This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

# DIFIX3D+: Improving 3D Reconstructions with Single-Step Diffusion Models

Jay Zhangjie Wu[1,2*] Yuxuan Zhang[1*] Haithem Turki[1] Xuanchi Ren[1,3,4] Jun Gao[1,3,4]
Mike Zheng Shou[2] Sanja Fidler[1,3,4] Zan Gojcic[1†] Huan Ling[1,3,4†]
[1]NVIDIA [2]Show Lab, National University of Singapore [3]University of Toronto [4]Vector Institute
https://research.nvidia.com/labs/toronto-ai/difix3d


Figure 1. **We demonstrate DIFIX3D+ on both in-the-wild scenes (*top*) and driving scenes (*bottom*).** Recent Novel-View Synthesis methods struggle in sparse-input settings or when rendering views far from the input camera poses. DIFIX distills the priors of 2D generative models to enhance reconstruction quality and can further act as a neural-renderer at inference time to mitigate the remaining inconsistencies. Notably, the same model effectively corrects NeRF [37] and 3DGS [20] artifacts.

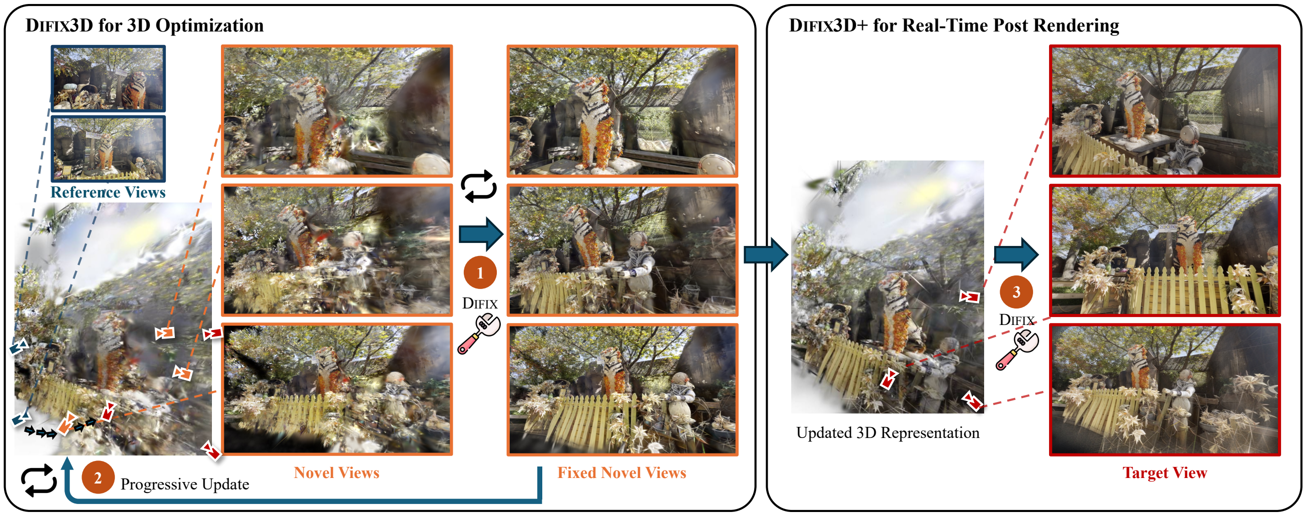
## Abstract

*Neural Radiance Fields and 3D Gaussian Splatting have revolutionized 3D reconstruction and novel-view synthesis task. However, achieving photorealistic rendering from extreme novel viewpoints remains challenging, as artifacts persist across representations. In this work, we introduce* DIFIX3D+, *a novel pipeline designed to enhance 3D reconstruction and novel-view synthesis through single-step diffusion models. At the core of our approach is* DIFIX, *a single-step image diffusion model trained to enhance and remove artifacts in rendered novel views caused by underconstrained regions of the 3D representation.* DIFIX *serves two critical roles in our pipeline. First, it is used during the reconstruction phase to clean up pseudo-training views that are rendered from the reconstruction and then distilled back into 3D. This greatly enhances underconstrained regions and improves the overall 3D representation quality. More importantly,* DIFIX *also acts as a neural enhancer during inference, effectively removing residual artifacts arising from imperfect 3D supervision and the limited capacity of current reconstruction models.* DIFIX3D+ *is a general solution, a single model compatible with both NeRF and 3DGS representations, and it achieves an average 2× improvement in FID score over baselines while maintaining 3D consistency.*

[*/†] Equal contribution/advising.

## 1. Introduction

Recent advances in neural rendering, particularly Neural Radiance Fields (NeRF) [37] and 3D Gaussian Splatting (3DGS) [20], represent an important step towards photore-




**Blue Cameras**: Training Views; **Red Cameras**: Target Views;
**Orange Cameras**: Intermediate Novel views along the progressive 3D updating trajectory (Sec. 4.2).

Figure 2. **DIFIX3D+ pipeline.** The overall pipeline of the DIFIX3D+ model involves the following stages: **Step 1**: Given a pretrained 3D representation, we render novel views and feed them to DIFIX which acts as a neural enhancer, removing the artifacts and improving the quality of the noisy rendered views (Sec. 4.1). The camera poses selected to render the novel views are obtained through pose interpolation, gradually approaching the target poses from the reference ones. **Step 2**: The cleaned novel views are distilled back to the 3D representation to improve its quality (Sec. 4.2). Steps 1 and 2 are applied in several iterations to progressively grow the spatial extent of the reconstruction and hence ensure strong conditioning of the diffusion model (DIFIX3D). **Step 3**: DIFIX additional acts as a real-time neural enhancer, further improving the quality of the rendered novel views.

alistic novel-view synthesis. However, despite their impressive performance near training camera views, these methods still suffer from artifacts such as spurious geometry and missing regions, especially when rendering less observed areas or more extreme novel views. The issue persists even for densely sampled captures collected under varying lighting conditions or with imperfect camera poses and calibration, hampering their suitability to real-world settings.

A core limitation of most NeRF and 3DGS approaches is their per-scene optimization framework, which requires carefully curated, view-consistent input data, and makes them susceptible to the *shape-radiance ambiguity* [86], where training images can be perfectly regenerated from a 3D representation that does not necessarily respect the underlying geometry of the scene. Without the data priors, these methods are also fundamentally limited in their ability to hallucinate plausible geometry and appearance in the underconstrained regions, and can only rely on the inherent smoothness of the underlying representation.

Unlike per-scene optimization based methods, large 2D generative models (e.g. diffusion models) are trained on internet-scale datasets, effectively learning the distribution of real-world images. Priors learned by these models generalize well to a wide range of scenes and use cases, and have been demonstrated to work on tasks such as inpainting [11, 64, 85] and outpainting [5, 62, 76]. However, the best way to lift these 2D priors to 3D remains unclear. Many contemporary methods query the diffusion model at each training step [25, 41, 72, 89]. These approaches primarily focus on optimizing object-centric scenes and scale poorly to larger environments with more expansive sets of possible camera trajectories [25, 41, 89]. Additionally, they are often time-consuming [72].

In this work, we tackle the challenge of using 2D diffusion priors to improve 3D reconstruction of large scenes in an efficient manner. To this end, we build upon recent advances in single-step diffusion [22, 32, 49, 77, 78], which greatly accelerate the inference speed of text-to-image generation. We show that these single-step models retain visual knowledge that can, with minimal fine-tuning, be adapted to "fix" artifacts present in NeRF/3DGS renderings. We use this fine-tuned model (DIFIX) during the reconstruction phase to generate pseudo-training views, which when distilled back into 3D, greatly enhance quality in underconstrained regions. Moreover, as the inference speed of these models is fast, we also directly apply DIFIX to the outputs of the improved reconstruction to further improve quality as a real-time post-processing step (DIFIX3D+).

We make the following contributions: **(i)** We show how to adapt 2D diffusion models to remove artifacts resulting from rendering a 3D neural representation, with minimal effort. The fine-tuning process takes only a few hours on a single consumer graphics card. Despite the short training time, the same model is powerful enough to remove artifacts in rendered images from both implicit representations such as NeRF and explicit representations like 3DGS. **(ii)** We propose an update pipeline that progressively refines the 3D representation by distilling back the improved novel views, thus ensuring multi-view consistency and significantly enhanced quality of the 3D representation. Compared to contemporary methods [26, 72] that query a diffusion model at each training time step, our approach is $>10\times$ faster. **(iii)** We demonstrate how single-step diffusion models enable near real-time post-processing that further improves novel view synthesis quality. **(iv)** We evaluate our approach across different datasets and present SoTA results, improving PSNR by $>$1dB and FID by $>2\times$ on average.

## 2. Related Work

The field of scene reconstruction and novel-view synthesis was revolutionized by the seminal NeRF [37] and 3DGS [20] works, which inspired a vast corpus of follow-up efforts. In the following, we discuss a non-exhaustive list of these approaches along axes relevant to our work.

**Improving 3D reconstruction discrepancies.** Most 3D reconstruction methods assume perfect input data, yet real-world captures often include slight inconsistencies that lead to artifacts and blurriness when distilled into a 3D representation. To address this, several methods improve NeRF's robustness to noisy camera inputs by optimizng camera poses [6, 21, 35, 39, 59, 69]. Other works focus on addressing lighting variations across images [34, 60, 73] and mitigating transient occlusions [48]. While these methods compensate for input data inconsistencies during training, they do not entirely eliminate them. This motivates our choice to apply our fixer also at render time, further improving quality in areas affected by these discrepancies (Sec. 4.2).

**Priors for novel view synthesis.** Numerous works address the limitations of NeRF and 3DGS in reconstructing under-observed scene regions. Geometric priors, introduced through regularization [38, 55, 75] or pretrained models that provide depth [7, 45, 63, 90] and normal [82] supervision, improve rendering quality in sparse-view settings. However, these methods are sensitive to noise, difficult to balance with data terms, and yield only marginal improvements in denser captures. Other works train feed-forward neural networks with posed multi-view data collected across numerous scenes. At render time, these approaches aggregate information from neighboring reference views to either enhance a previously rendered view [88] or directly predict a novel view [4, 31, 44, 79]. While these deterministic methods perform well near reference views, they often produce blurry results in ambiguous regions where the distribution of possible renderings is inherently multi-modal.

**Generative priors for novel view synthesis.** Recently, priors learned by the generative models have been increasingly used to enhance novel view synthesis. GANeRF [46] trains a per-scene generative adversarial network (GAN) that enhances NeRF's realism. Many other works use diffusion models that learn strong and generalizable priors from internet scale datasets. These diffusion models can either directly generate novel views with minimal fine-tuning [8, 13, 81, 83] or guide the optimization of a 3D representation. In the latter case, the diffusion model often serves as *scorer* that need to be queried during each optimization step [12, 25, 70, 72, 89], which significantly slows down training. In contrast, Deceptive-NeRF [27] and, concurrently with our work, 3D-Gaussian-Enhancer [28] use diffusion priors to enhance pseudo-observations rendered from the 3D representation, augmenting the training image set for fine-tuning the 3D representation. Since this approach avoids querying the diffusion model at every training step, the overhead is significantly reduced. While our work follows a similar direction, we diverge in two key aspects: (i) we introduce a progressive 3D update pipeline that effectively corrects artifacts even in extreme novel views while preserving long-range consistency and (ii) we use our model both during optimization and at render-time, leading to improved visual quality.

## 3. Background

**3D Scene Reconstruction and Novel-View Synthesis.** Neural Radiance Fields (NeRFs) have transformed the field of novel-view synthesis by modeling scenes as an emissive volume encoded within the weights of a coordinate-based multilayer perceptron (MLP). This MLP can be queried at any spatial location to return the view-dependent radiance $\boldsymbol{c} \in \mathbb{R}^3$ and volume density $\sigma \in \mathbb{R}$. The color of a ray $\mathbf{r}(\tau) = \boldsymbol{o} + t\boldsymbol{d}$ with origin $\boldsymbol{o} \in \mathbb{R}^3$ and direction $\boldsymbol{d} \in \mathbb{R}^3$ can then be rendered from the above representation by sampling points along the ray and accumulating their radiance through volume rendering as:

$$\mathcal{C}(\mathbf{p}) = \sum_{i=1}^{N} \alpha_i \mathbf{c}_i \prod_{j}^{i-1} (1 - \alpha_i) \tag{1}$$

where $\alpha_i = (1 - \exp(-\alpha_i \delta_i))$, $N$ denotes the number of samples along the ray, and $\delta_i$ is the step size used for quadrature.

Instead of representing scenes as a continuous neural field, 3D Gaussian Splatting [20] uses volumetric particles parameterized by their positions $\boldsymbol{\mu} \in \mathbb{R}^3$, rotation $\mathbf{r} \in \mathbb{R}^4$, scale $\mathbf{s} \in \mathbb{R}^3$, opacity $\eta \in \mathbb{R}$ and color $\mathbf{c}_i$. Novel views can be rendered from this representation using the same volume rendering formulation from Eq. (1), where

$$\alpha_i = \eta_i \exp\left[-\frac{1}{2}\left(\mathbf{p} - \boldsymbol{\mu}_i\right)^\top \boldsymbol{\Sigma}_i^{-1} \left(\mathbf{p} - \boldsymbol{\mu}_i\right)\right] \tag{2}$$

with $\boldsymbol{\Sigma} = \boldsymbol{R}\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{R}^T$ and $\mathbf{R} \in \mathrm{SO}(3)$ and $\mathbf{S} \in \mathbb{R}^{3\times 3}$ are the matrix representation of $\boldsymbol{r}$ and $\boldsymbol{s}$, respectively. The number $N$ of Gaussians that contribute to each pixel is determined through tile-base rasterization.

**Diffusion Models.** DMs [16, 54, 57] learn to model the data distribution $p_{\text{data}}(\mathbf{x})$ through *iterative denoising* and are trained with *denoising score matching* [16, 18, 33, 54, 56, 57, 61]. Specifically, to train a diffusion model, *diffused* versions $\mathbf{x}_\tau = \alpha_\tau \mathbf{x} + \sigma_\tau \boldsymbol{\epsilon}$ of the data $\mathbf{x} \sim p_{\text{data}}$ are generated, by progressively adding Gaussian noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$. Learnable parameters $\boldsymbol{\theta}$ of the denoiser model $\mathbf{F}_{\boldsymbol{\theta}}$ are optimized using the denoising score matching objective:

$$\mathbb{E}_{\mathbf{x}\sim p_{\text{data}},\tau\sim p_\tau,\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\boldsymbol{I})}\left[\|\mathbf{y} - \mathbf{F}_\theta(\mathbf{x}_\tau;\mathbf{c},\tau)\|_2^2\right], \quad (3)$$

where $\mathbf{c}$ represents optional conditioning information, such as a text prompt or image context. Depending on the model formulation, the target vector $\mathbf{y}$ is usually set as the added noise $\boldsymbol{\epsilon}$. Finally, $p_\tau$ denotes a uniform distribution over the diffusion time variable $\tau$. In practice a fixed discretization can be used [16]. In this setting, $p_\tau$ is often chosen as a uniform distribution, $p_\tau \sim \mathcal{U}(0, 1000)$. The maximum diffusion time $\tau = 1000$ is generally set such that the input data is fully transformed into Gaussian noise.

# 4. Boosting 3D Reconstruction with DM priors

Given a collection of RGB images and corresponding camera poses, our goal is to reconstruct a 3D representation that enables realistic novel view synthesis from arbitrary viewpoints, with particular emphasis on underconstrained regions distant from the input camera positions. To achieve this, we leverage the strong generative priors of a pretrained diffusion model during: **(i) optimization** to iteratively augment the training set with clean pseudo-views that improve the underlying 3D representation in distant and unobserved areas, and **(ii) inference** as a real-time post-processing step that further reduces artifacts caused by insufficient or inconsistent training supervision.

We first describe how to adapt a pretrained diffusion model into an image-to-image translation model that removes artifacts present in neural rendering methods (Sec. 4.1) and the data curation strategy used to fine-tune this model (Sec. 4.1.1). We then show how to use our fine-tuned diffusion model to improve the novel view synthesis quality of 3D representations in Sec. 4.2.

We visualize the overall DIFIX3D+ pipeline in Fig. 2 and the architecture of our DIFIX diffusion model in Fig. 3.

## 4.1. DIFIX: From a Pretrained Diffusion Model to a 3D Artifact Fixer

Given a rendered novel view $\tilde{I}$ that may contain artifacts from the 3D representation and a set of clean reference views $I_{\text{ref}}$, our model produces a refined novel view prediction $\hat{I}$. We build our model on top of a single-step diffusion model SD-Turbo [49], which has proven effective for image-to-image translation tasks [40], for efficiency reasons and to enable real-time post-processing during inference.

**Reference view conditioning.** We condition our model on a set of clean reference views $I_{\text{ref}}$, which in practice, we select as the closest training view. Inspired by video [1, 3, 9, 10, 13, 17, 53, 65, 66, 71, 84, 87] and multi-view diffusion models [24, 26, 29, 30, 42, 50, 51, 74], we adapt the self-attention layers into a *reference mixing layer* to capture cross-view dependencies. We start from concatenating novel view $\tilde{I}$ and reference views $I_{\text{ref}}$ on an additional view dimension and frame-wise encoded into latent space $\mathcal{E}((\tilde{I}, I_{\text{ref}})) = \mathbf{z} \in \mathbb{R}^{V\times C\times H\times W}$, where $C$ is the number of latent channels, $V$ is input number of views (reference views and target views) and $H$ and $W$ are the spatial latent dimensions. The *reference mixing layer* operates by first shifting the view axis to the spatial axis and reshaping back after the self-attention operation as follows (using `einops` [47] notation):

$$\mathbf{z}' \leftarrow \texttt{rearrange}(\mathbf{z}, \texttt{b c v (hw)} \rightarrow \texttt{b c (vhw)})$$
$$\mathbf{z}' \leftarrow l^i_\phi(\mathbf{z}', \mathbf{z}')$$
$$\mathbf{z}' \leftarrow \texttt{rearrange}(\mathbf{z}', \texttt{b c (vhw)} \rightarrow \texttt{b c v (hw)}),$$

where $l^i_\phi$ is a self-attention layer applied over the `vhw` dimension. This design allows us to inherit all module weights from the original 2D self-attention. We found this adaptation effective for capturing key information (e.g., objects, color, texture) from reference views, especially when the quality of the original novel view is severely degraded.

**Fine-tuning.** We fine-tune SD-Turbo [49] in a similar manner to Pix2pix-Turbo [40], using a frozen VAE encoder and a LoRA fine-tuned decoder. As in Image2Image-Turbo [40], we train our model to directly take the degraded rendered image $\tilde{I}$ as input, rather than random Gaussian noise, but apply a lower noise level ($\tau = 200$ instead of $\tau = 1000$). Our key insight is that the distribution of images degraded by neural rendering artifacts $\tilde{I}$ resembles the distribution of images $\mathbf{x}_\tau$ originally used to train the diffusion model at a specific noise level $\tau$ (Sec. 3). We validate this intuition by performing single-step "denoising" of rendered NeRF/3DGS images with artifacts, using a pre-trained SD-Turbo model. As shown in Fig. 4, $\tau = 200$ achieves the best results both visually and in terms of metrics.

**Losses.** We supervise our diffusion model with losses derived from readily available 2D supervision. We use the L2 difference between the model output $\hat{I}$ and the ground-truth image $I$ along with a perceptual LPIPS loss (as described in the supplement) in addition to a style loss term which encourages sharper details. We do so via a Gram matrix loss that defined as the L2 norm of the auto-correlation of VGG-16 features [43]:

$$\mathcal{L}_{\text{Gram}} = \frac{1}{L}\sum_{l=1}^{L}\beta_l \left\| G_l(\hat{I}) - G_l(I) \right\|_2, \quad (4)$$

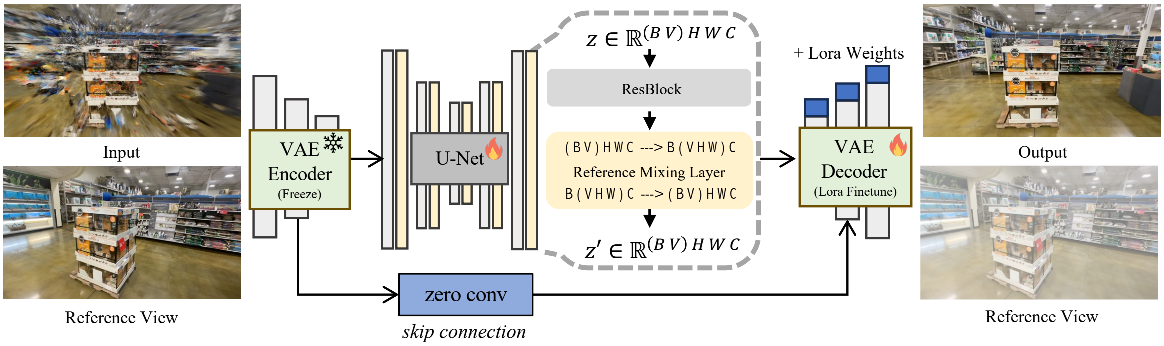



Figure 3. **DIFIX architecture.** DIFIX takes a noisy rendered image and a reference views as input (*left*), and outputs an enhanced version of the input image with reduced artifacts (*right*). DIFIX also generates identical reference views, which we discard in practice and hence depict transparent. The model architecture consists of a U-Net structure with a cross-view reference mixing layer (Sec. 4.1) to maintain consistency across reference views. DIFIX is fine-tuned from SD-Turbo, using a frozen VAE encoder and a LoRA fine-tuned decoder.


| $\tau$ | 1000 | 800 | 600 | 400 | 200 | 10 |
|---|---|---|---|---|---|---|
| PSNR | 12.18 | 13.63 | 15.64 | 17.05 | **17.73** | 17.72 |
| SSIM | 0.4521 | 0.5263 | 0.6129 | 0.6618 | **0.6814** | 0.6752 |

Figure 4. **Noise level.** To validate our hypothesis that the distribution of images with NeRF/3DGS artifacts is similar to the distribution of noisy images used to train SD-Turbo [49], we perform single-step "denoising" at varying noise levels. At higher noise levels (e.g., $\tau = 600$), the model effectively removes artifacts but also alters the image context. At lower noise levels (e.g., $\tau = 10$), the model makes only minor adjustments, leaving most artifacts intact. $\tau = 200$ strikes a good balance, removing artifacts while preserving context, and achieves the highest metrics.

with the Gram matrix at layer $l$ defined as:

$$G_l(I) = \phi_l(I)^\top \phi_l(I). \tag{5}$$

The final loss used to train our model is the weighted sum of the above terms: $\mathcal{L} = \mathcal{L}_{\text{Recon}} + \mathcal{L}_{\text{LPIPS}} + 0.5\mathcal{L}_{\text{Gram}}$.

#### 4.1.1 Data Curation

To supervise our model with the above loss terms, we require access to a large dataset consisting of pairs of images containing artifacts typical in novel-view synthesis and the corresponding "clean" ground truth images. A seemingly straightforward strategy would be to train a 3D representation with every $n$th frame and pair the remaining ground truth images with the rendered "novel" views. This **sparse reconstruction** strategy works well on the DL3DV dataset [23], which contains camera trajectories that allow us to sample novel views with significant deviation. However, it is suboptimal in most other novel view synthesis datasets [2, 36] where even held-out views largely observe the same region as the training views [70]. We therefore

| | Sparse Reconstruction | Cycle Reconstruction | Cross Reference | Model Underfitting |
|---|---|---|---|---|
| DL3DV [23] | ✓ | | | ✓ |
| Internal RDS | | ✓ | ✓ | ✓ |

Table 1. **Data curation.** We curate a paired dataset featuring common artifacts in novel-view synthesis. For DL3DV scenes [23], we employ sparse reconstruction and model underfitting, while for internal real driving scene (RDS) data, we utilize cycle reconstruction, cross reference, and model underfitting techniques.

explore various strategies to increase the amount of training examples (Tab. 1) :

**Cycle Reconstruction.** In nearly linear trajectories, such as those found in autonomous driving datasets, we first train a NeRF on the original path, and then render views from a trajectory shifted 1-6 meters horizontally (which we found to work well empirically). We then train a second NeRF representation against these rendered views and use this second NeRF to render degraded views for the original camera trajectory (for which we have ground truth).

**Model Underfitting.** To generate more salient artifacts than those obtained by merely holding out views, we underfit our reconstruction by training it with a reduced number of epochs (25%-75% of the original training schedule). We then render views from this underfitted reconstruction and pair them with the corresponding ground truth images.

**Cross Reference.** For multi-camera datasets, we train the reconstruction model solely with one camera and render images from the remaining held out cameras. We ensure visual consistency by selecting cameras with similar ISP.

### 4.2. DIFIX3D+: NVS with Diffusion Priors

Our trained diffusion model can be directly applied to enhance rendered novel views during inference (see (a) in Tab. 4). However, due to the generative nature of the model, this results in inconsistencies across different poses/frames, especially in under-observed and noisy regions where our model needs to hallucinate high-frequency details or even larger areas. An example is shown in Fig. 8, where the first column displays the NeRF result. Directly using DIFIX to

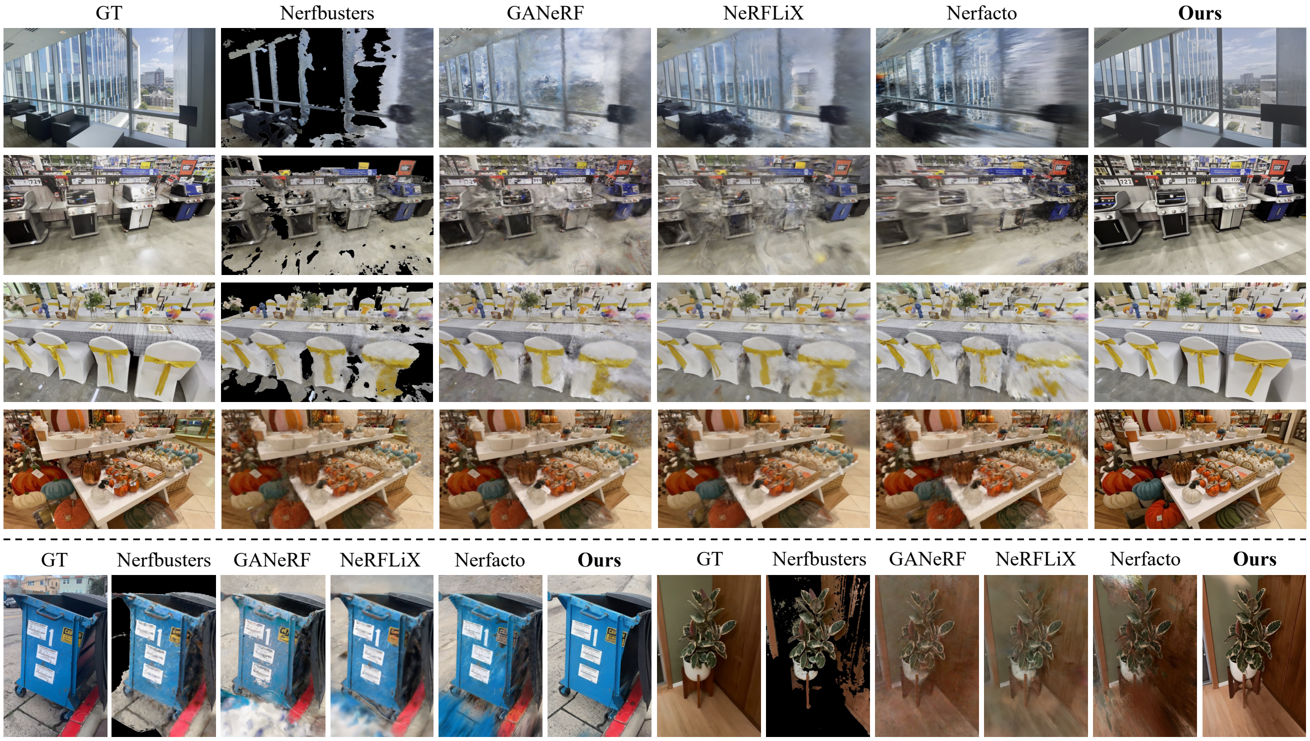



Figure 5. **In-the-wild artifact removal.** We show comparisons on held-out scenes from the DL3DV dataset [23] (*top*, above the dashed line) and the Nerfbusters [70] dataset (*bottom*). DIFIX3D+ corrects significantly more artifacts that other methods.

correct this novel view leads to inconsistent fixes. To address this issue, we distill the outputs of our diffusion model back into the 3D representation during training. This not only improves the multi-view consistency, but also leads to higher perceptual quality of the rendered novel views (see (b-c) in Tab. 4). Furthermore, we apply a final neural enhancer step during rendering inference, effectively removing residual artifacts. (see (d) in Tab. 4).

**DIFIX3D: Progressive 3D updates.** Strong conditioning of our diffusion-model on the rendered novel views and the reference views is crucial for achieving multi-view consistency and high fidelity to the input views. When the desired novel trajectory is too far from the input views, the conditioning signal becomes weaker and the diffusion model is forced to hallucinate more. We therefore adopt an iterative training scheme similar to Instruct-NeRF2NeRF [14] that progressively grows the set of 3D cues that can be rendered (multi-view consistently) to novel views and hence increases the conditioning for the diffusion model.

Specifically, given a set of target views, we begin by optimizing the 3D representation using the reference views. After every 1.5k iterations, we slightly perturb the ground-truth camera poses toward the target views, render the resulting novel view, and refine the rendering using the diffusion model trained in Sec. 4.1. The refined images are then added to the training set for another 1.5k iteration of training. By progressively perturbing the camera poses, refining the novel views, and updating the training set, this approach gradually improves 3D consistency and ensures high-quality, artifact-free renderings at the target views.

This progressive process allows us to progressively increase the overlap of 3D cues between the reference and target views, ultimately achieving consistent, artifact-free renderings. *See Supplementary Material for additional details about 3D update training.*

**DIFIX3D+: With Real-time Post Render Processing.** Due to the slight multi-view inconsistencies of the enhanced novel views that we are distilling, and the limited capacity of reconstruction methods to represent sharp details, some regions remain blurry (The second last column in Fig. 8). To further enhance the novel views, we use our diffusion model as the final post-processing step at render time, resulting in improvement across all perceptual metrics ((d) in Tab. 4), while maintaining a high degree of consistency. Since DIFIX is a single-step model, the additional rendering time is only 76 ms on a NVIDIA A100 GPU, over 10× faster than standard diffusion models with multiple denoising steps.

## 5. Experiments

We first evaluate DIFIX3D+ on in-the-wild scenes against several baselines and show its ability to enhance both NeRF and 3DGS-based pipelines (Sec. 5.1). We further evaluate the generality of our solution by enhancing automotive scenes (Sec. 5.2). We ablate our design in Sec. 5.3.

| Method | Nerfbusters Dataset PSNR↑ | SSIM↑ | LPIPS↓ | FID↓ | DL3DV Dataset PSNR↑ | SSIM↑ | LPIPS↓ | FID↓ |
|---|---|---|---|---|---|---|---|---|
| Nerfbusters [70] | 17.72 | 0.6467 | 0.3521 | 116.83 | 17.45 | 0.6057 | 0.3702 | 96.61 |
| GANeRF [46] | 17.42 | 0.6113 | 0.3539 | 115.60 | 17.54 | 0.6099 | 0.3420 | 81.44 |
| NeRFLiX [88] | 17.91 | 0.6560 | 0.3458 | 113.59 | 17.56 | 0.6104 | 0.3588 | 80.65 |
| Nerfacto [58] | 17.29 | 0.6214 | 0.4021 | 134.65 | 17.16 | 0.5805 | 0.4303 | 112.30 |
| DIFIX3D (Nerfacto) | 18.08 | 0.6533 | 0.3277 | 63.77 | 17.80 | 0.5964 | 0.3271 | 50.79 |
| DIFIX3D+ (Nerfacto) | **18.32** | **0.6623** | **0.2789** | **49.44** | **17.82** | **0.6127** | **0.2828** | **41.77** |
| 3DGS [20] | 17.66 | 0.6780 | 0.3265 | 113.84 | 17.18 | 0.5877 | 0.3835 | 107.23 |
| DIFIX3D (3DGS) | 18.14 | 0.6821 | 0.2836 | 51.34 | 17.80 | 0.5983 | 0.3142 | 50.45 |
| DIFIX3D+ (3DGS) | **18.51** | **0.6858** | **0.2637** | **41.77** | **17.99** | **0.6015** | **0.2932** | **40.86** |

Table 2. **Quantitative comparison on Nerfbusters and DL3DV datasets**. The best result is highlighted in **bold**, and the second-best is underlined.

### 5.1. In-the-Wild Artifact Removal

**DIFIX training.** We train DIFIX on a random selection of 80% of scenes (112 out of a total of 140) from the DL3DV [23] benchmark dataset. We generate 80,000 noisy-clean image pairs using the dataset curation strategies listed in Tab. 1, and simulate NeRF and 3DGS-based artifacts in a 1:1 ratio.

**Evaluation protocol.** We evaluate DIFIX3D+ with Nerfacto [58] and 3DGS [20] backbones on the 28 held out scenes from the DL3DV [23] benchmark and the 12 captures in the Nerfbusters [70] dataset. We partition each scene into a set of reference views used during training and evaluate on the left-out target views. We generate these splits for DL3DV by partitioning frames into two clusters based on camera position, ensuring a substantial deviation between reference and target views. We select reference and target views in the Nerfbusters dataset following their recommended protocol [70].

**Baselines.** We compare our Nerfacto and 3DGS DIFIX3D+ variants to their base methods. We also compare to Nerfbusters [70], which uses a 3D diffusion model to remove artifacts from NeRF[1], GANeRF [46], which train per-scene GAN that is used to enhance the realism of the scene representation, and NeRFLiX [88], which aggregates information from nearby reference views at inference time to improve novel view synthesis quality. We use the gsplat library[2] for 3DGS-based experiments and the official implementation for all other methods and baselines.

**Metrics.** We calculate PSNR, SSIM [67], LPIPS [19] as well as FID score [15] on novel views. More details are available in the Supplementrary Material.

**Results.** We provide quantitative results in Tab. 2. Our method outperforms all comparison methods by a significant margin across all metrics. Both DIFIX3D+ variants reduce LPIPS by 0.1 and FID by almost 3× relative to their respective NeRF and 3DGS backbones, highlighting a significant improvement in perceptual quality and visual fidelity. Furthermore, DIFIX3D+ also enhances PSNR, a pixel-wise metric sensitive to color shifting, by about 1db, indicating that DIFIX3D+ maintains a high degree of fidelity with original views (Sec. 4.2). We provide qualitative examples in Fig. 5 that show how DIFIX3D+ corrects significantly more artifacts that other methods, and additional videos in the supplement to further illustrate how we maintain a high degree of consistency across rendered frames.


Figure 6. **Qualitative results on the RDS dataset.** DIFIX for RDS was trained on 40 scenes and 100,000 paired data samples.

| Method | PSNR↑ | SSIM↑ | LPIPS↓ | FID↓ |
|---|---|---|---|---|
| Nerfacto | 19.95 | 0.4930 | 0.5300 | 91.38 |
| Nerfacto + NeRFLiX | 20.44 | 0.5672 | 0.4686 | 116.28 |
| Nerfacto + DIFIX3D | 21.52 | 0.5700 | 0.4266 | 77.83 |
| Nerfacto + DIFIX3D+ | **21.75** | **0.5829** | **0.4016** | **73.08** |

Table 3. **Comparison of quantitative results on RDS dataset**. The best result is highlighted in bold.

### 5.2. Automotive Scene Enhancement

**DIFIX training.** We construct an in-house real driving scene (RDS) dataset. The automotive capture rig contains three cameras with 40 degree overlaps between each camera. We train DIFIX with 40 scenes and generate 100,000 image pairs using the augmentation strategies listed in Tab. 1.

**Evaluation protocol.** We evaluate DIFIX3D+ with a Nerfacto backbone on 20 scenes (none of which are used during

[1]Nerfbusters [70] uses a visibility map extracted from a NeRF model trained on a combination of training and evaluation views and remove pixels that fall outside of that visibility map. This results in missing regions in Fig. 5

[2]https://github.com/nerfstudio-project/gsplat

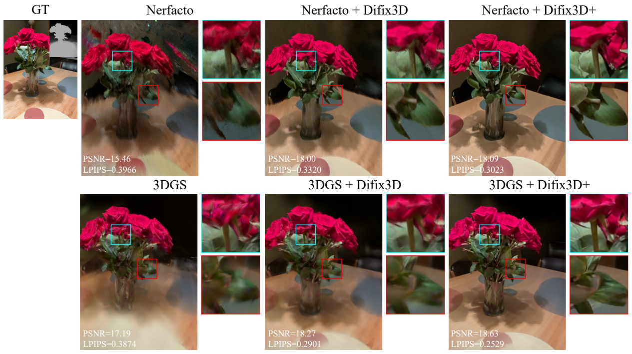



Figure 7. **Qualitative ablation of real-time post-render processing:** DIFIX3D+ uses an additional neural enhancer step that effectively removes residual artifacts, resulting in higher PSNR and lower LPIPS scores. The images displayed in green or red boxes correspond to zoomed-in views of the bounding boxes drawn in the main images.

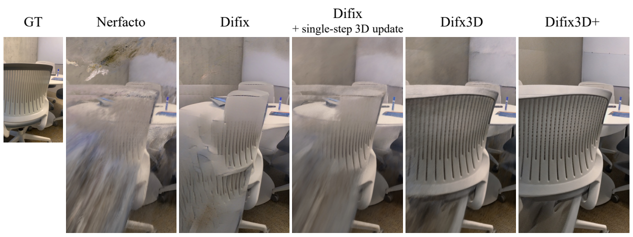



Figure 8. **Qualitative ablation results of DIFIX3D+**: The columns, labeled by method name, correspond to the rows in Tab. 4.

| Method | PSNR↑ | SSIM↑ | LPIPS↓ | FID↓ |
|---|---|---|---|---|
| Nerfacto | 17.29 | 0.6214 | 0.4021 | 134.65 |
| + (a) (**DIFIX**) | 17.40 | 0.6279 | 0.2996 | 49.87 |
| + (a) + (b) (DIFIX + single-step 3D update) | 17.97 | 0.6563 | 0.3424 | 75.94 |
| + (a) + (b) + (c) (**DIFIX3D**) | 18.08 | 0.6533 | 0.3277 | 63.77 |
| + (a) + (b) + (c) + (d) (**DIFIX3D+**) | **18.32** | **0.6623** | **0.2789** | **49.44** |

Table 4. **Ablation study of DIFIX3D+ on Nerfbusters dataset**. We compare a Nerfacto baseline to: (a) directly running DIFIX on rendered views without 3D updates, (b) distilling DIFIX outputs via 3D updates in a non-incremental manner, (c) applying the 3D updates incrementally, and (d) add DIFIX as a post-rendering step.

DIFIX training). We train NeRF with the center camera and evaluate the other two cameras as novel views.

**Baselines and metrics.** We compare our method to its NeRF baseline and NeRFliX [88]. We use the same evaluation metrics as in Sec. 5.1.

**Results.** Similar to Sec. 5.1, our method outperforms its baselines across all metrics (Tab. 3). Fig. 6 illustrates how our method reduces artifacts across views in a consistent manner.

## 5.3. Diagnostics

**Pipeline components.** We ablate our method by applying our pipeline components incrementally. We compare a Nerfacto baseline to: (a) directly running DIFIX on rendered views without 3D updates, (b) distilling DIFIX outputs via 3D updates in a non-incremental manner, (c) applying the 3D updates incrementally, and (d) add DIFIX as a post-rendering step. We show quantitative results in Tab. 4 averaged over the Nerfbusters [70] dataset. Qualitative ablation can be found in Fig. 8 and Fig. 7. Simply applying DIFIX to rendered outputs improves quality for renderings close to reference views but performs poorly in less observed regions, and causes flickering across rendered. Distilling diffusion outputs via 3D updates improves quality significantly but our incremental update strategy is essential, as evidenced by the degradation in LPIPS and FID when pseudo-views are added all at once. Visualization of post-rendering results is provided in Fig. 7, showcasing noticeable improvements in our outputs. These enhancements are further validated by the metric improvements shown in the last row of Tab. 4.

| Method | $\tau$ | SD Turbo Pretrain. | Gram | Ref | LPIPS↓ | FID↓ |
|---|---|---|---|---|---|---|
| pix2pix-Turbo | 1000 | ✓ | | | 0.3810 | 108.86 |
| DIFIX | 200 | ✓ | | | 0.3190 | 61.80 |
| DIFIX | 200 | ✓ | ✓ | | 0.3064 | 55.45 |
| DIFIX | 200 | ✓ | ✓ | ✓ | **0.2996** | **47.87** |

Table 5. **Ablation study of DIFIX components on Nerfbusters dataset**. Reducing the noise level, conditioning on reference views, and incorporating Gram loss improve our model.

**DIFIX training.** We validate our DIFIX training strategy by comparing to pix2pix-Turbo [40], which uses the same SD-Turbo backbone with a higher noise value ($\tau = 1000$ instead of $\tau = 200$) and to variants of our methods that omit reference view conditioning and Gram loss. Tab. 5 summarizes our results averages over the Nerfbusters dataset. Conditioning on reference views and with Gram loss further improves the result of our model. We note that simply decreasing the noise level from 1000 to 200 noticeably improves LPIPS and FID significantly, validating our findings in Fig. 4. The primary reason is that high noise level causes the model to generate more hallucinated pixels that contradict the ground truth, resulting in poorer generalization on the test dataset. *See Supplementary Material for visual examples.*

## 6. Conclusion

We introduced DIFIX3D+, a novel pipeline for enhancing 3D reconstruction and novel-view synthesis. At its core is DIFIX, a single-step diffusion model that can operate at near real time on modern NVIDIA GPUs. DIFIX improves 3D representation quality through a progressive 3D update scheme and enables real-time artifact removal during inference. Compatible with both NeRF and 3DGS, it achieves a $2\times$ improvement in FID scores over baselines while maintaining 3D consistency, showcasing its effectiveness in addressing artifacts and enhancing photorealistic rendering.

## References


[1] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025. 4

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022. 5

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 4

[4] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*, pages 14124–14133, 2021. 3

[5] Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wenzhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. Follow-your-canvas: Higher-resolution video outpainting with extensive content generation. *arXiv preprint arXiv:2409.01055*, 2024. 2

[6] Yu Chen and Gim Hee Lee. Dbarf: Deep bundle-adjusting generalizable neural radiance fields. In *CVPR*, pages 24–34, 2023. 3

[7] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, pages 12882–12891, 2022. 3

[8] Ruiqi Gao, Aleksander Holynski, Philipp Henzler, Arthur Brussee, Ricardo Martin-Brualla, Pratul Srinivasan, Jonathan T Barron, and Ben Poole. Cat3d: Create anything in 3d with multi-view diffusion models. *arXiv preprint arXiv:2405.10314*, 2024. 3

[9] Songwei Ge, Seungjun Nah, Guilin Liu, Tyler Poon, Andrew Tao, Bryan Catanzaro, David Jacobs, Jia-Bin Huang, Ming-Yu Liu, and Yogesh Balaji. Preserve Your Own Correlation: A Noise Prior for Video Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 4

[10] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu Video: Factorizing Text-to-Video Generation by Explicit Image Conditioning. *arXiv preprint arXiv:2311.10709*, 2023. 4

[11] Asya Grechka, Guillaume Couairon, and Matthieu Cord. Gradpaint: Gradient-guided inpainting with diffusion models. *Comput. Vis. Image Underst.*, 240(C), 2024. 2

[12] Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, pages 11808–11826. PMLR, 2023. 3

[13] 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, 2023. 3, 4

[14] Ayaan Haque, Matthew Tancik, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19740–19750, 2023. 6

[15] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 7, 13

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020. 4

[17] 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. *arXiv preprint arXiv:2210.02303*, 2022. 4

[18] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005. 4

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 694–711. Springer, 2016. 7, 13

[20] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 1, 3, 7

[21] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *ICCV*, pages 5741–5751, 2021. 3

[22] Shanchuan Lin, Anran Wang, and Xiao Yang. Sdxl-lightning: Progressive adversarial diffusion distillation. *arXiv preprint arXiv:2402.13929*, 2024. 2

[23] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22160–22169, 2024. 5, 6, 7

[24] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast Single Image to 3D Objects with Consistent Multi-View Generation and 3D Diffusion. *arXiv preprint arXiv:2311.07885*, 2023. 4

[25] 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 *ICCV*, pages 9298–9309, 2023. 2, 3

[26] 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 (ICCV)*, 2023. 3, 4

[27] Xinhang Liu, Jiaben Chen, Shiu-hong Kao, Yu-Wing Tai, and Chi-Keung Tang. Deceptive-nerf: Enhancing nerf reconstruction using pseudo-observations from diffusion models. *arXiv preprint arXiv:2305.15171*, 2023. 3

[28] Xi Liu, Chaoyi Zhou, and Siyu Huang. 3dgs-enhancer: Enhancing unbounded 3d gaussian splatting with view-consistent 2d diffusion priors. *arXiv preprint arXiv:2410.16266*, 2024. 3

[29] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. *arXiv preprint arXiv:2309.03453*, 2023. 4

[30] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3D: Single Image to 3D using Cross-Domain Diffusion. *arXiv preprint arXiv:2310.15008*, 2023. 4

[31] Yifan Lu, Xuanchi Ren, Jiawei Yang, Tianchang Shen, Zhangjie Wu, Jun Gao, Yue Wang, Siheng Chen, Mike Chen, Sanja Fidler, et al. Infinicube: Unbounded and controllable dynamic 3d driving scene generation with world-guided video models. *arXiv preprint arXiv:2412.03934*, 2024. 3

[32] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv preprint arXiv:2310.04378*, 2023. 2

[33] Siwei Lyu. Interpretation and generalization of score matching. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, page 359–366, Arlington, Virginia, USA, 2009. AUAI Press. 4

[34] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, pages 7210–7219, 2021. 3

[35] Andreas Meuleman, Yu-Lun Liu, Chen Gao, Jia-Bin Huang, Changil Kim, Min H Kim, and Johannes Kopf. Progressively optimized local radiance fields for robust view synthesis. In *CVPR*, pages 16539–16548, 2023. 3

[36] 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. 5

[37] 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. 1, 3

[38] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, 2022. 3

[39] Keunhong Park, Philipp Henzler, Ben Mildenhall, Jonathan T Barron, and Ricardo Martin-Brualla. Camp: Camera preconditioning for neural radiance fields. *ACM Transactions on Graphics (TOG)*, 42(6):1–11, 2023. 3

[40] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv preprint arXiv:2403.12036*, 2024. 4, 8, 14

[41] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *ICLR*, 2023. 2

[42] Guocheng Qian, Jinjie Mai, Abdullah Hamdi, Jian Ren, Aliaksandr Siarohin, Bing Li, Hsin-Ying Lee, Ivan Skorokhodov, Peter Wonka, Sergey Tulyakov, and Bernard Ghanem. Magic123: One Image to High-Quality 3D Object Generation Using Both 2D and 3D Diffusion Priors. *arXiv preprint arXiv:2306.17843*, 2023. 4

[43] Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. Film: Frame interpolation for large motion. In *European Conference on Computer Vision (ECCV)*, 2022. 4, 13

[44] Xuanchi Ren, Yifan Lu, Hanxue Liang, Zhangjie Wu, Huan Ling, Mike Chen, Sanja Fidler, Francis Williams, and Jiahui Huang. Scube: Instant large-scale scene reconstruction using voxsplats. *arXiv preprint arXiv:2410.20030*, 2024. 3

[45] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *CVPR*, pages 12892–12901, 2022. 3

[46] Barbara Roessle, Norman Müller, Lorenzo Porzi, Samuel Rota Bulò, Peter Kontschieder, and Matthias Nießner. Ganerf: Leveraging discriminators to optimize neural radiance fields. *ACM Transactions on Graphics (TOG)*, 42(6):1–14, 2023. 3, 7

[47] Alex Rogozhnikov. Einops: Clear and reliable tensor manipulations with einstein-like notation. In *International Conference on Learning Representations*, 2022. 4

[48] Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J Fleet, and Andrea Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. In *CVPR*, pages 20626–20636, 2023. 3

[49] Axel Sauer, Dominik Lorenz, Andreas Blattmann, and Robin Rombach. Adversarial diffusion distillation. In *European Conference on Computer Vision*, pages 87–103. Springer, 2025. 2, 4, 5

[50] 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. *arXiv preprint arXiv:2310.15110*, 2023. 4

[51] Yichun Shi, Peng Wang, Jianglong Ye, Mai Long, Kejie Li, and Xiao Yang. Mvdream: Multi-view diffusion for 3d generation. *arXiv preprint arXiv:2308.16512*, 2023. 4

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 14

[53] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-A-Video: Text-to-Video Generation without

Text-Video Data. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. 4

[54] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, 2015. 4

[55] Nagabhushan Somraj, Adithyan Karanayil, and Rajiv Soundararajan. SimpleNeRF: Regularizing sparse input neural radiance fields with simpler solutions. In *SIGGRAPH Asia*, 2023. 3

[56] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Proceedings of the 33rd Annual Conference on Neural Information Processing Systems*, 2019. 4

[57] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 4

[58] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, et al. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, pages 1–12, 2023. 7

[59] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. In *CVPR*, pages 4190–4200, 2023. 3

[60] Haithem Turki, Jason Y Zhang, Francesco Ferroni, and Deva Ramanan. Suds: Scalable urban dynamic scenes. In *CVPR*, 2023. 3

[61] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 23(7):1661–1674, 2011. 4

[62] Fu-Yun Wang, Xiaoshi Wu, Zhaoyang Huang, Xiaoyu Shi, Dazhong Shen, Guanglu Song, Yu Liu, and Hongsheng Li. Be-your-outpainter: Mastering video outpainting through input-specific adaptation. *arXiv preprint arXiv:2403.13745*, 2024. 2

[63] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *ICCV*, pages 9065–9076, 2023. 3

[64] Su Wang, Chitwan Saharia, Ceslee Montgomery, Jordi Pont-Tuset, Shai Noy, Stefano Pellegrini, Yasumasa Onoe, Sarah Laszlo, David J Fleet, Radu Soricut, et al. Imagen editor and editbench: Advancing and evaluating text-guided image inpainting. In *CVPR*, pages 18359–18369, 2023. 2

[65] Wenjing Wang, Huan Yang, Zixi Tuo, Huiguo He, Junchen Zhu, Jianlong Fu, and Jiaying Liu. VideoFactory: Swap Attention in Spatiotemporal Diffusions for Text-to-Video Generation. *arXiv preprint arXiv:2305.10874*, 2023. 4

[66] Yaohui Wang, Xinyuan Chen, Xin Ma, Shangchen Zhou, Ziqi Huang, Yi Wang, Ceyuan Yang, Yinan He, Jiashuo Yu, Peiqing Yang, Yuwei Guo, Tianxing Wu, Chenyang Si, Yuming Jiang, Cunjian Chen, Chen Change Loy, Bo Dai, Dahua Lin, Yu Qiao, and Ziwei Liu. LAVIE: High-Quality Video Generation with Cascaded Latent Diffusion Models. *arXiv preprint arXiv:2309.15103*, 2023. 4

[67] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4): 600–612, 2004. 7

[68] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 13

[69] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 3

[70] 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*, pages 18120–18130, 2023. 3, 5, 6, 7, 8, 13

[71] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 4

[72] 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. 2, 3

[73] Linning Xu, Vasu Agrawal, William Laney, Tony Garcia, Aayush Bansal, Changil Kim, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Aljaž Božič, Dahua Lin, Michael Zollhöfer, and Christian Richardt. VR-NeRF: High-fidelity virtualized walkable spaces. In *SIGGRAPH Asia*, 2023. 3

[74] Yinghao Xu, Hao Tan, Fujun Luan, Sai Bi, Peng Wang, Jiahao Li, Zifan Shi, Kalyan Sunkavalli, Gordon Wetzstein, Zexiang Xu, and Kai Zhang. DMV3D: Denoising Multi-View Diffusion using 3D Large Reconstruction Model. *arXiv preprint arXiv:2311.09217*, 2023. 4

[75] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *CVPR*, 2023. 3

[76] Jinze Yang, Haoran Wang, Zining Zhu, Chenglong Liu, Meng Wymond Wu, Zeke Xie, Zhong Ji, Jungong Han, and Mingming Sun. Vip: Versatile image outpainting empowered by multimodal large language model. In *ACCV*, 2024. 2

[77] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024. 2

[78] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024. 2

[79] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *CVPR*, 2021. 3

[80] Jason J Yu, Fereshteh Forghani, Konstantinos G Derpanis, and Marcus A Brubaker. Long-term photometric consistent novel view synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7094–7104, 2023. 14

[81] Wangbo Yu, Jinbo Xing, Li Yuan, Wenbo Hu, Xiaoyu Li, Zhipeng Huang, Xiangjun Gao, Tien-Tsin Wong, Ying Shan, and Yonghong Tian. Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. *arXiv preprint arXiv:2409.02048*, 2024. 3

[82] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. pages 25018–25032, 2022. 3

[83] David Junhao Zhang, Roni Paiss, Shiran Zada, Nikhil Karnad, David E Jacobs, Yael Pritch, Inbar Mosseri, Mike Zheng Shou, Neal Wadhwa, and Nataniel Ruiz. Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. *arXiv preprint arXiv:2411.05003*, 2024. 3

[84] David Junhao Zhang, Jay Zhangjie Wu, Jia-Wei Liu, Rui Zhao, Lingmin Ran, Yuchao Gu, Difei Gao, and Mike Zheng Shou. Show-1: Marrying pixel and latent diffusion models for text-to-video generation. *International Journal of Computer Vision*, pages 1–15, 2024. 4

[85] Guanhua Zhang, Jiabao Ji, Yang Zhang, Mo Yu, Tommi S Jaakkola, and Shiyu Chang. Towards coherent image inpainting using denoising diffusion implicit models. 2023. 2

[86] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2

[87] Daquan Zhou, Weimin Wang, Hanshu Yan, Weiwei Lv, Yizhe Zhu, and Jiashi Feng. MagicVideo: Efficient Video Generation With Latent Diffusion Models. *arXiv preprint arXiv:2211.11018*, 2023. 4

[88] Kun Zhou, Wenbo Li, Yi Wang, Tao Hu, Nianjuan Jiang, Xiaoguang Han, and Jiangbo Lu. Nerflix: High-quality neural view synthesis by learning a degradation-driven interviewpoint mixer. In *CVPR*, pages 12363–12374, 2023. 3, 7, 8

[89] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *CVPR*, 2023. 2, 3

[90] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. Fsgs: Real-time few-shot view synthesis using gaussian splatting. In *ECCV*, 2024. 3