometry-aware Difix via modular conditioning.** We extend the Difix image-repair module with lightweight T2I-Adapter branches that condition a single-step diffusion model on depth maps, Canny edge maps, or both. The resulting architecture is simple, modular, and designed to plug into NeRF- and 3DGS-style pipelines where novel-view artifacts and geometry errors are common.
- **Comprehensive multi-regime evaluation.** We conduct a systematic empirical study on both synthetic (NeRF-Blender) and real (DL3DV, ScanNet++) RGB-D data, comparing RGB-only, depth-guided, edge-guided, and depth+edge-guided variants un-

• *The authors are with the Department of Computer Science, University of Toronto.*

der idealized (clean geometry) and realistic (artifact-contaminated geometry) regimes. Across these settings, naive geometric conditioning fails to deliver robust gains in PSNR/SSIM/LPIPS and, when used to supervise 3DGS, often degrades novel-view quality and introduces new artifacts.
- **Identification of the *Artifact Propagation* failure mode.** We empirically identify and characterize a failure mode in which geometric artifacts in the rendered depth/edge maps are amplified by the conditioning pathway, causing the repair model to "solidify" floaters and other errors rather than remove them. We relate this phenomenon to the interaction between RGB priors and geometric cues in diffusion-based repair, and distill practical design guidelines for when geometric information is likely to help, when it is likely to hurt, and what conditioning strategies may be needed in future work to safely exploit geometry.

2 Related Work

2.1 Image Restoration with Conditional Diffusion Models

A common paradigm in contemporary image synthesis leverages conditional diffusion models that generate output images guided not only by text or image prompts, but also by structural inputs such as depth maps, edges, segmentation masks, or pose skeletons. Two representative frameworks in this space are ControlNet [7] and T2I-Adapter [6]. ControlNet augments large, frozen text-to-image diffusion backbones with trainable encoding modules connected via zero-initialized convolutions, enabling the model to respect spatial/structural constraints like depth or edge maps during generation. Similarly, T2I-Adapter learns lightweight adapters that align a pretrained T2I diffusion model's internal representation with external control signals, allowing color, layout, or structural control without retraining the full model. Importantly, both ControlNet and T2I-Adapter have been developed and most commonly demonstrated in standard text-to-image or image-to-image pipelines, where the conditioning input typically comes from a clean, user-provided structural image (e.g., a carefully prepared depth or edge map). For instance, a user provides a clean depth map to ensure the generated image respects that geometry. Our application is distinct: we use these techniques not for image generation from a clean prior, but for *repair* of an existing rendered output. This shifts the role of the control signal from a definitive prior to a potentially degraded feature extracted from a noisy source.

2.2 Artifact Correction in Neural Rendering

Artifact mitigation in 3D reconstruction falls into two main categories: in-optimization and post-hoc methods.

2.2.1 In-Optimization Methods

For neural-rendering methods such as NeRF and its derivatives (e.g., Mip-NeRF 360), early work focused on improved sampling, anti-aliasing, and regularization strategies (e.g., distortion-aware regularization and non-linear scene parameterization in Mip-NeRF 360 [8]) to suppress aliasing, sampling artifacts, and overfitting in challenging scenes. For approaches such as 3D Gaussian Splatting (3DGS), techniques involve regularizing Gaussian parameters (e.g., density, opacity, scale) and applying adaptive density/opacity control to avoid "floaters" and maintain coherent geometry, particularly under sparse-view or limited-data conditions. While effective, these techniques typically need careful per-scene tuning, potentially long optimization times, and can still struggle when data is limited.

2.2.2 Post-Hoc Methods

These methods treat the output as an image processing problem. Difix3D+ [5] pioneered the use of diffusion models for this task by training a model to translate a rendered image (from a potentially flawed 3DGS model) into a clean output image using an LPIPS and L2 loss. The core appeal is speed: once trained, the model offers instant image quality improvement without modifying the underlying 3D structure. However, previous iterations were strictly RGB-to-RGB, treating all artifacts purely as photometric noise regardless of their 3D origin.

2.3 Novel-view Synthesis Using Diffusion Models

Recent advances leverage diffusion models for novel-view synthesis. For instance, RenderDiffusion [9] introduces a 3D-aware diffusion framework that learns a latent 3D scene representation, trained using monocular 2D supervision. It renders it to generate consistent novel views of a single object. MOVIS [10] extends view-conditioned diffusion to multi-object and indoor scenes by injecting structure-aware features, like depth and object masks, into the model's U-Net, improving object placement and consistency. More recently, SplatDiff [11] introduced a pixel-splatting-guided video diffusion model that combines pixel splatting with a video diffusion model to merge geometry consistency with high-fidelity texture generation, achieving state-of-the-art results in single-view NVS. These works collectively show that applying the premise of combining geometry-aware representations with diffusion-based learned priors can improve NVS pipelines.

2.4 Structure-aware Regularization

Methods such as PlaNeRF [12] propose approaches to improve geometry reconstruction in Neural Radiance Fields (NeRF) without the need for explicit 3D supervision. Specifically, PlaNeRF uses a Singular Value Decomposition (SVD)-based plane regularizer that encourages reconstructed geometry to respect planar structure in the scene, relying only on RGB inputs and semantic maps. This method addresses common NeRF failure modes, such as overfitting and poor geometry in textureless or low-detail regions, and demonstrates improved structural accuracy on large-scale scenes. By doing so, PlaNeRF exemplifies a broader class of works that augment appearance-based reconstruction with implicit structural priors to yield more geometrically reliable outputs — an approach conceptually aligned with, but more principled than, naive conditioning strategies such as ours.

2.5 Acquiring Geometric Priors

To take advantage of geometric conditioning, accurate depth maps are required.

1) **Ground Truth from Rendering:** For synthetic scenes (e.g., NeRF Blender), one can obtain exact depth via ray casting. However, if the 3D representation (e.g., a 3DGS reconstruction) is imperfect — e.g., due to sparse views, under-constrained optimization, or density "floaters" — the resulting depth map may inherit those geometric errors (floating geometry, holes, incorrect depth values).
2) **Monocular Depth Estimation:** Foundation models like Depth Anything V2 [13] can predict high-quality, view-consistent depth from a single RGB image. That said, their behavior on images containing 3D reconstruction artifacts (like dense floaters) remains unverified and is a potential failure mode given that such distortions were likely absent from their training data, so depth predictions in those cases might be unreliable.

Our work integrates the rendered output, the core element of the artifact, back into the conditional input path, which fundamentally distinguishes our study from previous applications of T2I-Adapters.

3 Methodology and Architecture Design

Our methodology centers on extending the single-step Difix [5] framework with geometric conditioning using T2I-Adapters (Fig. 1).

3.1 Backbone Difix Architecture

We build on a fine-tuned Stable Diffusion Turbo (SD-Turbo) [14] model in the single-step Difix configuration. Given a rendered input image $I_{\text{render}} \in \mathbb{R}^{H\times W\times 3}$, a frozen VAE encoder maps it to a latent $z \in \mathbb{R}^{H/8\times W/8\times 4}$, and a U-Net predicts the noise at a single timestep t:

$$z = E_{\text{VAE}}(I_{\text{render}}), \qquad \epsilon_\theta(z, t, c_{\text{text}}) = \text{U-Net}(z, t, c_{\text{text}}),$$

where c_{text} is a text embedding (we use a fixed prompt as in Difix). The scheduler produces a denoised latent z_{clean}, which the decoder D_{VAE} maps back to the repaired image $\hat{I}$.

3.2 Geometry-aware Conditioning via T2I-Adapters

To inject structural cues, we attach T2I-Adapter branches [6] to the frozen U-Net encoder. Each branch takes as input a geometric condition $C_{\text{geo}} \in \{D, E\}$, where D is a normalized depth map and E is a Canny edge map [15]. The adapter produces multiscale feature maps $A = \{A_1, \ldots, A_4\}$ aligned with the encoder resolutions.

For encoder block i, we modify the activations as

$$F_i^{\text{U-Net}} \leftarrow F_i^{\text{U-Net}} + W_{\text{res}}\, A_i,$$

where $F_i^{\text{U-Net}}$ is the original feature map and W_{res} is a learned 1×1 convolution that matches channel dimensions. For the depth+edge variant, we use two independent adapters and sum their outputs:

$$F_i^{\text{U-Net}} \leftarrow F_i^{\text{U-Net}} + (F_D)_i(D) + (F_E)_i(E).$$

This design keeps the SD-Turbo backbone frozen and makes geometric conditioning fully modular.

3.3 Training Objective

We train the adapters and VAE decoder end-to-end to reconstruct a clean target image I_{GT} from a degraded render. Following Difix, we use a weighted combination of pixel-wise, perceptual, and style losses:

$$\begin{aligned}\mathcal{L}_{\text{total}} = \lambda_{\text{L2}}\, \|\hat{I} - I_{\text{GT}}\|_2^2 \\ + \lambda_{\text{LPIPS}}\, \text{LPIPS}(\hat{I}, I_{\text{GT}}) \\ + \lambda_{\text{Gram}}\, \mathcal{L}_{\text{Gram}}(\hat{I}, I_{\text{GT}}),\end{aligned}$$

where LPIPS [16] and the Gram-matrix style loss [17] are computed on VGG features. We set all weights to 1.0 and apply a short warm-up for the style term so that the model first learns basic reconstruction.

3.4 Training Data and Degradation Model

To mimic typical artifacts seen in neural rendering, we create degraded inputs by applying stochastic blur and rescaling to the ground-truth images I_{GT}: Gaussian blur with a random kernel radius in $[1, 3]$ (80% probability), and random down-sampling by a factor in $[2, 4]$ followed by nearest-neighbor up-sampling (50% probability). These corrupted views serve as I_{render} for the restoration stage.

4 Experimental Evaluation

Our evaluation proceeded in three logical phases, starting from unconstrained data and moving to controlled, ground-truth-heavy environments.

4.1 Phase 1: Depth Acquisition Challenges in Unconstrained Data

The initial experimental setup mirrored the application environment: using the DL3DV [18] dataset, which contains complex real-world scenes captured from multiple camera positions, but lacks reliable ground truth depth maps.

4.1.1 Depth Generation Pipeline

To generate training and conditioning data, we deployed two state-of-the-art approaches:

1) **3D-Rendered Depth:** A 3DGS model was trained on the scene. The predicted depth map D_{3D} was rendered directly from the optimized gaussians.
2) **Monocular Depth Estimation:** The input RGB views were processed by Depth Anything V2 to generate robust monocular depth D_{Mono}.

4.1.2 Observed Behavior

This phase revealed a critical challenge for geometric conditioning. The initial 3DGS reconstruction suffered heavily from floating geometry and incomplete surface coverage, particularly in the DL3DV scenes.

- D_{3D} contained severe geometric noise (e.g., floaters in the sky appeared as sharp, isolated objects at fixed depths). When provided as a condition, this noisy

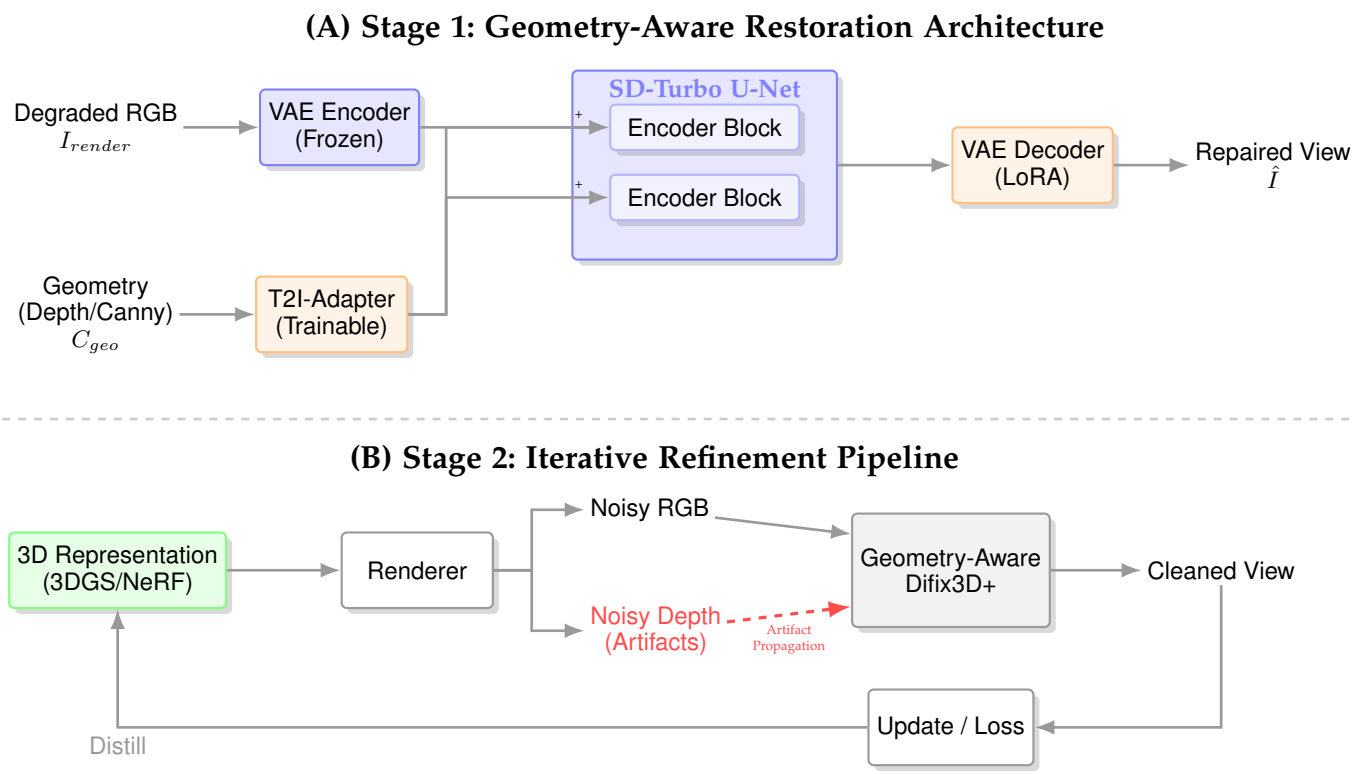
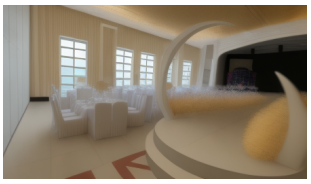


Fig. 1. **The Geometry-Aware Difix3D+ Framework. (A) Architecture:** We inject geometric cues (Depth/Edges) into a frozen SD-Turbo U-Net via trainable T2I-Adapters. **(B) Refinement Pipeline:** In the 3D optimization loop, rendered views are repaired by the model.

prior corrupted the learning objective, as the model was implicitly taught that these errors represented valid, textured geometry.
- D_{Mono}, while cleaner than D_{3D}, often struggled with occlusions and view-inconsistency due to monocular hallucination.

Due to the difficulty in disentangling artifact noise from conditioning corruption in this data-rich, but noisy, setup, we pivoted to a controlled environment.

TABLE 1
Novel view synthesis for a sample DL3DV scene

Target	Difix + T2I Output
	

4.2 Phase 2: Controlled Ablation on Synthetic Data

We performed ablation studies using the high-quality NeRF Synthetic Blender dataset (e.g., Lego, Drums, Materials), which provides perfect ground truth RGB and depth maps.

To rigorously evaluate the impact of geometric conditioning in isolation, we first conducted a Stage 1 evaluation. In this setting, the restoration models were trained and tested on pairs of ground truth images and their *artificially degraded* counterparts (simulated via Gaussian blur and random downsampling) rather than actual 3DGS renders. This allows us to assess the intrinsic image-repair capabilities of the conditioned diffusion models before integrating them into the iterative 3D optimization loop. For this experiment, all models were trained for 2000 steps using the same hyperparameters.

4.2.1 Metrics and Models (Stage 1)

We evaluate four variants against the ground truth I_{GT} on the test split of the synthetic dataset. Note that the metrics reported in Table 2 reflect this single-step image restoration task:

- **Baseline ($Difix$):** RGB-only (no geometric input).
- **Depth:** Conditioned on ground truth depth map.
- **Canny:** Conditioned on Canny edge map (extracted from the degraded input).
- **Depth + Canny:** Conditioned on both fused modalities.

Table 2 summarizes the performance across multiple standard image quality metrics for this restoration task.

Quantitative Analysis: On this controlled denoising experiment using perfect NeRF synthetic data, all geometry-aware variants remain very close to the RGB-only baseline. In fact, both the Canny and depth + Canny models achieve small but consistent improvements across PSNR, SSIM, LPIPS, and FID (Tables 2 and 3).

This indicates that the T2I-Adapter can make effective use of clean and reliable structural cues when the degradation is simple (blur/noise) and the underlying geometry is accurate.

TABLE 2
Stage 1 Quantitative Comparison: Restoration of Artificially Degraded NeRF Synthetic Scenes

Model Variant	PSNR ↑	SSIM ↑	LPIPS ↓	FID ↓
(Baseline)	31.8619	0.9632	0.0279	15.5294
Depth	31.8545	0.9624	0.0282	15.9502
Canny	**32.1287**	**0.9642**	**0.0272**	**14.7017**
Depth + Canny	32.1070	0.9638	0.0273	15.5019

TABLE 3
Image reconstruction results for NeRF Drums (Stage 1)

Target	Base (PSNR 31.64)	Canny (PSNR 32.32)
		

However, these per-view gains in Stage 1 do not translate into better 3D reconstruction. When using the repaired images to supervise 3D Gaussian Splatting on the Materials scene (Stage 2), the geometry-aware corrections, particularly depth + Canny, caused worse accumulated geometry and lower-quality novel views compared to the RGB-only baseline (Table 4).

In other words, optimizing local 2D image metrics at the Difix stage can still harm the end-to-end 3D pipeline by reinforcing or amplifying erroneous geometric signals, even when the per-frame metrics appear better.

4.3 Phase 3: Real-World Scans and Zero-Depth Conditioning (ScanNet++)

In the final phase, we extended our evaluation to the real-world ScanNet++ [19] dataset. Having observed in Phase 2 that conditioning on corrupted depth actively harms reconstruction via *Artifact Propagation*, we designed an experiment to test if the model could be trained to selectively utilize geometric priors only when they are reliable.

4.3.1 Zero-Depth Masking Strategy

We modified the Difix pipeline to accept a depth channel injected directly into the U-Net without the T2I-Adapter. However, unlike previous phases where we fed the corresponding (noisy) rendered depth during training, we implemented a depth masking strategy to decouple geometric noise from texture generation:

- **Ground Truth Path:** When the model calculates loss against the clean target image I_{GT}, we condition the adapter on the high-quality sensor depth D_{GT}. This teaches the adapter valid structural associations.
- **Degraded Source Path:** When the model receives the degraded input image (simulating a poor render), we explicitly set the input depth map to a **zero-tensor** ($D = \mathbf{0}$).

The rationale was to prevent the model from overfitting to the specific artifacts of a flawed depth estimator. By training the model to reconstruct images under two distinct regimes, perfect depth guidance and "blind" (zero-depth) restoration, we hypothesized the adapter would learn to act as a conditional enhancement that only activates when valid structure is present, reverting to RGB-only behavior otherwise.

As illustrated in Table 5, the restoration model converges effectively: it successfully learns to map degraded, low-resolution inputs to sharp, clean outputs while handling the zero-depth token without instability. The reconstruction quality in isolation is high, suggesting the model learned the dual-modality task.

4.3.2 Results and Analysis

Contrary to our hypothesis, this masked training strategy yielded worse performance than the baseline RGB-only Difix model. We observed that the adapter failed to generalize effectively between the "perfect depth" and "zero depth" modalities.

We attribute this failure to distribution mismatch and signal conflict:

1) **Inference Ambiguity:** During actual inference on a 3DGS render, we possess neither perfect ground truth depth nor a helpful zero-map; we have a noisy rendered depth. If we feed the zero-map, the model treats the input as "unknown geometry" but struggles because the zero-signal is technically a strong spatial feature (indicating empty space) that conflicts with the non-empty RGB content.
2) **Modality Switching:** The network likely struggled to learn a unified latent representation that accommodates both highly detailed structural guidance and a complete absence of it. The "zero-depth" tokens essentially acted as noise, consuming model capacity without providing informative gradients, ultimately degrading the base restoration capability.

5 Discussion: The Artifact Propagation Phenomenon

The central and most significant finding of our project is the identification and empirical characterization of **Artifact Propagation**. This phenomenon explains why a supposedly beneficial geometric prior actively damages the repair process in the neural rendering domain.

5.1 Conflict of Priors: RGB vs. Geometric

Standard generative models succeed because their inherent image prior is strong: they know what a natural image should look like.

- **RGB Prior's Strength:** A floater (a small, dense blob of noise floating in the sky) violates the natural image manifold. The RGB-only model correctly identifies this region as abnormal and relies on its learned prior to "denoise" it, typically by setting the region to the background color (sky, wall, etc.).

TABLE 4
Accumulation Results Across Different Training Epochs on NeRF Blender Materials

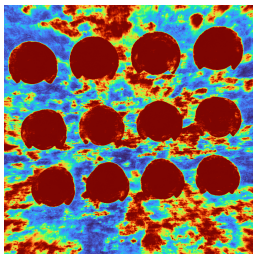
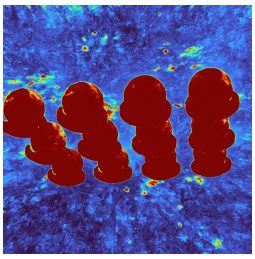
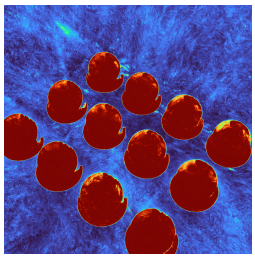
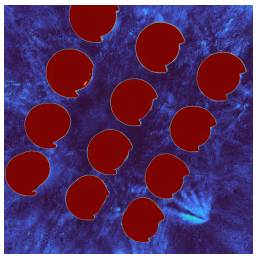
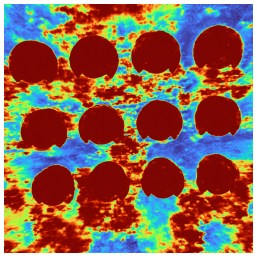
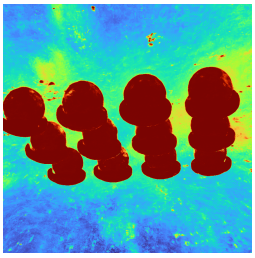
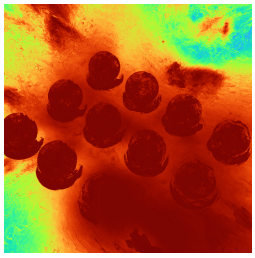
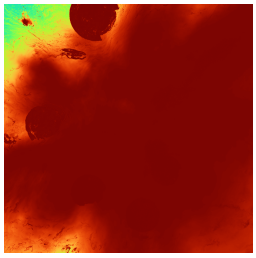
Model	2000 Epochs	10000 Epochs	20000 Epochs	30000 Epochs
Base				
Depth + Canny				

TABLE 5
Training Progression on ScanNet++ with Zero-Depth Conditioning Strategy

Target (GT)	Epoch 0 (PSNR 22.87)	Epoch 500 (PSNR 27.20)
		

- **Geometric Prior's Conflict:** When geometric conditioning is introduced, the floater now possesses a 3D location. It appears in the rendered depth map D_{render} as a sharp discontinuity at a specific depth z_{floater}. The T2I-Adapter extracts this spatial feature and relays a strong, explicit command to the U-Net that there is a distinct object at distance z_{floater}.

The core conflict is between the two loss mechanisms: the image generation loss wants to output a clean sky (deleting the floater), but the adapter injection implicitly enforces the geometric loss, forcing the model to hallucinate texture onto the geometrically defined floater region. This results in the model creating a clearly textured artifact (solidifying the floater), a visually worse outcome than the baseline's soft blending.

5.2 Mechanism of Degradation

The mechanism can be visualized as a cycle:

$$\text{3DGS Error} \xrightarrow{\text{Render}} \text{Artifact RGB} \wedge \text{Artifact Depth} \xrightarrow{\text{Adapter}} \text{Forced Texture Generation}$$

The T2I-Adapter effectively acts as a constraint enforcer that ensures the generated texture is consistent with the depth map. By passing the error in the depth domain, the adapter forces the generation process to adhere to the erroneous geometry, propagating the 3D bug into the 2D output. This

effect is visible in Table 4, where the depth + Canny model diverges from the true representation while the baseline maintains it. This discrepancy between slightly improved per-view 2D metrics and worse end-to-end 3D performance highlights a fundamental limitation of purely image-based evaluation: a model can increase PSNR/SSIM on individual frames while simultaneously degrading multi-view consistency. In our case, the adapter's strict adherence to erroneous depth signals reinforces incorrect geometry across views, leading to poorer accumulated 3D structure despite nominal improvements in 2D quality.

5.3 Impact of Canny Conditioning

The poor performance of Canny conditioning is related but distinct. Canny edge detection on blurred or noisy images is highly unstable, and minor color fluctuations can produce sharp, spurious edges. The Canny adapter then interprets these spurious lines as definitive structural boundaries. When the underlying ground truth is complex, this low-quality edge input guides the restoration away from the true contours, severely impacting high-frequency detail and leading to a significant drop in PSNR and SSIM.

5.4 Implications for Future 3D Restoration

This investigation serves as a critical negative result: simply gaining access to geometric data is insufficient for improving neural rendering repair. Future work must focus on methods to validate or filter the geometric prior before injection. If the primary artifact in neural rendering is geometric, the corrective signal must be free of that geometric noise, or the model must be trained to recognize and ignore it.

6 Conclusion and Future Work

We presented a comprehensive architectural investigation into integrating explicit geometric priors (depth and Canny edges) into the single-step Difix restoration pipeline via T2I-Adapters. Our goal was to test whether injecting structural cues could improve artifact correction in neural rendering.

Empirically, our results on controlled synthetic and real-world datasets demonstrate a pattern: geometric conditioning provides modest per-view improvements in 2D quality metrics (PSNR, LPIPS, SSIM) when the geometry is clean, but fails to translate into better end-to-end 3D reconstruction performance. We identified **Artifact Propagation** as the core failure mechanism: geometric artifacts inherent to the flawed 3D representation contaminate the conditioning channel, forcing the generative repair module to solidify, rather than eliminate, the errors.

This discrepancy between slightly improved 2D metrics and worse downstream 3D reconstruction performance highlights a limitation of solely image-based evaluation: a restoration model can increase the per-view PSNR/SSIM while reinforcing the very geometric errors that 3D Gaussian Splatting must optimize against, deteriorating multi-view consistency and downstream geometry quality. This study validates the strength of the RGB-only generative prior in detecting visual anomalies and highlights the crucial necessity of decoupling appearance (texture) and geometry (depth) in 3D-aware restoration systems.

6.1 Limitations

Our current implementation uses a simple sum for multi-modal fusion, which is sensitive to noise in any single channel. Furthermore, the reliance on rendering depth directly from the flawed 3D model means the conditioning signal inherently carries the primary bug.

6.2 Future Work

Based on the strong empirical evidence of Artifact Propagation, we propose several avenues for robust, geometry-aware repair:

1) **Confidence-Weighted Fusion:** Instead of simple addition, implement a mechanism where the geometric prior is modulated by a confidence map M. This map could be derived from depth uncertainty (e.g., standard deviation of accumulated transmittance in NeRF) or a separate learned error predictor. The fusion becomes:
$$F^i_{\text{U-Net}} = F^i_{\text{U-Net}} + M_i \odot (\mathcal{F}_{\text{geo}})^i(C_{geo})$$
The model would be trained to ignore geometric cues in regions of high uncertainty.
2) **Independent Structural Priors:** Use an external, highly robust monocular depth model (like Depth Anything V2) exclusively for geometric conditioning. If the monocular model successfully ignores 3DGS floaters, its output D_{Mono} might provide a cleaner signal, breaking the negative feedback loop.
3) **Perceptual Geometry Loss:** Instead of fully relying on the T2I-Adapter's implicit geometry adherence, introduce an explicit geometric loss term that compares the predicted depth map (if generated) against the expected clean depth, forcing the model to learn a clean geometry representation alongside the texture.

Acknowledgments

We thank the teaching staff of CSC2529 for their dedication and guidance throughout the project.

References

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *European Conference on Computer Vision (ECCV)*, 2020.

[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, 2023.

[3] D. Charatan, S. Li, A. Tagliasacchi, and V. Sitzmann, "pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction," in *CVPR*, 2024.

[4] Y. Chen, H. Xu, C. Zheng, B. Zhuang, M. Pollefeys, A. Geiger, T.-J. Cham, and J. Cai, "Mvsplat: Efficient 3d gaussian splatting from sparse multi-view images," *arXiv preprint arXiv:2403.14627*, 2024.

[5] J. Z. Wu, Y. Zhang, H. Turki, X. Ren, J. Gao, M. Z. Shou, S. Fidler, Z. Gojcic, and H. Ling, "Difix3d+: Improving 3d reconstructions with single-step diffusion models," 2025. [Online]. Available: https://arxiv.org/abs/2503.01774

[6] C. Mou, X. Wang, L. Xie, Y. Wu, J. Zhang, Z. Qi, Y. Shan, and X. Qie, "T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models," *arXiv preprint arXiv:2302.08453*, 2023.

[7] 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 (ICCV)*, 2023.

[8] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[9] T. Anciukevičius, Z. Xu, M. Fisher, P. Henderson, H. Bilen, N. J. Mitra, and P. Guerrero, "Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation," 2024. [Online]. Available: https://arxiv.org/abs/2211.09869

[10] R. Lu, Y. Chen, J. Ni, B. Jia, Y. Liu, D. Wan, G. Zeng, and S. Huang, "Movis: Enhancing multi-object novel view synthesis for indoor scenes," 2025. [Online]. Available: https://arxiv.org/abs/2412.11457

[11] X. Zhang, Y. Zhang, L. Mehl, M. Gross, and C. Schroers, "High-fidelity novel view synthesis via splatting-guided diffusion," 2025.

[12] F. Wang, A. Louys, N. Piasco, M. Bennehar, L. Roldão, and D. Tsishkou, "Planerf: Svd unsupervised 3d plane regularization for nerf large-scale scene reconstruction," 2023.

[13] L. Yang, B. Kang, Z. Huang, Z. Zhao, X. Xu, J. Feng, and H. Zhao, "Depth anything v2," *arXiv preprint arXiv:2406.09414*, 2024.

[14] A. Sauer, D. Lorenz, A. Blattmann, and R. Rombach, "Adversarial diffusion distillation," 2023. [Online]. Available: https://arxiv.org/abs/2311.17042

[15] J. Canny, "A computational approach to edge detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-8, no. 6, pp. 679–698, 1986.

[16] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," 2018. [Online]. Available: https://arxiv.org/abs/1801.03924

[17] L. A. Gatys, A. S. Ecker, and M. Bethge, "A neural algorithm of artistic style," 2015. [Online]. Available: https://arxiv.org/abs/1508.06576

[18] L. Ling, Y. Sheng, Z. Tu, W. Zhao, C. Xin, K. Wan, L. Yu, Q. Guo, Z. Yu, Y. Lu, X. Li, X. Sun, R. Ashok, A. Mukherjee, H. Kang, X. Kong, G. Hua, T. Zhang, B. Benes, and A. Bera, "Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision," 2023. [Online]. Available: https://arxiv.org/abs/2312.16256

[19] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai, "Scannet++: A high-fidelity dataset of 3d indoor scenes," 2023. [Online]. Available: https://arxiv.org/abs/2308.11417