École Polytechnique Fédérale de Lausanne

Large Scale 3D Scene Relighting using Pre-Trained Diffusion Models

by Efe Tarhan
MSc. Student in Communication Systems

Research Project Report

Dongqing Wang
Project Supervisor

EPFL IC IINFCOM IVRL
BC 328 (Bâtiment BC)
Station 14
1015 Lausanne

June 12, 2025

Abstract

In this project, an enhanced neural radiance field (NeRF) editing framework is proposed. The task is addressed by leveraging the strong priors of pretrained diffusion models, building upon the Delta Denoising Score (DDS) pipeline to enable color and texture modifications while preserving reflections within the scene. Wavelet transforms are employed as a selective filtering mechanism to pass only low-frequency components, allowing for the preservation of reflection cues during the editing process. A NeRF model is initially trained on a large-scale open-area scene to obtain a high-fidelity pretrained representation. Subsequent edits are performed via DDS-guided optimization, with the scene geometry kept frozen to maintain structural consistency. To further refine the editing process, surface normal prediction is incorporated, enhancing the geometric priors and promoting more accurate view-dependent appearance. The proposed framework demonstrates qualitatively superior results compared to previous methods, exhibiting improved visual fidelity, localized edits and reflection quality.

Contents

Chapter 1

Introduction

Neural Radiance Fields (NeRF) have become a foundational tool in 3D scene representation, offering photorealistic novel view synthesis from sparse multi-view imagery [9]. Despite their success, editing NeRFs remains a challenging problem due to the entangled representation of geometry and appearance. This limitation makes it a nontrivial task to perform semantic manipulations such as relighting, texture change, or object recoloring while preserving structural integrity and view consistency.

Recent work has explored the integration of pretrained diffusion models into the NeRF editing pipeline to address this challenge. Diffusion models, known for their semantic priors and high generative fidelity, provide supervision through score-based feedback mechanisms. Notably, Score Distillation Sampling (SDS) has been used for generating 3D NeRF representations with respect to provided text or image prompts [11]. However, SDS-based methods tend to apply coarse global changes and often compromise geometric and photometric consistency. Delta Denoising Score (DDS) [2] improves upon SDS by computing the difference in denoising scores between source and target prompts and backpropagating this gradient to the appearance parameters of a pretrained NeRF. This enables more localized and controlled edits without the need to regenerate training data. DreamCatalyst [4] further enhances this framework by introducing a latent-based regularization term that matches latent fidelity between source and target latens. It also proposes a dynamic loss weighting schedule to balance semantic alignment and identity preservation across editing iterations.

While DreamCatalyst achieves improved fidelity and consistency, it still faces limitations in handling fine-grained details and preserving global illumination effects such as reflections. To address these shortcomings, a wavelet-guided NeRF editing framework is proposed which extends DreamCatalyst through the integration of wavelet-based gradient filtering and surface normal supervision.

Wavelet transforms decompose gradients into frequency bands, enabling selective filtering of editing signals. By attenuating high-frequency components and preserving low-frequency bands, our method maintains global structures such as reflections and details while allowing localized semantic changes. Additionally, a surface normal prediction module is incorporated during training to guide the appearance head with geometric priors, further improving view-dependent rendering.

The remainder of this report is structured as follows: Chapter 2 reviews prior work in NeRFs, diffusion-guided generation, and score-based 3D editing. Chapter 3 details our proposed methodology, including DDS supervision, wavelet filtering and normal prediction. Chapter 4 presents our experimental setup and qualitative results. Chapter 5 explains the limitations faced during the project and possible future work to mitigate these problems. Chapter 6 concludes with a summary of findings and the project.

Chapter 2

Related Work

2.1 Neural Radiance Fields and Scene Representation

Neural Radiance Fields (NeRF), whose reconstruction pipeline is illustrated in Figure 2.1, have introduced a powerful framework for representing 3D scenes as continuous volumetric fields optimized from multi-view images, enabling high-quality novel view synthesis [9]. This implicit representation has since been extended and accelerated through various advancements, including point-based alternatives such as 3D Gaussian Splatting [3], which achieves real-time rendering by optimizing Gaussian primitives with learnable parameters for mean, variance, density, and spherical harmonics (SH) coefficients. These foundational techniques have laid the groundwork for robust 3D scene representations, initially for bounded objects and later extended to large-scale unbounded environments. In this work, I investigate how such representations can be adapted for generative modeling and editing, where recent pipelines aim to balance realism, structural consistency, and semantic controllability.

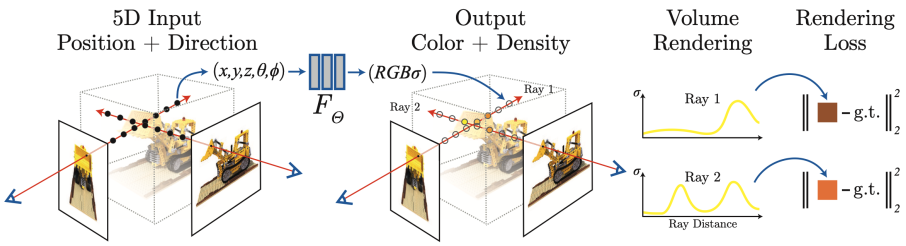

Figure 2.1: The NeRF-based scene reconstruction pipeline. Adapted from [9].

Recent developments have further explored alternative 3D scene representations that better support interactive applications, relighting, and dynamic content editing. Our approach builds directly on these NeRF-based models, adapting them into an editing pipeline that preserves the

underlying scene geometry, with a particular emphasis on maintaining view-dependent reflections, while enabling targeted modifications to the appearance of selected objects.

2.2 Conditional and Controlled Diffusion Guidance

Diffusion models have become a central component in generative modeling due to their stability, scalability, and ability to produce diverse outputs. Foundational works like Stable Diffusion [12] have paved the way for controlling generation through text prompts, steering outputs by manipulating representations in the latent space. The integration of conditional control into diffusion pipelines has significantly enhanced both editability and fidelity, as demonstrated by ControlNet [15], which enables spatial and semantic conditioning during the generative process.

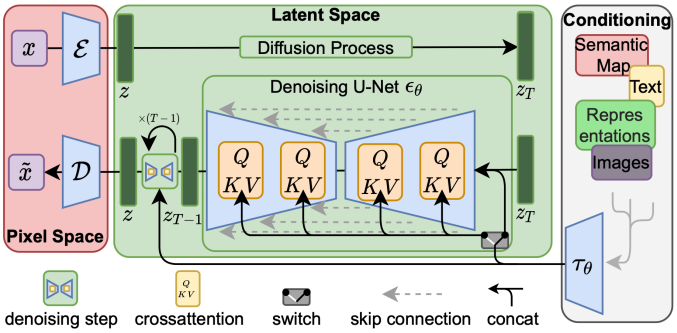

Figure 2.2: The Stable Diffusion pipeline for text conditioned image generation [12].

In the context of 3D generation and editing, methods such as Zero-1-to-3 [7] have demonstrated that strong priors from diffusion models can be effectively transferred to 3D tasks in a zero-shot setting, where models trained exclusively on synthetic 3D data are capable of producing plausible novel views from a single RGB image. Likewise, Magic3D [6] introduces a cascaded diffusion approach to synthesize high-resolution 3D content from text, emphasizing both geometric fidelity and appearance quality. While these approaches primarily target generation, their architectural components and optimization strategies are increasingly being repurposed for 3D editing, where the goal is to adapt existing scenes in a controlled and semantically guided manner.

Our approach builds on this developed research by incorporating conditional denoising supervision, particularly text conditioning, into the editing of pretrained NeRF scenes. By leveraging strong diffusion priors, precise modifications to color and texture while preserving critical structural elements such as geometry, fine details, and view-dependent reflections are enabled.

2.3 Text-Guided and Score-Based 3D Editing

Editing 3D content using text inputs has attracted significant attention in recent literature. DreamFusion [11] and ProlificDreamer [14] proposed pipelines for text-to-3D generation by leveraging pretrained 2D diffusion models, utilizing score distillation methods such as Score Distillation Sampling (SDS) and Variational Score Distillation (VSD), respectively. These approaches achieve results in terms of semantic alignment and visual fidelity by backpropagating gradients from the diffusion model to update the parameters of a NeRF-based 3D representation. Although they are well-suited for generative tasks, these methods are not ideal for editing existing scenes. Since they are tuned for generative tasks, the edits are often leads to low-quality, overly smoothed, and spatially imprecise edits. Fine details and local structural elements are frequently degraded, making them less effective for applications that require precise and localized modifications.

Instruct-NeRF2NeRF [1] introduces a diffusion-based 3D editing framework that modifies NeRF scenes by updating the training dataset rather than directly optimizing NeRF parameters. While this enables high-level instruction-driven edits, it suffers from spatial inconsistency, as iteratively replacing and re-optimizing data makes it difficult to maintain coherence across views. Delta Denoising Score (DDS) [2] addresses this by introducing a direct optimization framework that leverages pretrained diffusion models to compute denoising score differences between source and target prompts. These gradients are backpropagated to update NeRF's appearance parameters, enabling localized, semantically guided edits without regenerating the dataset or retraining the model. The DDS-based update scheme is illustrated in Figure 2.3.

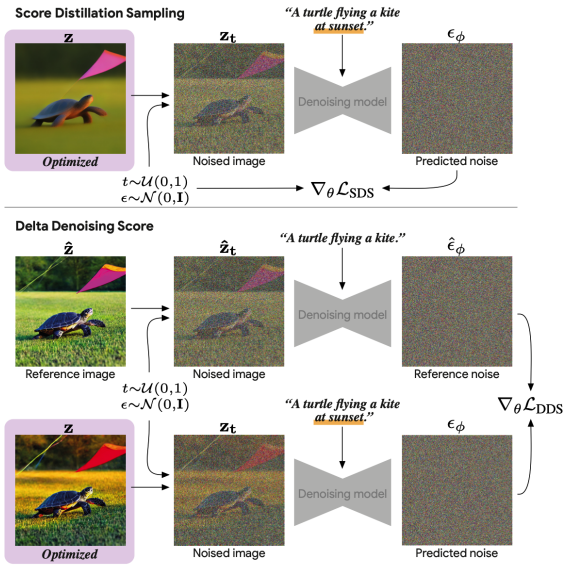

Figure 2.3: Pipeline comparison between SDS and DDS-based editing [2].

By directly supervising the NeRF through diffusion-based gradients, DDS offers finer appearance

control while preserving structural consistency. It bridges the gap between the global, coarse edits of DreamFusion and the dataset-based updates of Instruct-NeRF2NeRF. DreamCatalyst [4], built upon the DDS framework, further improves editing fidelity by introducing a latent regularization between the rendered scene and the original images. It also employs iteration-dependent loss weighting, focusing on semantic edits during early stages and gradually shifting toward identity preservation in later iterations.

Additional works such as Contrastive Denoising Score (CDS) [10] and Posterior Distillation Sampling (PDS) [5] extend the DDS framework by introducing improved loss formulations and sampling strategies. CDS enforces stronger contrastive supervision by comparing positive and negative prompt pairs, improving semantic separation and edit sharpness. PDS, on the other hand, refines the sampling process by distilling the posterior distribution from diffusion trajectories, enabling more stable and expressive guidance during optimization. These methods highlight the growing refinement of score-based NeRF editing pipelines, moving toward more controllable and photorealistic scene editing.

In this work, the DreamCatalyst methodology is extended by incorporating wavelet-based filtering into the gradient updates to enhance the preservation of fine details during editing. Furthermore, the effect of freezing geometry layers and enabling surface normal prediction is tested during the experiments. The CDS methodology has been adapted to 3D editing scenario as an extension method. These components are explained in detail in the following chapter.

Chapter 3

Methodology

In this chapter, the proposed NeRF editing pipeline is detailed. The method builds upon the Delta Denoising Score (DDS) framework and DreamCatalyst architecture, and is extended through the introduction of wavelet-based gradient filtering, surface normal prediction, and the contrastive score-based loss described in the CDS. These components collectively enable semantically meaningful, localized edits while preserving geometric integrity and view-dependent details such as reflections.

3.1 Delta Denoising Score and DreamCatalyst

The Delta Denoising Score (DDS) [2] introduces a score-based NeRF editing method that leverages pretrained diffusion models to compute the gradient between the denoising scores of a source prompt y_s and a target prompt y_t. Given an input latent z and timestep t, DDS minimizes the squared norm of the difference in predicted noise:

$$\nabla_\theta \mathscr{L}_{\text{DDS}} = \nabla_\theta \mathbb{E}_{t,\epsilon} \left[\left\| \epsilon_\theta(\mathbf{x}_t^{\text{src}}, \mathbf{y}_t^{\text{src}}, t) - \hat{\epsilon}_\theta(\mathbf{x}_t^{\text{tgt}}, \mathbf{y}_t^{\text{tgt}}, t) \right\|_2^2 \right] \tag{3.1}$$

This loss is backpropagated to update the appearance-related parameters of a pretrained NeRF model while keeping the geometry fixed.

DreamCatalyst [4] improves this framework by introducing an identity preservation mechanism that ensures the edited scene maintains fidelity to the original content. To enforce edit stability and visual consistency, it incorporates an additional latent reconstruction loss:

$$\mathscr{L}_{\text{latent}} = \left\| \mathbf{z}_t^{\text{src}} - \mathbf{z}_t^{\text{tgt}} \right\|^2, \tag{3.2}$$

where $\mathbf{z}_t^{\text{tgt}}$ is the diffusion-model latent of the synthesized image from the NeRF and $\mathbf{z}_t^{\text{src}}$ is the corresponding latent representation of the ground truth image. This ensures consistency in the semantic and perceptual space, rather than the pixel domain.

Moreover, DreamCatalyst proposes a dynamic loss weighting strategy where the total loss combines the DDS and latent terms:

$$\mathscr{L}_{\text{total}} = \lambda_{\text{DDS}}(i) \cdot \mathscr{L}_{\text{DDS}} + \lambda_{\text{latent}}(i) \cdot \mathscr{L}_{\text{latent}}, \tag{3.3}$$

with iteration-dependent weights $\lambda_{\text{DDS}}(i)$ and $\lambda_{\text{latent}}(i)$ that gradually shift focus from semantic editability to representation fidelity across iterations i. In practice, λ_{DDS} decays and λ_{latent} increases over time to prioritize latent preservation in the later optimization stages.

These foundational components form the base of our editing pipeline, upon which further enhancements, such as wavelet-domain gradient control and surface normal supervision, are introduced to preserve fine details and structural consistency.

3.2 Wavelet Transform

Wavelet transforms provide a powerful framework for analyzing signals and images at multiple spatial and frequency scales. Unlike the Fourier transform, which offers only frequency information, wavelets offer both frequency and spatial localization, making them particularly suitable for tasks such as compression, denoising, and in our case, gradient filtering for reflection-preserving NeRF editing.

Types of Wavelets and Wavelet Order

There exist several wavelet families, each offering unique trade-offs between spatial resolution, frequency localization, and symmetry. The most commonly used families include:

- **Haar**: A simple, piecewise constant wavelet with excellent computational efficiency but poor frequency resolution.
- **Daubechies (dbN)**: A widely used family of orthogonal wavelets characterized by smoothness and compact support. The order N determines the number of vanishing moments and the

wavelet's ability to represent smooth transitions.

- **Symlets and Coiflets**: Variants of Daubechies designed to offer improved symmetry and reconstruction quality, especially useful in image processing tasks.

Higher-order wavelets, such as db8, offer increased smoothness and are more effective in preserving global features like soft lighting or reflections, which are critical in NeRF rendering.

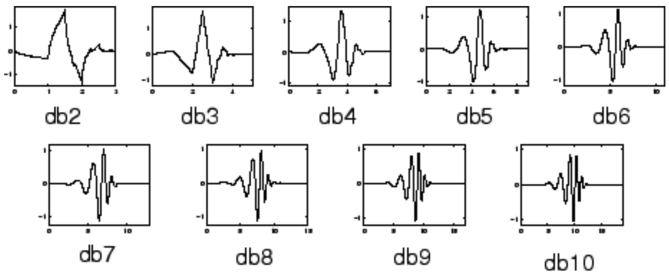

Figure 3.1: Example wavelet basis functions for different families and orders [8].

Decomposition Level J

In the discrete wavelet transform (DWT), an image is recursively decomposed into low-frequency (approximation) and high-frequency (detail) components at each level. These include:

- **LL**: Approximation coefficients (low-low), capturing the global structure.
- **LH, HL, HH**: Detail coefficients capturing horizontal, vertical, and diagonal edges, respectively.

The number of decomposition levels J defines how many times this process is applied. A higher J results in more fine-grained separation of frequency bands. In our setup, typical values of $J = 1, 2,$ or 3 are used. Larger values allow better suppression of high-frequency editing artifacts, thereby preserving global scene features and preventing over-modification of detailed structures like reflections.

3.3 Wavelet-based Gradient Filtering

In conventional DDS-based editing, gradients from the denoising score difference are computed in the latent space and directly backpropagated to the appearance features of the NeRF. However, such raw updates often lead to changes in high-detail features such as texture and reflections, especially when the prompt is under-specified.

To mitigate this, a wavelet-based gradient filtering approach is introduced into the DDS pipeline. Specifically, the gradient of the DDS loss is transformed using a Discrete Wavelet Transform (DWT), decomposing it into subbands across multiple frequency levels. Let $\nabla_\phi \mathscr{L}_{\text{DDS}}$ denote the original gradient in the latent space. The transformed gradient becomes:

$$\{g_{\text{LL}}, g_{\text{LH}}, g_{\text{HL}}, g_{\text{HH}}\} = \text{DWT}(\nabla_\phi \mathscr{L}_{\text{DDS}}) \tag{3.4}$$

A filtering mask is then applied to suppress or attenuate specific frequency bands. In our case, high-frequency subbands (LH, HL, HH) are suppressed to retain global structures such as lighting and reflections, while the low-frequency component (LL) is preserved:

$$g'_{\text{LL}} = g_{\text{LL}}, \quad g'_{\text{LH}} = \alpha \cdot g_{\text{LH}}, \quad g'_{\text{HL}} = \alpha \cdot g_{\text{HL}}, \quad g'_{\text{HH}} = \alpha \cdot g_{\text{HH}} \tag{3.5}$$

where $\alpha = 0$ as the suppression coefficient or perfect low pass filter (LPF).

The filtered gradient is then reconstructed using the inverse DWT (IDWT):

$$\nabla_\phi \tilde{\mathscr{L}}_{\text{DDS}} = \text{IDWT}(g'_{\text{LL}}, g'_{\text{LH}}, g'_{\text{HL}}, g'_{\text{HH}}) = \text{IDWT}(g_{\text{LL}}, 0, 0, 0) \tag{3.6}$$

This filtered gradient $\nabla_\phi \tilde{\mathscr{L}}_{\text{DDS}}$ replaces the original DDS gradient in the optimization step. In contrast, the latent loss is left unfiltered and directly backpropagated:

$$\nabla_\phi \mathscr{L}_{\text{latent}} = 2\left(\mathbf{z}_t^{\text{src}} - \mathbf{z}_t^{\text{tgt}}\right) \tag{3.7}$$

Thus, the total optimization objective becomes:

$$\phi^* = \arg\min_\phi \lambda_{\text{DDS}}(t) \cdot \tilde{\mathscr{L}}_{\text{DDS}} + \lambda_{\text{latent}}(t) \cdot \mathscr{L}_{\text{latent}} \tag{3.8}$$

This targeted filtering preserves view-consistent reflective structures and avoids distortions introduced by uncontrolled score gradients, resulting in cleaner and more photorealistic edits.

3.4 Surface Normal Prediction and Geometry Freezing

To further enhance structural consistency and reinforce the geometric prior of the pretrained NeRF scene, a surface normal prediction module from the Nerfstudio framework is incorporated. This

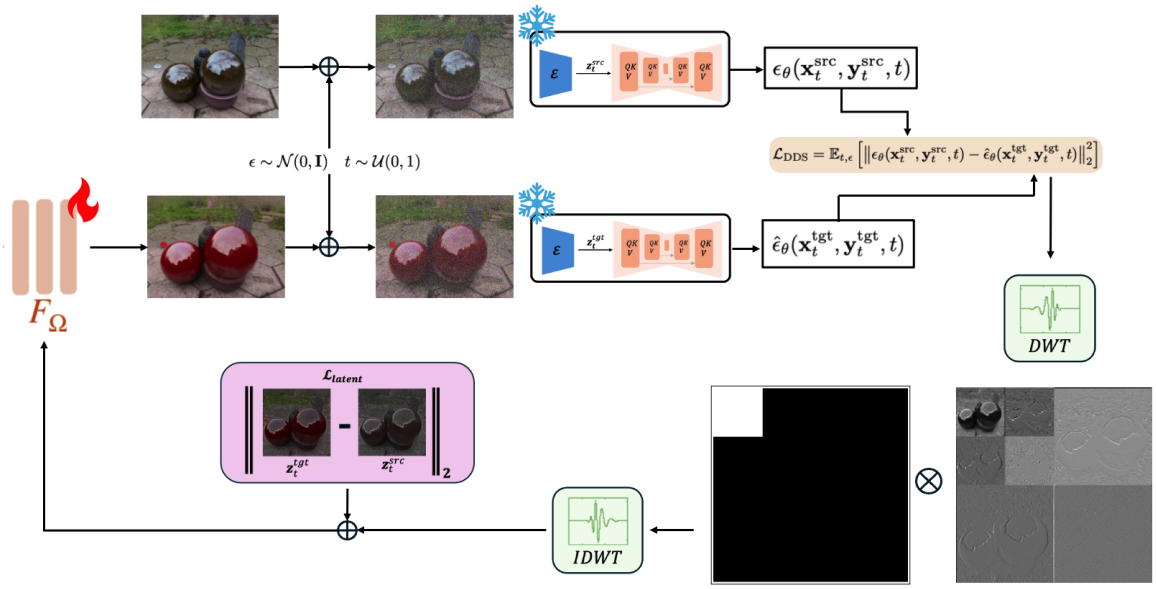

Figure 3.2: The Wavelet-integrated DreamCatalyst pipeline.

module introduces an auxiliary supervision branch that learns to predict surface normals as an additional modality. By encouraging accurate surface orientation estimates, the predicted normals improve the quality of the underlying geometry and lead to more consistent view-dependent effects. This not only benefits the fidelity of the original NeRF but also contributes to higher-quality and structurally coherent edits during the optimization process.

Additionally, for extra trials and experiments, the NeRF geometry layers are frozen throughout the editing process. Only the appearance-related parameters ϕ are updated, ensuring that the spatial structure of the scene remains unchanged.

3.5 Contrastive Denoising Score (CDS) Integration

To further improve the alignment between source and target prompts, the Contrastive Denoising Score (CDS) [10] pipeline is also explored and integrated. As shown in Figure 3.3, CDS extends the DDS framework by incorporating a contrastive loss on the attention maps of the denoising UNet. For each denoising step, positive and negative pairs of attention features are sampled, and trained using a cross-entropy based PatchNCE Loss, encouraging alignment for latent patches from the same regions and discouraging alignment from different regions to localize the edits.

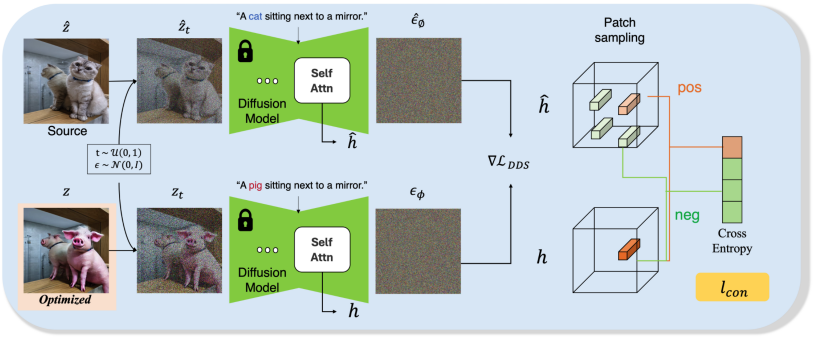

Figure 3.3: The CDS pipeline for 2D semantic editing, adapted from [10]. Attention maps are regularized using contrastive loss across source and target prompts.

While DDS directly penalizes noise prediction discrepancies, CDS regularizes the editing procedure by limiting the regions where the latens are spread. This improves semantic alignment and localizes optimization. In our setup, the 2D CDS method has been implemented for NeRF editing pipeline of Nerfstudio where similar to Dreamatalyst, the obtained loss w.r.t the latent variable provided to the 2D diffusion module is backpropagated to the parameters of the NeRF model.

Chapter 4

Results

This chapter presents the experimental results evaluating the proposed diffusion-guided NeRF editing framework. The experiments focus on three main aspects: (i) the impact of different wavelet types used in the gradient filtering process, (ii) the influence of optimization settings such as iteration count and guidance scale, (iii) qualitative comparison across diverse scenes to assess the consistency, fidelity, and detail level of the edits, and lastly (iv) implementation of CDS method. All experiments are performed on pretrained real NeRF scenes from unbounded environments.

4.1 Wavelet Type Ablation

To evaluate the role of wavelet type in the filtering process, several wavelet families (e.g., Haar, Daubechies) applied to the DDS gradients are being used for the experiments. Figures 4.1, 4.2, and 4.3 show editing outcomes on three scenes—*gardenspheres*, *toycar*, and *sedan*—with consistent source and target prompts.

In the gardenspheres scene (Figure 4.1), which involves recoloring reflective spheres from gray to red, the best results are achieved using third-order Daubechies (`db4`) wavelets. These preserve reflection structures more faithfully than Haar or `db2` wavelets, which tend to distort specular highlights. When using lower-order wavelets or filters that suppress higher-frequency bands, artifacts such as residual sky color (e.g., persistent blue regions) become apparent. In contrast, order-3 wavelets mitigate this issue. Notably, all wavelet-based methods successfully preserve reflections, whereas DreamCatalyst fails to retain reflective details.

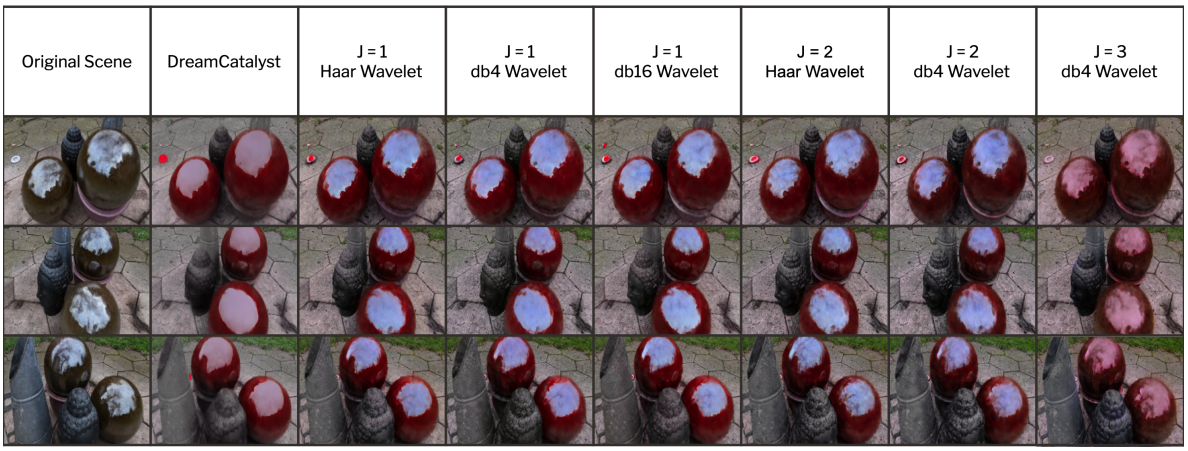

Figure 4.1: Wavelet ablation on the "gardenspheres" scene. Best result achieved using 3rd-order db4 wavelets. Source: "a photo of two reflective spheres." Target: "a photo of two red reflective spheres."

In the toycar scene (Figure 4.2), which involves recoloring the car to orange, both Haar and db4 wavelets yield compelling results. However, increasing the wavelet order leads to a gradual reversion of the edits, diminishing the intended color transformation. It is also important to note that each column in the table corresponds to a separate experiment with different learning rate and text guidance settings. Compared to these wavelet-based approaches, the original DreamCatalyst output suffers from reduced contrast, desaturated colors, and loss of fine details, such as the car's reflection in the mirror, which are better preserved when wavelets are used.

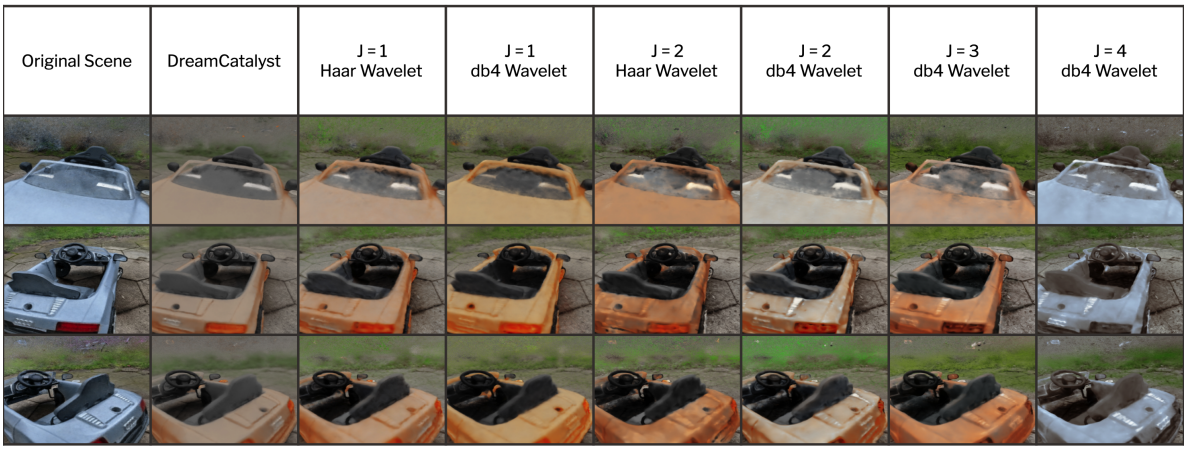

Figure 4.2: Wavelet ablation on the "toycar" scene. Best results observed with Haar or db4 wavelets. Source: "a photo of a toy car." Target: "a photo of an orange toy car."

In the sedan scene (Figure 4.3), the most consistent and effective color transformation from black to red is achieved using second-order Haar wavelets. While higher-order and multi-level wavelets enhance reflection detail, they also begin to degrade the target colors in certain regions, revealing a trade-off between preserving fine structural detail and maintaining accurate color edits.

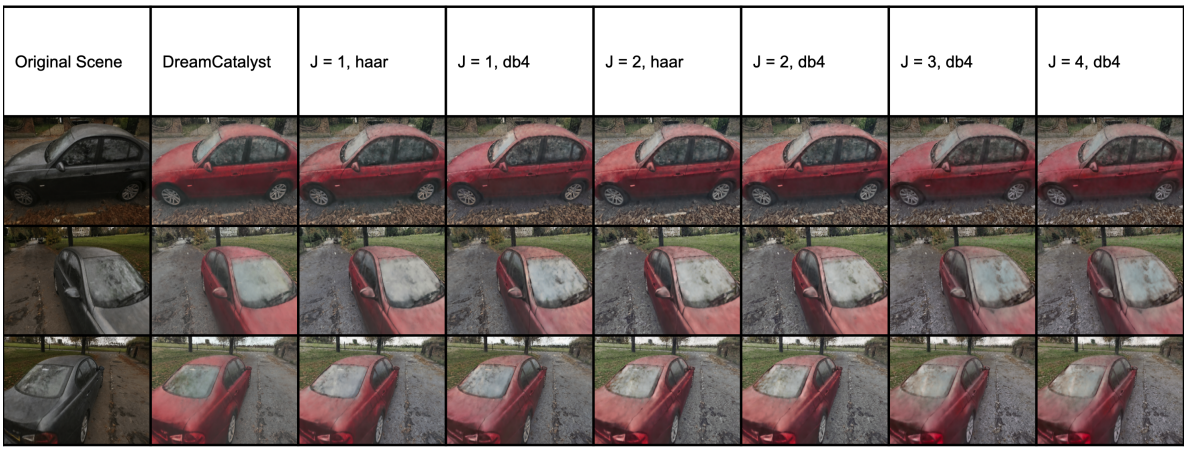

Figure 4.3: Wavelet ablation on the "sedan" scene. Best results obtained with 2nd-order Haar wavelets. Source: "a photo of a black car." Target: "a photo of a red car."

4.2 Effect of Optimization Settings

To assess the influence of optimization parameters, the number of iterations and the guidance scale S_g used in DDS are examined.

Figure 4.4 shows the progression of edits at increasing iterations. While initial stages yield meaningful semantic changes, over-optimization sometimes causes semantic reversion or blurring, particularly in the *toycar* case, likely due to overly strong identity constraints. This observation motivates spatially adaptive regularization as discussed in Chapter 5.

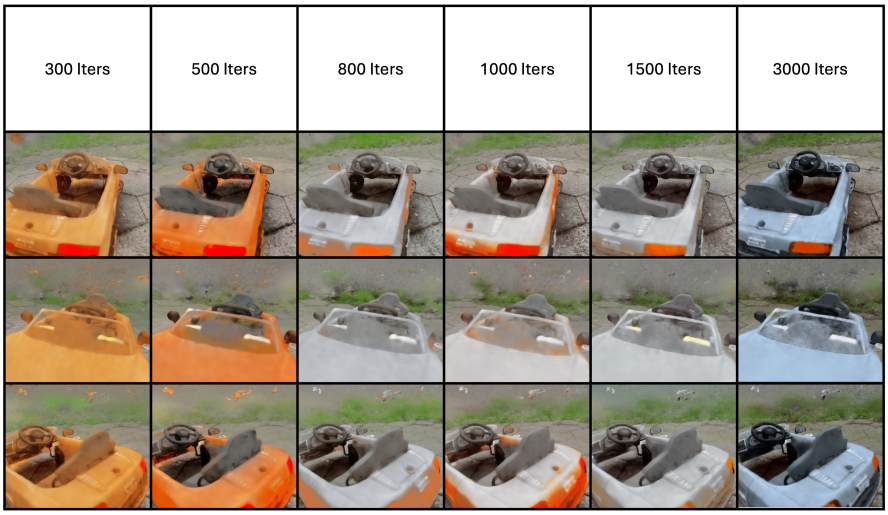

Figure 4.4: Effect of iteration count on edit quality in the "toycar" scene. Later iterations can introduce blur or revert edits.

Figure 4.5 illustrates the effect of varying guidance scale S_{Guidance}. Low values result in weak edits, while high values introduce saturation artifacts. Moderate scales strike a balance between photorealism and semantic accuracy.

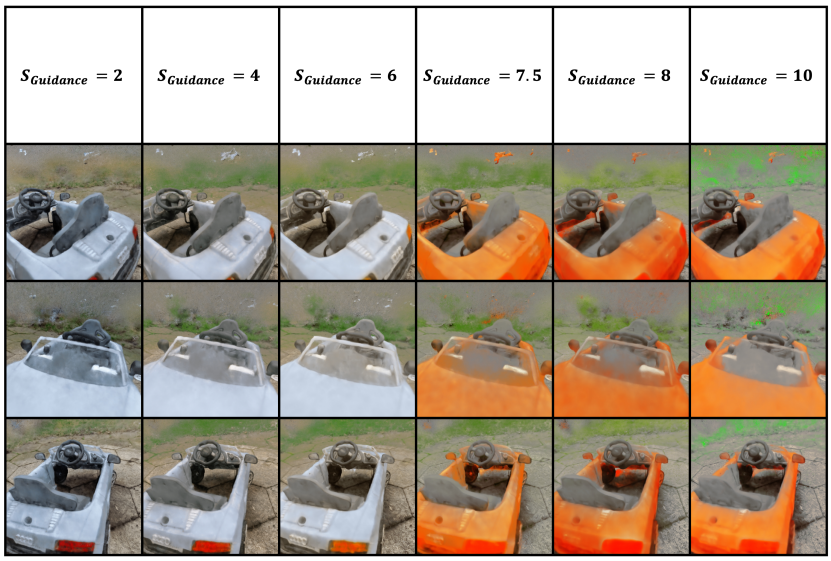

Figure 4.5: Effect of guidance scale S_g on the "toycar" scene. Larger values introduce artifacts; optimal edits occur at moderate scales.

4.3 Alternative Geometry Representations

The effectiveness of DreamCatalyst is also examined on scenes rendered using Gaussian Splatting. Figure 4.6 demonstrates that low-fidelity geometry in these representations impairs editing quality and introduces significant visual artifacts. These findings indicate the need for improved regularization or smoothing for splat-based representations in DDS pipelines.

Figure 4.6: Editing result on a Gaussian Splatting-based scene. Poor geometry fidelity leads to severe artifacts and degraded edits.

4.4 Normal Prediction Analysis

To enhance editability through geometric priors, surface normal prediction was introduced. Figure 4.7 and Figure 4.8 compare outputs with and without predicted normals (top row: with normals, bottom: without). Despite expectations, normal prediction did not lead to observable improvements. While it aimed to regularize scene geometry in the final editing stages, its impact remained minimal in practice.

Figure 4.7: Effect of normal prediction on a Gaussian Splatting scene. Top: with normals, Bottom: without. No substantial improvement is observed.

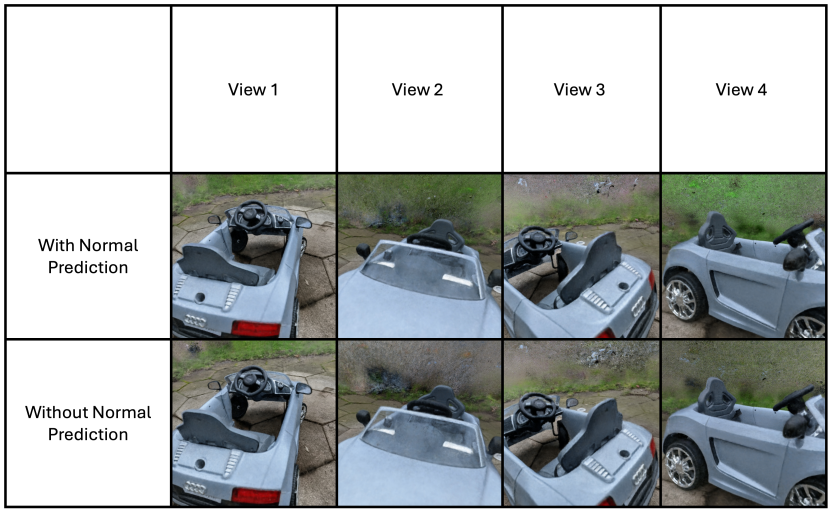

Figure 4.8: Effect of normal prediction on the "toycar" scene. Top: with normals, Bottom: without. No significant visual gain is observed.

4.5 Contrastive Denoising Score (CDS) Evaluation

To evaluate the performance of the CDS pipeline [10] in semantic NeRF editing, we apply it to the *gardenspheres* scene with a range of color-based prompts: "red reflective spheres," "green reflective spheres," "blue reflective spheres," and "yellow reflective spheres." The qualitative results are shown in Figure 4.9, with one view per row and one target prompt per column.

The results demonstrate that CDS effectively performs editing with strong color fidelity and spatial consistency. Compared to DDS, which uses denoising score gradients directly, CDS regularizes the editing process via attention map alignment between source and target prompts.

Notably, the yellow sphere edits exhibit saturation artifacts and mild leakage onto nearby regions, indicating that additional regularization or alpha-blending techniques may further enhance precision. Overall, the results confirm that CDS is capable of handling complex view-consistent edits in highly reflective scenes.

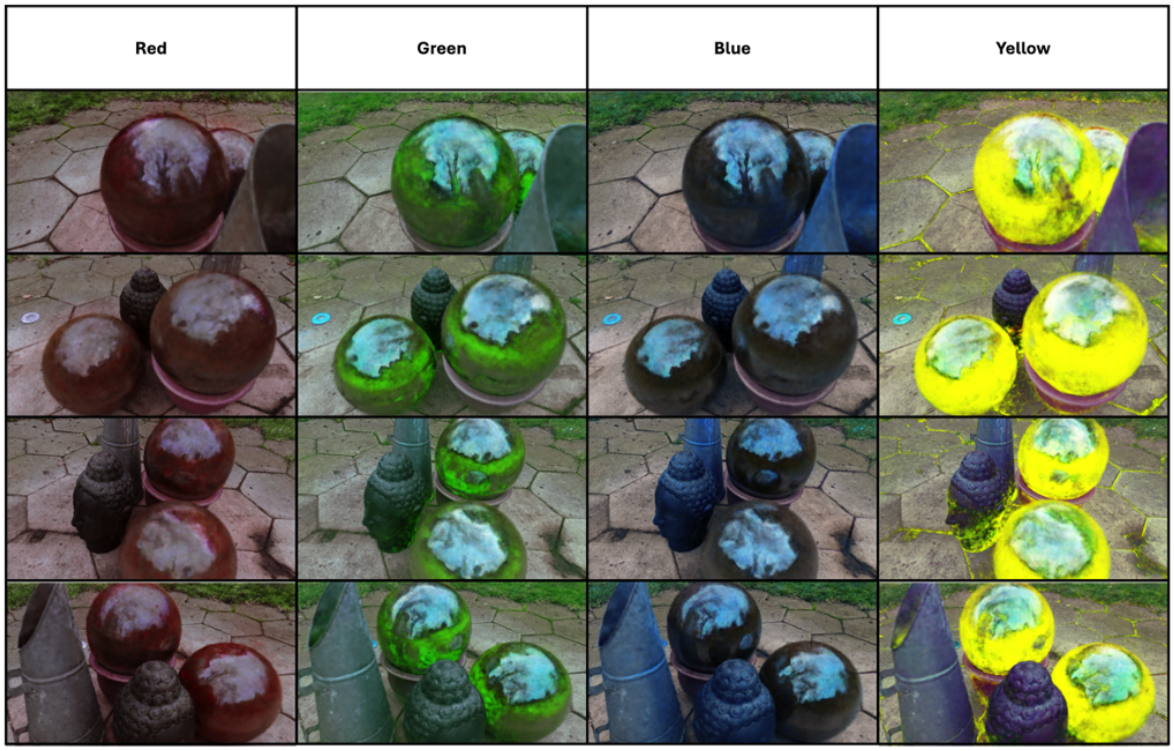

Figure 4.9: Multi-view editing results using the CDS pipeline [10] on the "gardenspheres" scene. Each column corresponds to a different target prompt: red, green, blue, and yellow reflective spheres. Each row shows a different view of the scene.

Although it is stated that the CDS method is more effective for localized gradients it can be seen from the results that it is not fully obtained. Possible reasons will be discussed in the next section.

Chapter 5

Limitations & Future Work

While the proposed framework demonstrates promising results, several limitations were encountered during the project that suggest directions for future improvement.

One primary limitation stems from the quality of the underlying NeRF reconstructions. The experiments conducted using default Nerfstudio settings yielded suboptimal geometry and appearance fidelity, often leading to the loss of fine details and structural inaccuracies. These limitations directly impact the effectiveness of the editing pipeline, as the quality of the base NeRF, serving as the scene prior, strongly influences the outcome of the optimization.

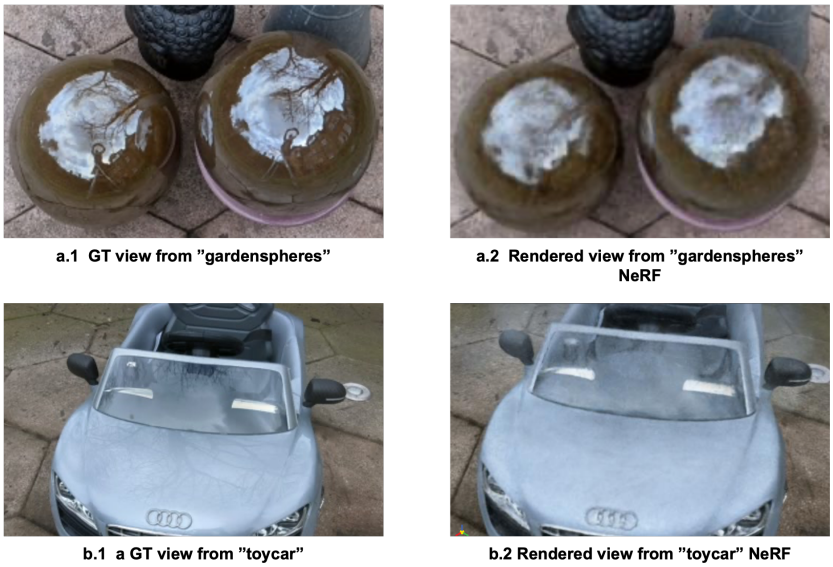

Figure 5.1: Qualitative comparison highlighting prior inconsistency in DreamCatalyst edits. (a.1) Original rendering from the "gardenspheres" dataset; (a.2) edited result with the target prompt. (b.1) Original rendering from the "toycar" dataset; (b.2) corresponding edited result.

To address this, future work could integrate recent advances such as NeRF-Casting [13], which improves rendering of reflective surfaces by explicitly modeling and blending reflected rays based on predicted surface roughness and reflectivity. NeRF-Casting introduces physically-informed enhancements to view-dependent appearance by learning spatially-varying material properties and using them to guide the blending of reflected RGB values. Preliminary integration of this method was attempted in the current work; however, it could not be fully implemented due to time constraints. Completing this integration could significantly enhance the predicted normals and improve the base geometry, leading to more stable and photorealistic editing results.

Another notable challenge was the scene-dependent instability observed in certain editing tasks. In some cases, the DreamCatalyst framework exhibited a tendency to revert or weaken previously successful semantic edits as optimization progressed. This behavior is illustrated in Figure 4.4, where color modifications achieved in early iterations were gradually undone in later stages. A potential cause of this issue is the identity regularization term, which is globally applied and may inadvertently suppress and even revert changes since the identity comparison is made with the original scene dataset. Future work could investigate improved spatially-aware or intelligent regularization, like CLIP encoding, to resolve such conflicts.

In addition, the wavelet-based gradient filtering mechanism, while effective in preserving low-frequency information such as shading and reflections, was applied uniformly across all spatial regions. A more adaptive filtering strategy could be explored in future work, where frequency suppression is modulated based on local semantic content, edge strength, or user guidance. This would enable finer control over which features are protected versus edited, supporting more expressive and targeted 3D scene modifications.

Lastly, the CDS (Contrastive Denoising Score) module introduced in Chapter 3 could not be fully integrated into the optimization loop due to time constraints. While the visual results show potential, the method currently suffers from issues such as gradient leakage and optimization instability. Its contrastive loss formulation requires careful tuning of hyperparameters (e.g., ratio of the loss values, number of iterations, learning rate, etc.) to achieve stable and meaningful gradients. Future work should systematically explore these design choices to unlock the full potential of attention-regularized semantic edits.

Together, these directions point toward a more robust, physically-informed, and semantically controllable NeRF editing framework.

Chapter 6

Conclusion

In this work, a diffusion-guided NeRF editing framework was presented, building upon the Dream-Catalyst architecture. The editing process was enhanced through wavelet-based gradient filtering, allowing low-frequency components to be preserved while suppressing high-frequency artifacts. Surface normal prediction was incorporated to improve geometric consistency, and geometry layers were kept frozen to maintain structural integrity during optimization.

Qualitative improvements in visual fidelity, reflection preservation, and semantic alignment were demonstrated across diverse scenes. Although limitations were identified and discussed in the preceding chapter, the results underscore the effectiveness of combining frequency-domain filtering with geometry-aware supervision. The proposed methodology provides a signal-processing-driven alternative for more controllable and structurally consistent 3D scene editing.

Bibliography

[1] Ayaan Haque et al. “Instruct-NeRF2NeRF: Editing 3D Scenes with Instructions”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 19683–19693. DOI: 10.1109/ICCV51070.2023.01808.

[2] Amir Hertz, Kfir Aberman, and Daniel Cohen-Or. “Delta Denoising Score”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 2328–2337. DOI: 10.1109/ICCV51070.2023.00221.

[3] Bernhard Kerbl et al. “3D Gaussian Splatting for Real-Time Radiance Field Rendering”. In: *ACM Transactions on Graphics* 42.4 (June 2023). URL: https://repo-sam.inria.fr/fungraph/3d-gaussian-splatting/.

[4] Jiwook Kim et al. “DreamCatalyst: Fast and High-Quality 3D Editing via Controlling Editability and Identity Preservation”. In: *International Conference on Learning Representations (ICLR)*. 2025. URL: https://arxiv.org/abs/2407.11394.

[5] Juil Koo, Chanho Park, and Minhyuk Sung. “Posterior Distillation Sampling”. In: *CVPR*. 2024.

[6] Chen-Hsuan Lin et al. “Magic3D: High-Resolution Text-to-3D Content Creation”. In: *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 300–309. DOI: 10.1109/CVPR52729.2023.00037.

[7] Ruoshi Liu et al. “Zero-1-to-3: Zero-shot One Image to 3D Object”. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2023, pp. 9264–9275. DOI: 10.1109/ICCV51070.2023.00853.

[8] MathWorks. *Introduction to the Wavelet Families*. https://www.mathworks.com/help/wavelet/gs/introduction-to-the-wavelet-families.html. Accessed: 2025-06-07. 2023.

[9] Ben Mildenhall et al. “NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis”. In: *ECCV*. 2020.

[10] Hyelin Nam et al. “Contrastive Denoising Score for Text-guided Latent Diffusion Image Editing”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. June 2024, pp. 9192–9201.

[11] Ben Poole et al. “DreamFusion: Text-to-3D using 2D Diffusion”. In: *International Conference on Learning Representations (ICLR)*. 2023. URL: https://doi.org/10.48550/arXiv.2209.14988.

[12] Robin Rombach et al. “High-Resolution Image Synthesis with Latent Diffusion Models”. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2022, pp. 10674–10685. DOI: 10.1109/CVPR52688.2022.01042.

[13] Dor Verbin et al. “NeRF-Casting: Improved View-Dependent Appearance with Consistent Reflections”. In: *SIGGRAPH Asia 2024 Conference Papers*. SA ’24. Tokyo, Japan: Association for Computing Machinery, 2024. URL: https://doi.org/10.1145/3680528.3687585.

[14] Zhengyi Wang et al. “ProlificDreamer: High-Fidelity and Diverse Text-to-3D Generation with Variational Score Distillation”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2023. URL: https://doi.org/10.48550/arXiv.2305.16213.

[15] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. “Adding Conditional Control to Text-to-Image Diffusion Models”. In: *IEEE International Conference on Computer Vision (ICCV)*. 2023.