3DGS-Enhancer: Enhancing Unbounded 3D Gaussian Splatting with View-consistent 2D Diffusion Priors

Xi Liu* **Chaoyi Zhou*** **Siyu Huang**
Visual Computing Division
School of Computing
Clemson University
{xi9, chaoyiz, siyuh}@clemson.edu

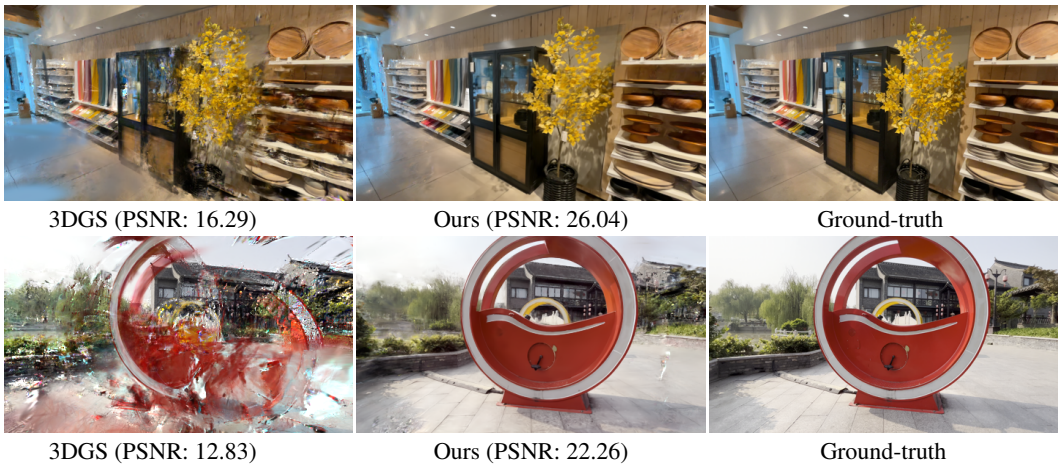


Figure 1: The 3DGS-Enhancer improves 3D Gaussian splatting representations on unbounded scenes with sparse input views.

Abstract

Novel-view synthesis aims to generate novel views of a scene from multiple input images or videos, and recent advancements like 3D Gaussian splatting (3DGS) have achieved notable success in producing photorealistic renderings with efficient pipelines. However, generating high-quality novel views under challenging settings, such as sparse input views, remains difficult due to insufficient information in under-sampled areas, often resulting in noticeable artifacts. This paper presents 3DGS-Enhancer, a novel pipeline for enhancing the representation quality of 3DGS representations. We leverage 2D video diffusion priors to address the challenging 3D view consistency problem, reformulating it as achieving temporal consistency within a video generation process. 3DGS-Enhancer restores view-consistent latent features of rendered novel views and integrates them with the input views through a spatial-temporal decoder. The enhanced views are then used to fine-tune the initial 3DGS model, significantly improving its rendering performance. Extensive experiments on large-scale datasets of unbounded scenes demonstrate that 3DGS-Enhancer yields superior reconstruction performance and high-fidelity rendering results compared to state-of-the-art methods. The project webpage is https://xiliu8006.github.io/3DGS-Enhancer-project.

Corresponding author: Siyu Huang

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

1 Introduction

Novel-view synthesis (NVS) has decades of history in computer vision and graphics communities, aiming to generate views of a scene from multiple input images or videos. Recently, 3D Gaussian splatting (3DGS) [18] has excelled in producing photorealistic renderings with a highly efficient rendering pipeline. However, rendering high-quality novel views far from existing viewpoints remains very challenging, as often encountered in sparse-view settings, due to insufficient information in under-sampled areas. As shown in Figure 1, noticeable ellipsoid-like and hollow artifacts manifest when there are only three input views. Due to these common low-quality rendering results in practice, it is essential to enhance 3DGS to ensure its viability for real-world applications.

To our knowledge, few prior studies have specifically focused on enhancement methods aimed at improving the rendering quality of NVS. Most existing enhancement work for NVS [19, 43] focuses on incorporating additional geometric constraints such as depth and normal into the 3D reconstruction process to fulfill the gap between the observed and unobserved regions. For example, DNGaussian [19] applies a hard-and-soft depth regularization to the geometry of radiance fields. However, these methods heavily rely on the effectiveness of additional constraint and are often sensitive to noises. Another line of work leverages generative priors to regularize the NVS pipeline. For instance, ReconFusion [40] enhances Neural Radiance Fields (NeRFs) [25] by synthesising the geometry and texture for the unobserved regions. Although it can generate photo-realistic novel views, the view consistency is still challenging when the generated views are far away from the input ones.

In this work, we exploit the 2D generative priors, *e.g.*, the latent diffusion models (LDMs) [31], for 3DGS representation enhancement. LDM has demonstrated powerful and robust generation capabilities in various image generation [31] and restoration tasks [42]. Nevertheless, the main challenge lies in the poor 3D view consistency among generated 2D images, which significantly hinders the 3DGS training process that requires highly precise view consistency. Although some efforts have been made, such as the Score Distillation Sampling (SDS) loss [29] that distills the optimization objective of a pre-trained diffusion model, it fails to generate the 3D representation allowing rendering high-fidelity images

Motivated by the analogy of the visual consistency between multi-view images and the temporal consistency between video frames, we propose to reformulate the challenging 3D consistency problem as an easier task of achieving temporal consistency within video generation, so we can leverage the powerful video diffusion models for restoring high-quality and view-consistent images. We propose a novel 3DGS enhancement pipeline, dubbed 3DGS-Enhancer. The core of 3DGS-Enhancer is a video LDM consisting of an image encoder that encodes latent features of rendered views, a video-based diffusion model that restores temporally consistent latent features, and a spatial-temporal decoder that effectively integrates the high-quality information in original rendered images with the restored latent features. The initial 3DGS model will be finetuned by these enhanced views to improve its rendering performance. The proposed 3DGS-Enhancer can be trajectory-free to reconstruct the unbound scenes from sparse views and generate the natural 3D representation for the invisible area between two known views. A cocurrent work V3D [7] also leverages latent video diffusion models [4] for generating object-level 3DGS models from single images. In contrast, our 3DGS-Enhancer focuses on enhancing any existing 3DGS models and thus can be applied to more generalized scenes, *e.g.*, the unbounded outdoor scenes.

In experiments, we generate large-scale datasets with pairs of low-quality and high-quality images on hundreds of unbounded scenes, based on DL3DV [20], for comprehensively evaluating the novelly investigated 3DGS enhancement problem. Empirical results demonstrate that the proposed 3DGS-Enhancer method achieves superior reconstruction performance on various challenging scenes, yielding more distinct and vivid rendering results. The code and the generated dataset will be publicly available. The contributions of this paper are summarized as follows.

1. To the best of our knowledge, this is the first work to tackle the problem of enhancing low-quality 3DGS rendering results, an issue that widely exists in practical 3DGS applications.

2. We propose a novel pipeline 3DGS-Enhancer that addresses the 3DGS enhancement problem. 3DGS-Enhancer reformulates the 3D-consistent image restoration task as temporally consistent video generation, such that powerful video LDMs can be leveraged for generating both high-quality and 3D-consistent images. Novel 3DGS fine-tuning strategies are

also devised for an effective integration of the enhanced views with the original 3DGS representation.

3. We conduct extensive experiments on large-scale datasets of unbounded scenes to demonstrate the effectiveness of the proposed methods over existing state-of-the-art few-shot NVS methods.

2 Related Work

Radiance fields for novel view synthesis. Novel view synthesis (NVS) aims to generate unseen viewpoints from a set of input images and camera information. Radiance fields methods, like NeRFs [25], encode 3D scenes as radiance fields and use volume rendering for novel views, achieving high-fidelity results but at the cost of lengthy training and inference times. Improvements such as Mip-NeRF [1, 2] enhance rendering quality through anti-aliasing, while others [6, 9, 46, 26] focus on speeding up the processes. Recently, 3D Gaussian splatting (3DGS) [18] has emerged, offering competitive rendering quality and significantly higher efficiency by representing scenes as 3D Gaussian spheres and using a fast differentiable splatting pipeline [49]. However, 3DGS still requires high-quality and numerous input views for optimal reconstruction, which is often impractical.

Few-shot novel view synthesis. Leveraging additional information is essential for generating novel views from sparse input images. Various approaches incorporate different regularization techniques to prevent 3D geometry from overfitting to the training views. [19, 10, 27, 22] introduce extra geometric information, such as depth maps or coarse mesh, to enhance the robustness and performance of 3D reconstruction from sparse views. [5, 8] leverage the learned priors from multi-view stereo datasets as general priors to improve performance in sparse view reconstruction tasks. FreeNeRF [43] integrates frequency and occlusion regularization during training to mitigate overfitting issues in few-shot neural rendering. Similarly, DietPixelNeRF [16] employs a semantic view consistency loss to ensure that all views share consistent semantics, thereby alleviating overfitting. However, these methods are highly sensitive to the network's performance, where incorrect depth estimations or inaccurate mesh reconstructions can significantly degrade the final output.

Diffusion priors for novel view synthesis. Recently, utilizing diffusion models as priors for few-shot novel view synthesis has proven to be an effective approach. DreamFusion [29] employs Score Distillation Sampling (SDS) with a pre-trained diffusion model to guide 3D object generation from text prompts [35, 32, 45]. Some works [21, 33, 34] embed 3D awareness into 2D diffusion models to generate multi-view images, though these methods typically require large datasets [48] and significant training resources [16, 27]. ReconFusion [40] leavarge the 2D diffusion priors to recover a high-fidelity NeRF from sparse input views. More advanced approaches leverage video diffusion models [4, 12, 13, 23] for few-shot NVS. For instance, AnimateDiff [11] fine-tunes diffusion models with additional camera motions using LoRA [14], while methods like SVD-MV [4], V3D [36] and IM-3D [23] propose camera-controlled video diffusion models for object-level 3D generation. In contrast, our approach offers greater generalizability for unbounded outdoor scenes.

Radiance fields enhancement. Several existing studies focus on enhancing NeRFs by addressing the limited detail preservation issue caused by insufficient or low-quality input data. NeRF-SR [37] and Refsr-nerf [15] use a super-resolution network to upscale the training view images, allowing novel views to be synthesized at higher resolutions with appropriate details. Alignerf [17] introduce optical-flow network to solve the misalignment problem to enhance the performance. Some other approaches incorporate 2D diffusion priors into 3D reconstructions. For instance, DiffusionNeRF [41] leverages a diffusion model to learn gradients of logarithms of RGBD patch priors, serving as regularized geometry and color for a scene. Nerfbusters [39] use diffusion priors to remove ghostly artifacts in the 3D gaussians. Our work aim to addresses the radiance fields enhancement problem by proposing a novel framework 3DGS-Enhancer, achieving superior enhancement performance for low-quality unbounded 3DGS representations.

3 Preliminary of 3D Gaussian Splatting

Here, we briefly review the formulation and rendering process of 3DGS [18]. 3DGS represents a scene as a set of anisotropic 3D Gaussian spheres, allowing high-fidelity NVS with extremely low rendering latency. A 3D Gaussian sphere includes a center position $\mu \in \mathbb{R}^3$, a scaling factor $\mathbf{s} \in \mathbb{R}^3$,

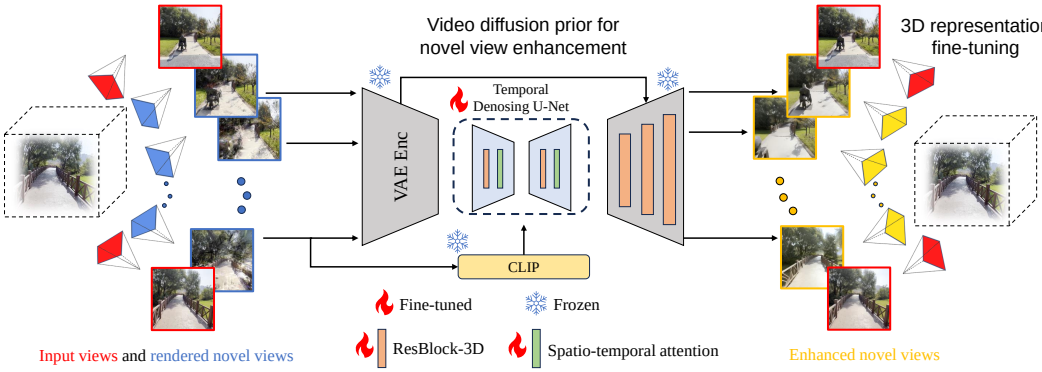


Figure 2: An overview of the proposed 3DGS-Enhancer framework for 3DGS representation enhancement. We learn 2D video diffusion priors on a large-scale novel view synthesis dataset to enhance the novel views rendered from the 3DGS model on a novel scene. Then, the enhanced views and input views jointly fine-tune the 3DGS model.

and a rotation quaternion $\mathbf{q} \in \mathbb{R}^4$, such that the Gaussian distribution is

$$G(x) = e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)}, \tag{1}$$

where $\Sigma = RSS^T R^T$, S is the scaling matrix determined by $\mathbf{s}$ and R is the rotation matrix determined by $\mathbf{q}$. To additionally model the view-dependent appearance, the Gaussian sphere also includes spherical harmonics (SH) coefficients $\mathcal{C} \in \mathbb{R}^k$, where k is the number of SH functions, and an $\alpha \in \mathbb{R}$ for opacity. The color and opacity are also calculated by the Gaussian distribution illustrated in Eq. 1.

For rendering, all the 3D Gaussian spheres are projected onto the 2D camera planes via a differentiable Gaussian splatting pipeline [49]. Given the viewing transform matrix W and Jacobian matrix J of the affine approximation of the projective transformation, the covariance matrix Σ' in camera coordinates is calculated as

$$\Sigma' = JW\Sigma W^T J^T. \tag{2}$$

The differentiable splatting method efficiently projects the 3D Gaussian spheres to 2D Gaussian distributions, ensuring fast α-blending for rendering and color supervision. For each pixel, the color is rendered by M Gaussian spheres that overlap with the pixel on the 2D camera planes, sorted in the depth distance as

$$C = \sum_{i \in M} \mathcal{C}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_i). \tag{3}$$

4 Method

4.1 3DGS-Enhancer: An Overview

This work studies the 3DGS enhancement problem. More specifically, given a 3DGS model trained on a scene consisting of input views $\{I_1^{\text{ref}}, I_2^{\text{ref}}, \ldots, I_{N_{\text{ref}}}^{\text{ref}}\}$ and corresponding camera poses $\{\boldsymbol{p}_1^{\text{ref}}, \boldsymbol{p}_2^{\text{ref}}, \ldots, \boldsymbol{p}_{N_{\text{ref}}}^{\text{ref}}\}$, the goal of this work is to enhance a set of low-quality novel views $\{I_1, I_2, I_3, \ldots, I_{N_{new}}\}$ rendered by the 3DGS model. The enhanced images further fine-tune the 3DGS model to improve its reconstruction and rendering quality.

This work novelly reformulates the challenging task of 3D-consistent image restoration as the task of video restoration, in light of the analogy between the multi-view consistency and the video temporal consistency. We propose a novel framework named 3DGS-Enhancer that employs a video LDM comprising an image encoder, a video-based diffusion model, and a spatial-temporal decoder to enhance the rendered images while preserving a high 3D consistency. 3DGS-Enhancer also adopts novel fine-tuning strategies to selectively integrate the views enhanced by the video LDM into the 3DGS fine-tuning process. An illustration of the 3DGS-Enhancer framework is shown in Figure 2. We discuss more details of the framework in the following.

4.2 Video Diffusion Prior for Temporal Interpolation

In this section, we introduce the video diffusion model for achieving 3D-consistent 2D image restoration. To lift the consistency between the generated 2D video frames and the high-quality reference views, we further propose to formulate the video restoration task as a video interpolation task, where the first frame and the last frame of inputs to the video diffusion model are two reference views. This formulation provides stronger guidance for the video restoration process. Let $\{\boldsymbol{p}^{\text{ref}}_{i-1}, \boldsymbol{p}^{s}_{1}, \boldsymbol{p}^{s}_{2}, \ldots, \boldsymbol{p}^{s}_{T}, \boldsymbol{p}^{\text{ref}}_{i}\}$ be the camera poses sampled from the trajectory fitted between two reference views, the images rendered accordingly are $v = \{I^{\text{ref}}_{i-1}, I_1, I_2, \ldots, I_T, I^{\text{ref}}_{i}\}$. $v \in \mathbb{R}^{(T+2)\times 3\times H\times W}$ serves as the input to the video diffusion model, *e.g.*, a pre-trained image-guided stable video diffusion (SVD) model [4] that adopts cross-frame spatio-temporal attention module and 3D residual convolution in the diffusion U-Net. Unlike SVD, which repeats the single input image feature extracted by CLIP [30] for T times as the conditional inputs, we input $\boldsymbol{v}$ to the CLIP encoder to get a sequence of conditional inputs $\boldsymbol{c}_{\text{clip}}$ and add it to the video diffusion model through cross attention. Meanwhile, we input $\boldsymbol{v}$ to the VAE encoder to get latent feature $\boldsymbol{c}_{\text{vae}}$ and add it into the diffusion model through a classifier-free guidance strategy to incorporate richer color information. The diffusion U-Net ϵ_θ predicts the noise ϵ for each diffusion step t, and the training objective is

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}\left[\|\epsilon - \epsilon_\theta(z_t, t, \boldsymbol{c}_{\text{clip}}, \boldsymbol{c}_{\text{vae}})\|\right]. \tag{4}$$

where $z_t = \alpha_t z + \sigma_t \epsilon * ..$ where z is the gt latent, $\epsilon \in \mathcal{N}(0, I)$, α_t and σ_t define a noise at timestep t. The learned video diffusion model generates a sequence of enhanced image latents z_v corresponding to the rendered low-quality views v.

4.3 Spatial-Temporal Decoder

Although the video diffusion model can generate enhanced image latents z_v, we observe that there are artifacts such as temporal inconsistency, blurring, and color shift in outputs of the original decoder of video LDM. To address this issue, we propose a modified spatial-temporal decoder (STD). STD makes the following improvements over the original VAE Decoder: 1) **Temporal decoding manner.** STD adopts additional temporal convolution layers to ensure the temporal consistency between decoded outputs. Similar to our video diffusion model, the first and the last input frames are the reference view images, and the intermediate inputs are the generated views; 2) **Effective integration of rendered views.** STD adopts additional conditional inputs, same as those of the video diffusion model, allowing the decoder to better leverage the original rendered images. Inspired by [44, 47], these conditional inputs are fed into STD through Controllable Feature Warping (CFW) modules [38], such that their high-frequency patterns are better preserved. 3) **Color correction.** To address the color shift issue, we apply color normalization to the decoded images by following StableSR [38]. However, we observe that highly blurred and low-quality images in the conditional inputs can undermine the color correction effects. To mitigate this, we use the first reference view to calculate the mean and variance, and then align all the other decoded images with this reference view. Let I^g_i be the i-th decoded images with a mean $\mathbf{u}_{\hat{I}^g_0}$ and a variance $\sigma_{\hat{I}^g_0}$, $\hat{I}^g_0$ be the reference view with a mean $\mu_{I^g_i}$ and a variance $\sigma_{I^g_i}$, the corrected image I^c_i is computed by:

$$I^c_i = \frac{I^g_i - \mu_{I^g_i}}{\sigma_{I^g_i}} \cdot \sigma_{\hat{I}^g_0} + \mathbf{u}_{\hat{I}^g_0}. \tag{5}$$

The optimization objective of STD consists of an L1 reconstruction loss and an LPIPS perceptual loss between I^g and ground-truth $\hat{I^g}$, and an adversarial loss, as

$$\mathcal{L}_{\text{STD}} = \mathcal{L}_{\text{rec}}(I^g, \hat{I^g}) + \mathcal{L}_{\text{LPIPS}}(I^g, \hat{I^g}) + \mathcal{L}_{\text{adv}}(I^g). \tag{6}$$

where $\mathcal{L}_{\text{adv}}$ is the adversarial loss that discriminates between real image $\hat{I^g}$ and fake image I^g.

4.4 Fine-tuning Strategies of 3D Gaussian Splatting

Confidence-aware 3D Gaussian splatting. Unlike existing sparse-view NVS methods, our approach does not rely on depth estimation networks for depth regularization. Instead, we take a purely 2D visual method by utilizing a video diffusion model to enhance images rendered from a low-quality 3DGS model. Despite this significant enhancement in the quality of the rendered views, we propose

Figure 3: The red circle indicates the area with high confidence, meaning the generated videos can contribute more information. Conversely, the green quadrilateral highlights the area with low confidence, suggesting that the generated video should not tend to optimize this area.

to rely more on the reference views rather than the restored novel views when fine-tuning the 3DGS model, since the 3DGS model is highly sensitive to slight inaccuracies in the restored views. These inaccuracies could be amplified during the fine-tuning process.

To minimize the negative impact of generated images on Gaussian training, we propose confidence-aware 3D Gaussian splatting. This strategy involves two levels of confidence, image level and pixel level. For the image level, the generated images that are closer to real images have lower confidence. For pixel level, the larger the mean covariance of all the Gaussians used to render this pixel, the higher its confidence.

Image level confidence. In the task of novel view synthesis, if noise exists in two image views, a close distance between them increases the likelihood of generating conflicts and disrupting the 3D consistency of the scene. Therefore, for novel views that are close to the reference view, it is crucial to carefully optimize the 3D Gaussians to mitigate the adverse effects of noise. Conversely, when a novel view is far from all known views, it has a smaller likelihood of disturbing already well-reconstructed areas. Based on this reasoning, we normalize the distance from novel views to reference views between 0 and 1. The farther a viewpoint is from the reference view, the higher its confidence.

Pixel level confidence. Inspired by ActiveNeRF [28], which uses Gaussian distributions in NeRF to estimate uncertainty and identify views with the highest information gain, we aim to find the pixels that can provide the highest information gain from the generated images. As shown in Fig 3, we observed that well-reconstructed areas are typically represented by Gaussians with very small volumes, calculated using the scaling vector $s \in \mathbb{R}$. Based on this observation, we propose a method to calculate pixel-level confidence.

The unique representation of 3D Gaussians allows us to render an H×W×3 image using a process similar to rendering colors, where each channel corresponds to one of the three components of the scaling vector $\mathbf{s}$. In 3DGS-Enhancer, we multiply these three channels of the scale map to obtain pixel-level confidence. For each pixel in the generated images, higher confidence results in greater weight in supervising the training of the 3DGS model.

Given a set of 3D Gaussian, the 3-channel C_{conf} confidence map is rendered as same as colour rendering, and the formula is defined as follows

$$C_{conf} = \sum_{i \in M} \mathbf{s}_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_i). \tag{7}$$

Table 1: **A quantitative comparison of few-shot 3D reconstruction**. Experiments on DL3DV and LLFF follow the setting of [43]. Experiments on Mip-NeRF 360 follow the setting of [40].

Method	3 views			6 views			9 views		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DL3DV (130 training scenes, 20 test scenes)									
Mip-NeRF [1]	10.92	0.191	0.618	11.56	0.199	0.608	12.42	0.218	0.600
RegNeRF [27]	11.46	0.214	0.600	12.69	0.236	0.579	12.33	0.219	0.598
FreeNeRF [43]	10.91	0.211	0.595	12.13	0.230	0.576	12.85	0.241	0.573
3DGS [18]	10.97	0.248	0.567	13.34	0.332	0.498	14.99	0.403	0.446
DNGaussian [19]	11.10	0.273	0.579	12.67	0.329	0.547	13.44	0.365	0.539
3DGS-Enhancer (ours)	**14.33**	**0.424**	**0.464**	**16.94**	**0.565**	**0.356**	**18.50**	**0.630**	**0.305**

And the 1 channel pixel level confidence map $P_c = \sqrt[3]{C_{conf}[0] \times C_{conf}[1] \times C_{conf}[2]}$. Overall, in our training process for 3D Gaussians, the loss functions were defined as

$$\mathcal{L}_{\text{3DGS}} = I_c \cdot (P_c \odot \|C - \hat{C}\|_1 + SSIM(C, \hat{C})) \tag{8}$$

where SSIM is the Structural Similarity Index and $\odot$ is Hadamard's product, I_c is the image-level confidence map and $\hat{C}$ is the real pixel value.

5 Experiments

5.1 3DGS-Enhance Dataset

Given that the enhancement of 3DGS representations is a new task, we create a dataset to simulate various artifacts of the 3DGS representations. This dataset also serves as a more comprehensive benchmark for evaluating the performance of few-shot NVS methods. Existing few-shot NVS algorithms [43, 19] primarily focus on face-forward evaluations [24], where the test views have significant overlap with the input views. However, this evaluation method is not suitable for large-scale unbounded outdoor scenes. Therefore, we propose a dataset processing strategy that allows us to post-process any existing multi-view dataset to generate a large number of training image pairs that include typical artifacts caused by few-shot NVS.

More specifically, for each scene, we have n views $I_{\text{train}} = \{I_1, I_2, \ldots, I_n\}$, which serve as the input for a high-quality 3DGS model. We uniformly sample a small number m of views I_{low} from I_{train}, which serve as the input for the low-quality 3DGS model. By linearly fitting the high-quality camera poses $p_i^{\text{train}} = \{p_1^{\text{train}}, p_2^{\text{train}}, \ldots, p_{n^*}^{\text{train}}\}$, we randomly sample a camera trajectory $p_i^{\text{render}} = \{p_1^{\text{render}}, p_2^{\text{render}}, \ldots, p_{n^*}^{\text{render}}\}$ on p_i^{train} and render the image pairs using both high-quality and low-quality 3DGS models. This creates a set of high-quality and low-quality image pairs used for the training of our video diffusion model.

We apply this dataset processing strategy to DL3DV [20], a large-scale outdoor dataset containing 10K scenes. We randomly select 130 scenes from the original DL3DV dataset and form more than 150,000 image pairs. We randomly select another 20 scenes from DL3DV to form the test sets, evaluating the corss-scene capability of our method. More implementation details of the method can be found in the supplementary material.

5.2 Comparison with State-of-the-Arts

The quantitative and qualitative results on the DL3DV test set with 3 6 and 9 input views are shown in Table 1 and Figure 4. Our approach outperforms all the other baselines in PSNR, SSIM, and LPIPS scores. NeRF-based methods including Mip-NeRF [1] and FreeNeRF [43] produce blurry novel views due to smoothing inconsistencies. In contrast, 3DGS [18] generates elongated elliptical artifacts due to local minima convergence. DNGuassian [19] reduces artifacts with depth regularization but results in blurry and noisy novel views.

The first example in Figure 4 demonstrates 3DGS-Enhacer's capability to remove artifacts while preserving view consistency. By interpolating input views using a video diffusion model, we incorporate more information while enrusing a high view consistency, enabling high-quality novel

Figure 4: A visual comparison of rendered images on scenes from DL3DV [20] test set with the 3-view setting.

views and avoiding local minima. The second example highlights 3DGS-Enhancer's advantage in recovering high-frequency details. Our dataset processing strategy and video diffusion model enable an understand of strong multi-view prior across various scenes. As a result, very challenging cases such as the trees can be restored with sharp details. In summary, comparisons with baseline methods demonstrate our approach's potential to significantly improve the unbounded 3DGS representations, synthesizing high-fidelity novel views for open environments.

To demonstrate the generalizability of our method for out-of-distribution dataset, we train the methods on the DL3DV-10K dataset [20] and test them on the Mip-NeRF360 dataset [2]. The results, as

3DGS (PSNR: 16.15) Ours (PSNR: 22.32) Ground-truth

3DGS (PSNR: 13.90) Ours (PSNR: 22.32) Ground-truth

Figure 5: A visual comparison of cross-dataset generalization ability, where the methods are trained on the DL3DV-10K dataset [20] and tested on the Mip-NeRF360 dataset [2].

summarized in Table 2 and Fig 5, show that our method outperforms the baseline approaches, highlighting its remarkable generalization capabilities in unbounded environments.

Table 2: A quantitative comparison of methods on the unseen Mip-NeRF360 dataset [2].

Method	6 views			9 views		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
Mip-NeRF360 (all test scenes)						
Mip-NeRF	13.08	0.159	0.637	13.73	0.189	0.628
RegNeRF	12.69	0.175	0.660	13.73	0.193	0.629
FreeNeRF	12.56	0.182	0.646	13.20	0.198	0.635
3DGS	11.53	0.144	0.651	12.65	0.187	0.607
DNGaussian	11.81	0.208	0.689	12.51	0.228	0.683
3DGS-Enhancer (ours)	**13.96**	**0.260**	**0.570**	**16.22**	**0.399**	**0.454**

5.3 Ablation Study

Real image as reference views. Table 3 shows the quantitative comparisons of different components in 3DGS-Enhancer framework. The video diffusion model provides strong multi-view priors. However, due to its native restrictions, we directly feed the original input views into the 3DGS fine-tuning process. This results in more reliable and view-consistent information from the input domain to facilitate 3DGS fine-tuning, as demonstrated by the "Real image" in Table 3.

Confidence aware reweighting. Distant views are less likely to cause artifacts, so we normalize their distance to reference views between [0, 1], giving higher confidence of video diffusion results to farther viewpoints. This strategy is denoted by "Image confidence" in Table 3. Pixel-level confidence, as denoted by "Pixel confidence" in Table 3, is based on the density of small-volume Gaussians in well-reconstructed areas, using a color rendering pipeline to calculate volumes. Both pixel and image-level confidence strategies improve results individually, and their combination yields the best performance.

Video diffusion and STD. Figure 6 visualizes the effects of video diffusion and STD module, respectively. Video diffusion removes most of the artifacts, and STD module enhances fine-grained and high-frequency textures, resulting in more vivid novel view renderings, which are closer to the ground truth. Table 4 shows the improvment for each modules.

Table 3: An ablation study of the four modules of our 3DGS-Enhancer framework, where all results are averaged across 3, 6, 9, and 12 input views on DL3DV dataset [20].

Video diffusion	Real image	Image confidence	Pixel confidence	PSNR↑	SSIM↑	LPIPS↓
✓	-	-	-	14.33	0.476	0.422
✓	✓	-	-	17.01	0.553	0.361
✓	✓	✓	-	17.29	0.570	0.354
✓	✓	-	✓	17.16	0.564	0.351
✓	✓	✓	✓	**17.34**	**0.574**	**0.351**

Table 4: An ablation study of STD (temporal layers) and color correction module on the DL3DV test dataset with a 9-view setting.

Video diffusion	STD (temporal layers)	color correction	PSNR ↑	SSIM ↑	LPIPS ↓
✓	-	-	18.11	0.591	0.312
✓	✓	-	18.44	0.625	0.306
✓	✓	✓	**18.50**	**0.630**	**0.305**

Input Diffusion STD Ground-truth

Figure 6: An ablation study of the video diffusion model components in our 3DGS-Enhancer framework.

6 Conclusions, Limitations, and Future Work

This paper has introduced 3DGS-Enhancer, a unified framework that applies view-consistency prior from video diffusion and use trajectory interpolation method to enhance unbounded 3DGS representations. By combining image and pixel-level confidence with 3DGS fine-tuning, we have achieved state-of-the-art performance in NVS enhancement. However, our approach relies on adjacent views for continuous interpolation, it cannot be easily adapted to single-view 3D model generation. Moreover, the confidence-aware 3DGS fine-tuning strategies are relatively simple and straightforward. In the future, it is interesting to integrate confidence maps directly with the video generation model, enabling the generation of images that are more in line with the real 3D world without the need for post-processing. Meanwhile, utilizing the efficient data generation capability of 3DGS to construct a massively scaled dataset for our video generation model presents a prime opportunity to enhance the model's 3D consistency. This approach also facilitates the 2D models to understand the 3D world directly from 2D images without additional geometric constraints. Regarding the social impact, the goal of this work is to advance the fields of 3D reconstruction and NVS. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

7 Acknowledgement

The authors gratefully acknowledge the Clemson University Palmetto Cluster for providing the high-performance computing resources that supported the computations of this work.

References

[1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5855–5864, 2021.

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022.

[3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *ICCV*, 2023.

[4] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[5] David Charatan, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. pixelsplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19457–19467, 2024.

[6] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European conference on computer vision*, pages 333–350. Springer, 2022.

[7] Zilong Chen, Yikai Wang, Feng Wang, Zhengyi Wang, and Huaping Liu. V3d: Video diffusion models are effective 3d generators. *arXiv preprint arXiv:2403.06738*, 2024.

[8] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis from sparse views of novel scenes. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2021.

[9] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022.

[10] Guangcong, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[11] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *International Conference on Learning Representations*, 2024.

[12] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P. Kingma, Ben Poole, Mohammad Norouzi, David J. Fleet, and Tim Salimans. Imagen video: High definition video generation with diffusion models, 2022.

[13] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022.

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[15] Xudong Huang, Wei Li, Jie Hu, Hanting Chen, and Yunhe Wang. Refsr-nerf: Towards high fidelity and super resolution view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8244–8253, June 2023.

[16] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5885–5894, October 2021.

[17] Yifan Jiang, Peter Hedman, Ben Mildenhall, Dejia Xu, Jonathan T. Barron, Zhangyang Wang, and Tianfan Xue. Alignerf: High-fidelity neural radiance fields via alignment-aware training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 46–55, June 2023.

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023.

[19] Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Dngaussian: Optimizing sparse-view 3d gaussian radiance fields with global-local depth normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20775–20785, 2024.

[20] Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. *arXiv preprint arXiv:2312.16256*, 2023.

[21] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9298–9309, 2023.

[22] Xiaoxiao Long, Cheng Lin, Peng Wang, Taku Komura, and Wenping Wang. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *European Conference on Computer Vision*, pages 210–227. Springer, 2022.

[23] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, Natalia Neverova, Andrea Vedaldi, Oran Gafni, and Filippos Kokkinos. Im-3d: Iterative multiview diffusion and reconstruction for high-quality 3d generation. *International Conference on Machine Learning, 2024*, 2024.

[24] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019.

[25] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[26] Muhammad Husnain Mubarik, Ramakrishna Kanungo, Tobias Zirr, and Rakesh Kumar. Hardware acceleration of neural graphics. In *Proceedings of the 50th Annual International Symposium on Computer Architecture*, pages 1–12, 2023.

[27] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022.

[28] Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. Activenerf: Learning where to see with uncertainty estimation. In *European Conference on Computer Vision*, pages 230–246. Springer, 2022.

[29] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023.

[30] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[32] Kyle Sargent, Zizhang Li, Tanmay Shah, Charles Herrmann, Hong-Xing Yu, Yunzhi Zhang, Eric Ryan Chan, Dmitry Lagun, Li Fei-Fei, Deqing Sun, and Jiajun Wu. ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. *CVPR, 2024*, 2023.

[33] Ruoxi Shi, Hansheng Chen, Zhuoyang Zhang, Minghua Liu, Chao Xu, Xinyue Wei, Linghao Chen, Chong Zeng, and Hao Su. Zero123++: a single image to consistent multi-view diffusion base model, 2023.

[34] Yichun Shi, Peng Wang, Jianglong Ye, Long Mai, Kejie Li, and Xiao Yang. MVDream: Multi-view diffusion for 3d generation. In *The Twelfth International Conference on Learning Representations*, 2024.

[35] Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653*, 2023.

[36] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitrii Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. SV3D: Novel multi-view synthesis and 3D generation from a single image using latent video diffusion. *arXiv*, 2024.

[37] Chen Wang, Xian Wu, Yuan-Chen Guo, Song-Hai Zhang, Yu-Wing Tai, and Shi-Min Hu. Nerf-sr: High quality neural radiance fields using supersampling. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 6445–6454, 2022.

[38] Jianyi Wang, Zongsheng Yue, Shangchen Zhou, Kelvin CK Chan, and Chen Change Loy. Exploiting diffusion prior for real-world image super-resolution. In *arXiv preprint arXiv:2305.07015*, 2023.

[39] Frederik Warburg, Ethan Weber, Matthew Tancik, Aleksander Holynski, and Angjoo Kanazawa. Nerfbusters: Removing ghostly artifacts from casually captured nerfs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 18120–18130, October 2023.

[40] Rundi Wu, Ben Mildenhall, Philipp Henzler, Keunhong Park, Ruiqi Gao, Daniel Watson, Pratul P Srinivasan, Dor Verbin, Jonathan T Barron, Ben Poole, et al. Reconfusion: 3d reconstruction with diffusion priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21551–21561, 2024.

[41] Jamie Wynn and Daniyar Turmukhambetov. DiffusioNeRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models. In *CVPR*, 2023.

[42] Bin Xia, Yulun Zhang, Shiyin Wang, Yitong Wang, Xinglong Wu, Yapeng Tian, Wenming Yang, and Luc Van Gool. Diffir: Efficient diffusion model for image restoration. *ICCV*, 2023.

[43] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[44] Xi Yang, Chenhang He, Jianqi Ma, and Lei Zhang. Motion-guided latent diffusion for temporally consistent real-world video super-resolution. In *European Conference on Computer Vision*, pages 224–242. Springer, 2025.

[45] Taoran Yi, Jiemin Fang, Junjie Wang, Guanjun Wu, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Qi Tian, and Xinggang Wang. Gaussiandreamer: Fast generation from text to 3d gaussians by bridging 2d and 3d diffusion models. In *CVPR*, 2024.

[46] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *arXiv*, 2021.

[47] Shangchen Zhou, Peiqing Yang, Jianyi Wang, Yihang Luo, and Chen Change Loy. Upscale-a-video: Temporal-consistent diffusion model for real-world video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2535–2545, 2024.

[48] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018.

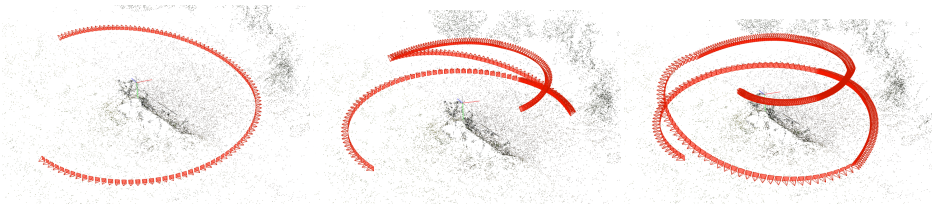
[49] Matthias Zwicker, Hanspeter Pfister, Jeroen Baar, and Markus Gross. Surface splatting. *Proceedings of the ACM SIGGRAPH Conference on Computer Graphics*, 2001, 08 2001.

8 Appendix

8.1 Details of 3DGS Enhancement Dataset

For our 3DGS Enhancement Dataset, constructed based on DL3DV, we randomly select 120 scenes to create the training set for our video diffusion model and 30 scenes as the test set. By following previous works, we use the standard train/test split, selecting every 8th frame of the remaining frames for evaluation.

To create image pairs simulating the artifacts due to the lack of input views in novel view synthesis problem, we render the image pairs from pairs of low-high quality 3DGS models. Specifically, the input views for the high-quality model consist of all images in the original dataset, while the inputs for the low-quality model are a subset uniformly sampled from the original dataset. To add more complexity, we sample the subset according to a certain number (e.g., 3, 6, 9) or a certain ratio (e.g., 5%). With the aim to fully capture the distribution of artifacts created by the sparse input views and train the video diffusion model with smoother inputs, we propose a heuristic trajectory fitting algorithm, as shown in Figure 7, proving a sequence of cameras by interpolating the low or high-quality model's input views. Specifically, if the original camera trajectories are smooth and simple, such as those of DL3DV, we use the high-quality input views as the reference to fit the trajectories. For complex trajectories, such as those in Mip-NeRF 360, we use the low-quality input to avoid significantly poor rendering results, which would lead to unreasonable artifact distributions. As a result, we render a large number of image pairs with and without artifacts, as shown in Figure 8, at a resolution of 512×512, leading to powerful video diffusion priors with high view consistency and photo-realism.



Input: 3 Input: 6 Input: 9

Figure 7: The fitting trajectories under different number of input views.

8.1.1 Training details

Our video diffusion model includes a pre-trained VAE to encode an image sequence into a latent sequence and decode the latent sequence back into the image sequence. It also includes a U-Net with learnable temporal layers, which employs cross-frame attention modules and 3D CNN modules to ensure frame-consistent outputs. The input of video diffusion model is a image sequence segment that includes 25 images with different sample steps from the image sequences rendered from the low-quality 3DGS model. The first and the last frames in this segment are replaced with images rendered from the high-quality 3DGS model. During fine-tuning, our video diffusion model is conditioned on these image sequence segments and trained to synthesize the corresponding segments rendered from the high-quality 3DGS model.

Our video diffusion model is fine-tuned with a learning rate of 0.0001, incorporating 500 steps for warm-up, followed by a total of 80,000 training steps. The batch size is set to 1 in each GPU, where each batch consisted of 25 images at 512x512 resolution. To optimize the training process, the Adam optimizer is employed. Additionally, a dropout rate of 0.1 is applied to the conditions between the first and last frames and the training process utilize CFG (classifier-free guidance) to train the diffusion model. The training is conducted on 2 NVIDIA A100-80G GPUs over 3 days. The STD is fine-tuned with a learning rate of 0.0005 and 50,000 training steps. The batch size is set to 1 in each GPU, where each batch consists of 5 images at 512x512 resolution, but for inference, it was increased to 25. The fine-tuning process is conducted on 2 NVIDIA A100-80G GPUs in 2 days. The entire pipeline's inference and training speeds were evaluated and are presented in Table 5.

Figure 8: The low and high quality image pairs created in our 3DGS Enhancement dataset.

Table 5: A comparison of per-scene training time and rendering FPS between methods. For our method, the LQ-3DGS reconstruction takes 10.5 minutes, stable video diffusion inference for 50 novel views requires 2.0 minutes, and the HQ-3DGS reconstruction takes 12.0 minutes.

Method	**Per-scene training time** ↓	**Rendering FPS** ↑
Mip-NeRF	10.7h	0.09
RegNeRF	2.5h	0.09
FreeNeRF	3.8h	0.09
3DGS	10.5min	100
DNGaussian	3.3min	100
3DGS-Enhancer (ours)	24.5min	100

8.2 Details of Comparison Baselines

For the evaluation datasets, we compare against the standard 3D Gaussian Splatting [18] (which is also the reconstruction pipeline used in our work), and the state-of-the-art few-view NVS regularization methods, including Mip-NeRF [1], FreeNeRF [43], Zip-NeRF [3], and RegNeRF [27]. We also compare to some few-shot NVS methods using generative priors including ZeroNVS [32], and ReconFusion [40].

For the evaluation of MipNeRF, FreeNeRF, RegNeRF, and DNGaussian on DL3DV and Mip-NeRF 360 dataset, we follow the original configurations and code shared by the authors. Additionally, we use random point cloud as the initialization for 3DGS, following the implementations from DNGaussian. We also decrease the batch size for RegNeRF from 4096 to 512 according to the limited computation resource.