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

Supplementary Material

We provide additional implementation details in Sec. A and further results in Sec. B. We discuss limitations and future work in Sec. C.

A. Additional Implementation Details

A.1. Loss Functions

We supervise our diffusion model with losses derived from readily available 2D supervision in the RGB image space, avoiding the need for any sort of 3D supervision that is hard to obtain:

- *Reconstruction loss.* Which we define as the L2 loss between the model output $\hat{I}$ and the ground-truth image I:

$$\mathcal{L}_{\text{Recon}} = \|\hat{I} - I\|_2. \tag{6}$$

- *Perceptual loss.* We incorporate an LPIPS [19] loss based on the L1 norm of the VGG-16 features $\phi_l(\cdot)$ to enhance image details, defined as:

$$\mathcal{L}_{\text{LPIPS}} = \frac{1}{L}\sum_{l=1}^{L}\alpha_l \left\|\phi_l(\hat{I}) - \phi_l(I)\right\|_1, \tag{7}$$

- *Style loss.* We use the Gram matrix loss based on VGG-16 features [43] to obtain sharper details. We define the loss as the L2 norm of the auto-correlation of VGG-16 features [43]:

$$\mathcal{L}_{\text{Gram}} = \frac{1}{L}\sum_{l=1}^{L}\beta_l \left\|G_l(\hat{I}) - G_l(I)\right\|_2, \tag{8}$$

  with the Gram matrix at layer l defined as:

$$G_l(I) = \phi_l(I)^\top \phi_l(I). \tag{9}$$

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

A.2. Progressive 3D updates

Please refer to the pseudocode in Algorithm 1 for further details.

A.3. Evaluation Metrics

We employ several evaluation metrics to quantitatively assess the model's performance in novel view synthesis. These metrics include Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [68], Learned Perceptual Image Patch Similarity (LPIPS) [19], and Fréchet Inception Distance (FID) [15]. Following the evaluation procedure outlined by Nerfbusters [70], we calculate a visibility map and mask out the invisible regions when computing the metrics.

Algorithm 1: Progressive 3D Updates for Novel View Rendering

```
Input: Reference views $V_{\text{ref}}$, Target views $V_{\text{target}}$, 3D
       representation $R$ (e.g., NeRF, 3DGS),
       Diffusion model $D$ (DIFIX), Number of
       iterations per refinement $N_{\text{iter}}$, Perturbation
       step size $\Delta_{\text{pose}}$
Output: High-quality, artifact-free renderings at
        $V_{\text{target}}$
Initialize: Optimize 3D representation $R$ using $V_{\text{ref}}$.
while not converged do
    /* Optimize the 3D
       representation                        */
    for $i = 1$ to $N_{iter}$ do
        Optimize $R$ using the current training set.
    /* Generate novel views by
       perturbing camera poses               */
    for each $v \in V_{target}$ do
        Find the nearest camera pose of $v$ in the
          training set.
        Perturb the nearest camera pose by $\Delta_{\text{pose}}$.
        Render novel view $\hat{v}$ using $R$.
        Refine $\hat{v}$ using diffusion model $D$.
        Add refined view $\hat{v}$ to the training set.
return Refined renderings at $V_{\text{target}}$.
```

PSNR. The Peak Signal-to-Noise Ratio (PSNR) is widely used to measure the quality of reconstructed images by comparing them to ground truth images. It is defined as:

$$\text{PSNR} = 10 \cdot \log_{10}\left(\frac{\text{MAX}^2}{\text{MSE}}\right), \tag{10}$$

where MAX represents the maximum possible pixel value (e.g., 255 for 8-bit images), and MSE is the mean squared error between the predicted image I_{pred} and the ground truth image I_{gt}. Higher PSNR values indicate better reconstruction quality.

SSIM. The Structural Similarity Index (SSIM) evaluates the perceptual similarity between two images by considering luminance, contrast, and structure. It is computed as:

$$\text{SSIM}(I_{\text{pred}}, I_{\text{gt}}) = \frac{(2\mu_{\text{pred}}\mu_{\text{gt}} + C_1)(2\sigma_{\text{pred,gt}} + C_2)}{(\mu_{\text{pred}}^2 + \mu_{\text{gt}}^2 + C_1)(\sigma_{\text{pred}}^2 + \sigma_{\text{gt}}^2 + C_2)}, \tag{11}$$

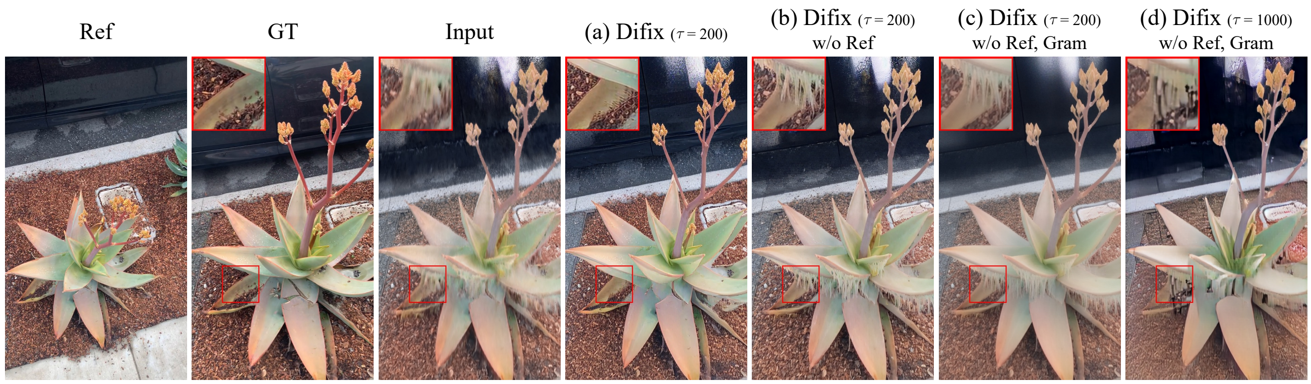
Ref
GT
Input
(a) Difix ($\tau = 200$)
(b) Difix ($\tau = 200$) w/o Ref
(c) Difix ($\tau = 200$) w/o Ref, Gram
(d) Difix ($\tau = 1000$) w/o Ref, Gram

Figure S1. **Visual comparison of DIFIX components.** Reducing the noise level τ ((c) *vs.* (d)), incorporating Gram loss ((b) *vs.* (c)), and conditioning on reference views ((a) *vs.* (b)) all improve our model.

where μ and σ^2 represent the mean and variance of the pixel intensities, respectively, and $\sigma_{\text{pred,gt}}$ is the covariance. The constants C_1 and C_2 stabilize the division to avoid numerical instability.

LPIPS. The Learned Perceptual Image Patch Similarity (LPIPS) metric evaluates the perceptual similarity between two images based on feature embeddings extracted from pre-trained neural networks. It is defined as:

$$\text{LPIPS}(I_{\text{pred}}, I_{\text{gt}}) = \sum_{l} \|\phi_l(I_{\text{pred}}) - \phi_l(I_{\text{gt}})\|_2^2, \qquad (12)$$

where ϕ_l represents the feature maps from the l-th layer of a pre-trained VGG-16 network [52]. Lower LPIPS values indicate greater perceptual similarity.

FID. The Fréchet Inception Distance (FID) measures the distributional similarity between generated images and real images in the feature space of a pre-trained Inception network. It is computed as:

$$\text{FID} = \|\mu_{\text{gen}} - \mu_{\text{real}}\|_2^2 + \text{Tr}(\Sigma_{\text{gen}} + \Sigma_{\text{real}} - 2(\Sigma_{\text{gen}}\Sigma_{\text{real}})^{\frac{1}{2}}), \qquad (13)$$

where $(\mu_{\text{gen}}, \Sigma_{\text{gen}})$ and $(\mu_{\text{real}}, \Sigma_{\text{real}})$ denote the means and covariances of the feature distributions for the generated and real images, respectively. Lower FID values indicate better alignment between the generated and real image distributions. We report the FID score calculated between the novel view renderings and the corresponding ground-truth images across the entire testing set.

A.4. Data Curation

To curate paired training data, we employ a range of strategies including sparse reconstruction, cycle reconstruction, cross-referencing, and intentional model underfitting. The curated paired data generated through these strategies is visualized in Fig. S2. The simulated corrupted images exhibit common artifacts observed in extreme novel views, such as blurred details, missing regions, ghosting structures, and spurious geometry. This curated dataset provides a robust learning signal for the DIFIX model, enabling the model to effectively correct artifacts in underconstrained novel views and enhance the quality of 3D reconstruction.

B. Additional Results

B.1. Ablation Study of DIFIX

In addition to the quantitative results presented in Tab. 5, we provide visual examples in Fig. S1 to demonstrate the effectiveness of our key design choices in DIFIX. Compared to using a high noise level (*e.g.*, pix2pix-Turbo [40]), reducing the noise level significantly removes artifacts and improves overall visual quality ((c) *vs.* (d)). Incorporating Gram loss enhances fine details and sharpens the image ((b) *vs.* (c)). Furthermore, conditioning on a reference view corrects structural inaccuracies and alleviates color shifts ((a) *vs.* (b)). Together, these advancements culminate in the superior results achieved by DIFIX.

B.2. Evaluation of Multi-View Consistency

We evaluate our model using the Thresholded Symmetric Epipolar Distance (TSED) metric [80], which quantifies the number of consistent frame pairs in a sequence. As shown in Tab. S1, our model achieves higher TSED scores than reconstruction-based methods (*e.g.*, Nerfacto) and other baselines, demonstrating superior multi-view consistency in novel view synthesis. Notably, the final post-processing step (DIFIX3D+) enhances image sharpness without compromising 3D coherence.

Method	Nerfacto	NeRFLiX	GANeRF	DIFIX3D	DIFIX3D+
TSED ($T_{error} = 2$)	0.2492	0.2532	0.2399	0.2601	**0.2654**
TSED ($T_{error} = 4$)	0.5318	0.5276	0.5140	0.5462	**0.5515**
TSED ($T_{error} = 8$)	0.7865	0.7789	0.7844	**0.7924**	0.7880

Table S1. **Multi-view consistency evaluation on the DL3DV dataset.** A higher TSED score indicates better multi-view consistency.

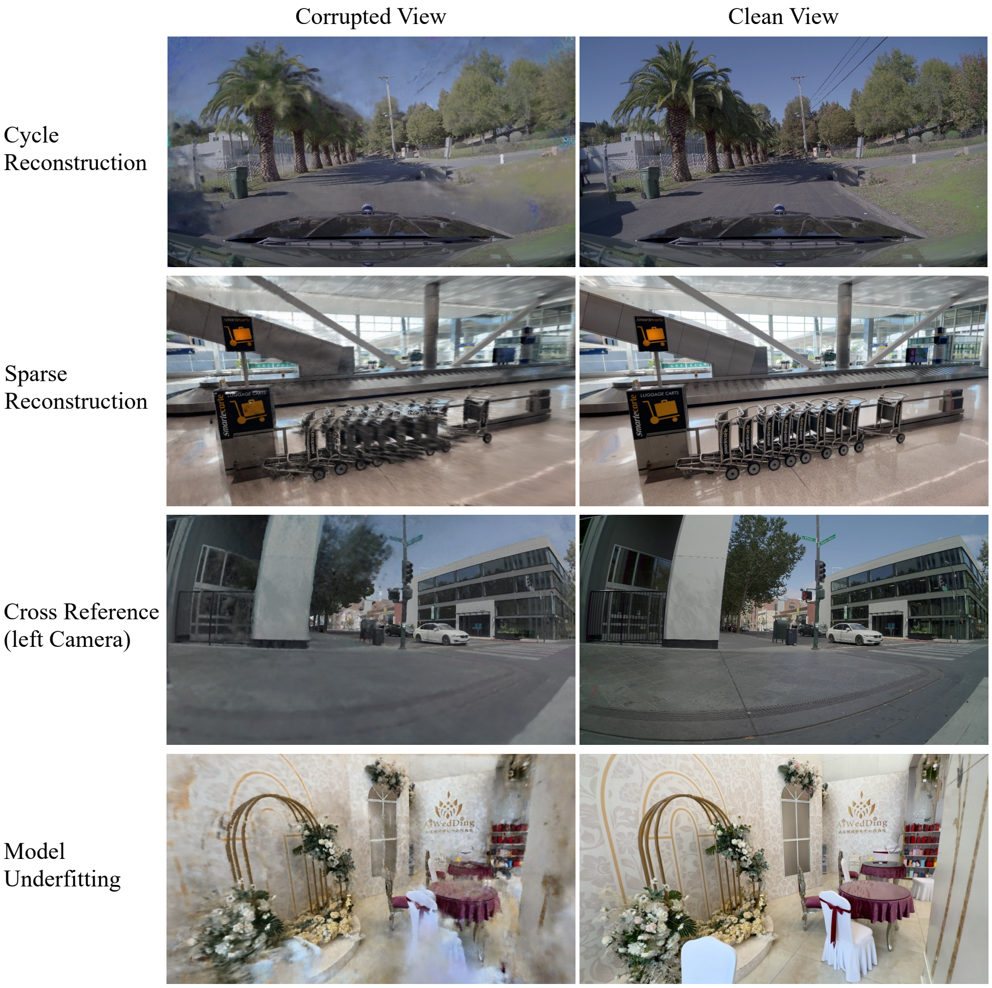

Figure S2. **Visualization of the paired dataset:** We utilize a variety of strategies to simulate corrupted training data, including sparse reconstruction, cycle reconstruction, cross-referencing, and intentional model underfitting. The curated paired dataset provides a strong learning signal for the DIFIX model.

C. Limitation and Future Work

We present DIFIX3D+, a novel pipeline designed to advance 3D reconstruction and novel-view synthesis. However, as a 3D enhancement model, the performance of DIFIX3D+ is inherently limited by the quality of the initial 3D reconstruction. It currently struggles to enhance views where 3D reconstruction has entirely failed. Addressing this limitation through the integration of modern diffusion model priors represents an exciting direction for future research. To prioritize speed and approach near real-time post-rendering processing, DIFIX is derived from a single-step image diffusion model. Additional promising avenues include scaling DIFIX to a single-step video diffusion model, enabling enhanced long-context 3D consistency.