DiSR-NeRF: Diffusion-Guided View-Consistent Super-Resolution NeRF

Jie Long Lee Chen Li Gim Hee Lee
Department of Computer Science, National University of Singapore
ljielong@comp.nus.edu.sg, lichen@u.nus.edu, gimhee.lee@nus.edu.sg

Abstract

We present DiSR-NeRF, a diffusion-guided framework for view-consistent super-resolution (SR) NeRF. Unlike prior works, we circumvent the requirement for high-resolution (HR) reference images by leveraging existing powerful 2D super-resolution models. Nonetheless, independent SR 2D images are often inconsistent across different views. We thus propose Iterative 3D Synchronization (I3DS) to mitigate the inconsistency problem via the inherent multi-view consistency property of NeRF. Specifically, our I3DS alternates between upscaling low-resolution (LR) rendered images with diffusion models, and updating the underlying 3D representation with standard NeRF training. We further introduce Renoised Score Distillation (RSD), a novel score-distillation objective for 2D image resolution. Our RSD combines features from ancestral sampling and Score Distillation Sampling (SDS) to generate sharp images that are also LR-consistent. Qualitative and quantitative results on both synthetic and real-world datasets demonstrate that our DiSR-NeRF can achieve better results on NeRF super-resolution compared with existing works. Code and video results available at the project website[1].

1. Introduction

Novel view synthesis is a long-standing problem in computer vision with significant real-world applications. Recently, neural radiance fields (NeRFs) have emerged as a powerful representation, achieving state-of-the-art performance in novel view synthesis. Since the pioneering work on NeRFs by [28], many follow up works have explored the improvement of NeRF's speed [4, 7, 31, 42], fidelity [1, 2, 49], scale [26, 43], robustness [6, 21, 55], and generalizability [3, 16, 19, 46, 52, 53, 56]. However, one important aspect of NeRFs that has not been well explored is super-resolution. In real-world scenarios, imaging devices may be limited in resolution (*i.e.*, drones, CCTVs, etc.) and consequently high-resolution multi-view images may be unavailable. With low-resolution inputs, NeRF struggles to represent the high-quality details of the underlying 3D scenes. In this work, we tackle the task of NeRF super-resolution, which aims to learn high-resolution implicit representation of 3D scenes from only low-resolution images. One possible direction is to design a generative 3D super-resolution model, which however, requires large datasets of high-resolution multi-view images for training. Collecting such large-scale, high-resolution multi-view data is labor-intensive and requires expensive equipment to obtain accurate scans. On the other hand, large 2D high-resolution image datasets such as LAION-5B [38] are publicly available and have been used to train powerful 2D super-resolution models. We thus propose to leverage knowledge from the 2D super-resolution models to circumvent the requirements for HR images.

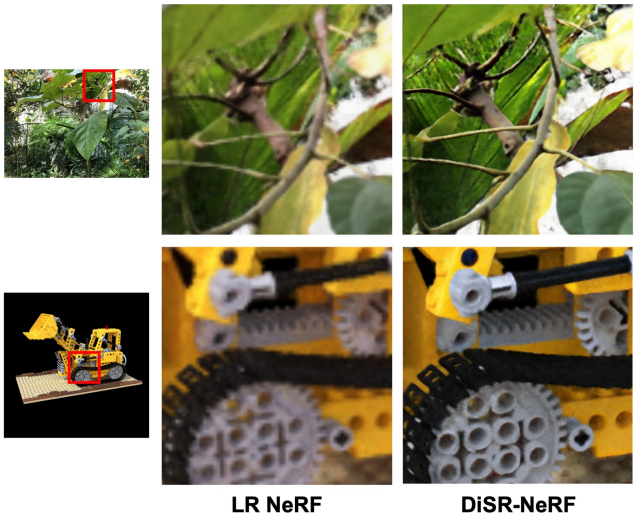

Figure 1. Our DiSR-NeRF distils super resolution priors from a 2D diffusion upscaler to generate high quality details from low resolution NeRFs.

Naively upscaling individual LR training images with 2D super-resolution methods produces SR images that may not be consistent across views. A NeRF trained on such images produces blurred details as SR details may not agree

[1] https://github.com/leejielong/DiSR-NeRF

across different camera views. Recent works such as Super-NeRF [8] resolve this by searching the latent space of a 2D upscaler for view-consistent SR results, however the framework only enforces low-resolution view-consistency. In this work, we propose DiSR-NeRF, a Diffusion-guided Super-Resolution NeRF method which produces NeRFs with high resolution and view-consistent details.

Our method comprises two key components. 1) We propose the two-stage **Iterative 3D Synchronization (I3DS)** to solve the cross-view inconsistency problem. We first refine the rendered images from the LR NeRF with a diffusion-based 2D super-resolution model, and subsequently synchronize the details into 3D through standard NeRF training. The alternating process between the two stages guides the NeRF to converge to view-consistent details. 2) We introduce the **Renoised Score Distillation (RSD)** objective to get the best from both worlds of ancestral sampling and Score Distillation Sampling (SDS). Particularly, we observe that the default ancestral sampling used in diffusion-based super resolution can generate details that are structurally inconsistent with the conditioned LR image (*cf*. Fig. 6 in the experiments). This may aggravate the inconsistency across different views. On the other hand, the Score Distillation Sampling (SDS) objective commonly used in Text-to-3D generation can produce LR-consistent features but with limited details. The optimization target of RSD is designed as the intermediate denoised latents of the ancestral sampling trajectory. This transforms the ancestral sampling process into an optimization framework, generating coarse-to-fine details over the course of optimization. As a result, RSD is able to achieve sharper details compared to SDS while also producing LR-consistent features compared to ancestral sampling.

Our method requires only low-resolution multi-view images of a target scene, thus alleviating the cumbersome need for high-resolution reference images or large scale multi-view HR image datasets. Our qualitative and quantitative results show that DiSR-NeRF can outperform existing baselines to achieve effective super-resolution NeRF. Our contributions are as follows:

- We introduce DiSR-NeRF, a method that achieves high quality super-resolution NeRF using only LR training images and a pretrained 2D diffusion upscaler.
- We propose Iterative 3D Synchronization (I3DS) that achieves convergence to view-consistent SR details.
- We design Renoised Score Distillation (RSD), a score-distillation objective that produces sharper details and maintains consistency towards the conditioned LR image.

2. Related Works

2.1. 2D Image Super-Resolution

Image super-resolution is inherently an ill-posed problem since there can be a distribution of possible SR solutions for a LR image. Most state-of-the-art approaches therefore seek to learn the conditional distribution $p(\mathbf{x} \mid \mathbf{y}, \mathbf{w})$ of possible SR images $\mathbf{x}$ that correspond to the conditioned LR image $\mathbf{y}$ using a generative model parameterized by $\mathbf{w}$.

GANs. Generative Adversarial Networks (GANs) are a class of generative models that have demonstrated impressive results in image SR. GAN-based SR models [17, 47, 48] utilize adversarial training objectives to produce SR images that appear photorealistic to human visual perception. However, GAN-based SR models are often difficult to train and may encounter mode collapse.

Normalizing Flows. Flow-based SR models [15, 20, 24, 54] employ conditional normalizing flows to model the conditional SR image distribution. Unlike GANs, flow-based SR models learn to explicitly compute the probability density of SR images conditioned on the LR image. However, flow-based SR models are restricted to invertible architectures due to bijectivity constraints and lack expressiveness compared to other approaches.

Diffusion Models. Diffusion-based SR models [11, 18, 36, 37, 51] learn to generate SR images from pure Gaussian noise using a trained denoising network to iteratively restore structure and details while being conditioned on a text prompt and LR image. In our work, we use the Stable Diffusion $\times 4$ Upscaler (SD$\times$4), which is a pretrained latent diffusion upscaler [36] to guide the generation of high-resolution details in 3D. Latent diffusion models use a pretrained variational autoencoder (VAE) to project images to a lower-dimensional latent space for diffusion. Executing the diffusion process in the latent space improves speed and GPU memory usage.

2.2. Super-Resolution NeRF

Super-Resolution NeRF is currently an area that has yet to be well-explored. Some works improve NeRF details via anti-aliasing or ray supersampling [1, 2, 44], but remain fundamentally limited by the level of detail available in the input images. Other works achieve super-resolution NeRF [13, 44] under a reference-guided setting, requiring HR reference images of the target scene to be available. Such requirements can be impractical when only low resolution imaging solutions are available.

Super-NeRF [8] is a recent work with similar motivation to ours, and it achieves high-resolution detail generation for LR NeRF by searching the latent space of ESRGAN [47] for view-consistent solutions. However, the proposed framework only explicitly constraints view consistency in the LR domain. In contrast to Super-NeRF, we utilize the

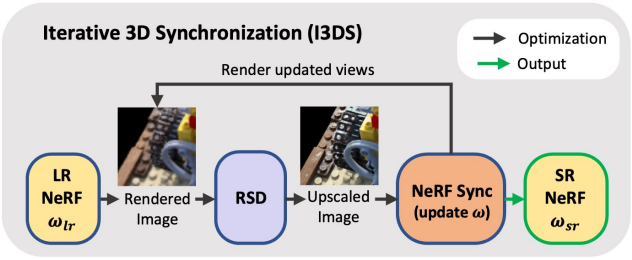

Figure 2. I3DS seperates upscaling and NeRF fitting in seperate, alternate stages. NeRF renders are upscaled via RSD in the upscaling stage, and upscaled images are used as training images to learn view-consistent details. The two stage process is repeated over several cycles to achieve detail convergence.

score function of diffusion-based upscaler models to produce sharp and coherent SR details. Furthermore, we also introduce a 3D synchronization mechanism to converge on view-consistent features.

3. Our Method

3.1. Preliminaries

Diffusion Models. Diffusion models [10, 34, 39–41] are generative models which transform a sample from a noise distribution towards a data distribution using a fixed forward process and a learned reverse process. The forward process introduces noise to data based on a predetermined noising schedule and gradually destroys detail and structure. A data sample $\mathbf{z}_0$ can be noised into $\mathbf{z}_t$ using the closed-form formula [10] of the forward process:

$$\mathbf{z}_t = \sqrt{\bar{\alpha}_t}\mathbf{z}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon} \tag{1}$$

which produces noised latents $\mathbf{z}_t$ at timestep $t \in [0, T]$ given timestep-dependent noising coefficient $\bar{\alpha}_t$ and Gaussian noise sample $\boldsymbol{\epsilon} \in \mathcal{N}(0, \mathbf{I})$.

The reverse process restores structure from noise. It is learned by a network $\boldsymbol{\phi}$ that is trained to denoise a noised latent $\mathbf{z}_t$ by predicting its noise component $\boldsymbol{\epsilon}_\phi(\mathbf{z}_t, y, t)$ conditioned upon y (text prompt/image) and timestep t. In diffusion image generators, y comprises the embedded text prompt, while in diffusion upscalers, a lightly noised LR image and the image noising level is also included. Diffusion models are typically trained with an evidence lower bound (ELBO) objective [10] given by:

$$\mathcal{L} = \mathbb{E}_{t\sim U(0,T),\epsilon\sim\mathcal{N}(\mathbf{0},\mathbf{I})}[\gamma(t)\|\boldsymbol{\epsilon}_\phi(\mathbf{z}_t, y, t) - \boldsymbol{\epsilon}\|_2^2]. \tag{2}$$

The sampling process which is also referred to as ancestral sampling, iteratively denoises a Gaussian noise sample $\mathbf{z}_T$ according to:

$$\mathbf{z}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{z}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_\phi(\mathbf{z}_t, y, t)\right) + \sigma_t\boldsymbol{\epsilon} \tag{3}$$

for a DDPM [10] scheduler, where σ_t is the standard deviation of Gaussian noise samples. In ancestral sampling, the latent $\mathbf{z}_t$ is gradually denoised into $\mathbf{z}_0$ from the data distribution via a sample trajectory given by:

$$(\mathbf{z}_T, \mathbf{z}_{T-1}, ..., \mathbf{z}_1, \mathbf{z}_0), \tag{4}$$

where $\mathbf{z}_0$ is the generated output image.

Score Distillation Sampling. Diffusion models are also score-based generative models [41, 45] which learn a score function $\nabla_\mathbf{z}\log p_t(\mathbf{z})$ as the gradient of the log probability density with respect to data. Score Distillation Sampling (SDS) [35, 45] is an optimization objective that uses the score function of pretrained diffusion models to provide gradients that guide the optimization of a differentiable image parametrization [30] $\mathbf{z} = g(\theta)$ towards the mode of a conditional probability distribution $p(\mathbf{z}_t \mid y)$. SDS has demonstrated remarkable results in Text-to-3D generation [5, 14, 22, 23, 25, 35, 45, 50, 58], but is known to encounter issues such as over-saturation and over-smoothing [50].

Given an image $\mathbf{z}_0 = g(\theta)$, a pretrained diffusion model predicts the noise component $\boldsymbol{\epsilon}(\mathbf{z}_t, y, t)$ which is related to the score function $\nabla_{\mathbf{z}_t}\log p(\mathbf{z}_t \mid y)$ by:

$$\boldsymbol{\epsilon}(\mathbf{z}_t, y, t) = -\sigma_t\nabla_{\mathbf{z}_t}\log p(\mathbf{z}_t \mid y). \tag{5}$$

Consequently, [35] proposes the SDS objective as:

$$\nabla_\theta\mathcal{L}_{SDS} = \mathbb{E}_{t,\epsilon}[\gamma(t)(\boldsymbol{\epsilon}_\phi(\mathbf{z}_t, y, t) - \boldsymbol{\epsilon})\frac{\partial\mathbf{z}_0}{\partial\theta}]. \tag{6}$$

In 3D, $\mathbf{z}_0 = g(\theta)$ is a NeRF render where $g(\cdot)$ corresponds to the volume rendering function and θ represents the NeRF parameters. In 2D, $g(\cdot)$ is an identity transform and θ is the pixel values of the image under optimization. In this paper, we use the 2D formulation of $g(\theta)$.

3.2. Iterative 3D Synchronization (I3DS)

In our initial experiments, we observed that directly applying SDS on NeRF renders produces blurred details and would fail to converge on detailed reconstructions. We postulate that this is because SDS supervision is only provided over a small local patch of rays in each training step, which does not optimize NeRF towards globally consistent features. Rendering multiple patches concurrently is also intractable due to significant GPU memory required. To resolve this issue, we propose Iterative 3D Synchronization (I3DS) which disentangles upscaling and NeRF synchronization by performing both processses in seperate alternate stages.

Fig. 2 illustrates our I3DS framework. The first stage is the **upscaling-stage**. Starting from a NeRF $\boldsymbol{\omega}_{lr}$ pretrained from low resolution inputs, we render images in 4× resolution from all training poses. Each rendered image is then

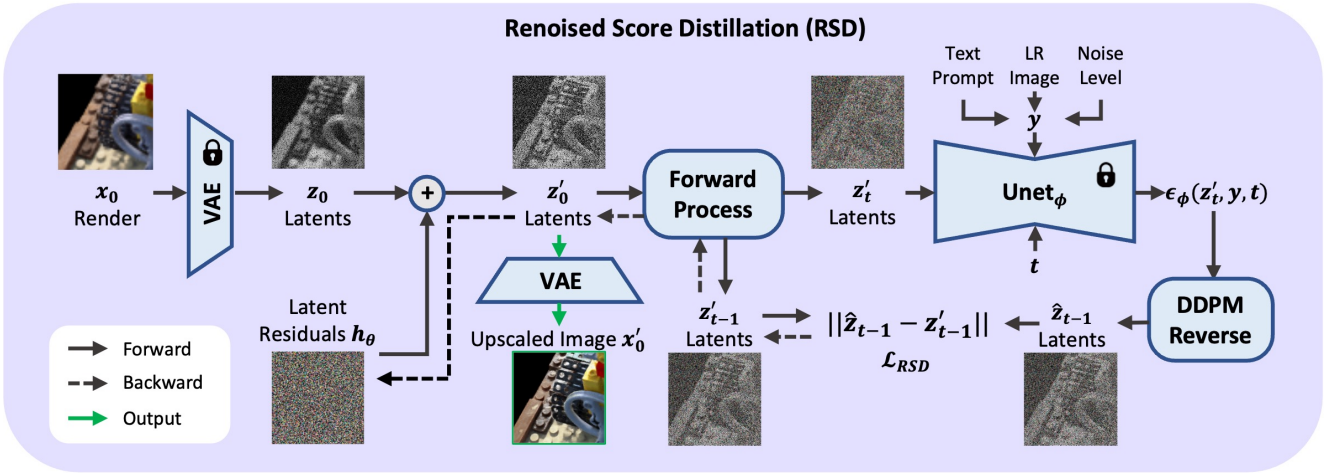

Figure 3. Our RSD produces LR-consistent HR details by optimizing $\mathbf{z}'_{t-1}$ towards predicted denoised latents $\hat{\mathbf{z}}_{t-1}$ following a linearly decreasing time schedule. After optimization, the residuals $\mathbf{h}_\theta$ contain HR details that is added to $\mathbf{z}_0$ to obtain upscaled latents $\mathbf{z}'_0$, which is decoded into LR-consistent upscaled images $\mathbf{x}'_0$. Refer to the text in Sec. 3.3 for more details.

independently upscaled using RSD (refer to Sec. 3.3) with SD×4 to generate high resolution details. Since the initial rendered images are less detailed, the upscaling process can generate varying HR details which may not be multi-view consistent. The second stage is the **synchronization-stage**. This stage resolves multi-view inconsistency by using the RSD-refined images as training inputs for the NeRF $\boldsymbol{\omega}$ that is initialized from $\boldsymbol{\omega}_{lr}$. During synchronization, NeRF $\boldsymbol{\omega}$ is updated using the standard NeRF training procedure [28] where rays are randomly sampled across all training views. Here, we leverage the view-consistent property of NeRF to capture coherent details across views. This stage transfers view-aligned details generated in the upscaling stage into the 3D representation.

There is a synergistic effect between the upscaling and synchronization stages. The synchronization-stage enables NeRF to capture view-consistent details and inconsistent details are naturally discarded. Consequently, the upscaling-stage receives increasingly detailed and view-consistent input images rendered from NeRF. This allows RSD to generate additional details with lower cross-view inconsistency. Furthermore, since RSD is always conditioned on the original LR training images, the I3DS process does not degrade into degenerate solutions. Over successive iterations, I3DS updates NeRF $\boldsymbol{\omega}$ to learn highly detailed and view-consistent features to produce SR NeRF $\boldsymbol{\omega}_{sr}$.

Our I3DS process shares some similarity to NeRF editing approaches [9, 33] which also utilize an alternating training regime to exploit the multi-view consistent property of NeRFs. Nonetheless, we differ from [9, 33] by performing optimization processes in both stages. Furthermore, we optimize all renders concurrently and replace all training images in a single batch after RSD optimization to achieve efficient parallelization and reduce unnecessary image-to-latent encoding. This design consideration achieves 4× reduction in optimization duration.

3.3. Renoised Score Distillation (RSD)

The upscaling-stage in I3DS involves upscaling NeRF renders in 2D. A straightforward solution would be to use ancestral sampling, which is the de-facto method for diffusion-based upscalers. However, we find that using ancestral sampling with I3DS does not lead to high quality results. Although ancestral sampling produces sharp SR images, structural features may deviate from the LR image conditioning which aggravates cross-view inconsistencies (*cf*. Fig. 6 for visualization). Another approach would be to use SDS optimization for 2D upscaling in lieu of ancestral sampling. However, SDS produces images that are less detailed than ancestral sampling despite being LR consistent. The over-smoothed results of SDS optimization, which has also been observed in [50], limits I3DS from producing high quality NeRFs. These observations lead us to the idea of getting the best of both worlds – we propose Renoised Score Distillation (RSD) to incorporate elements of SDS optimization into the ancestral sampling process to achieve detailed SR images that are also LR consistent.

As discussed in Sec. 3.1, ancestral sampling follows a latent trajectory $(\mathbf{z}_T, \mathbf{z}_{T-1}, ..., \mathbf{z}_1, \mathbf{z}_0)$ that gradually transforms a Gaussian noise sample $\mathbf{z}_T$ to a data sample $\mathbf{z}_0$. Given a noisy latent $\mathbf{z}_t$ obtained from encoding and noising a source image $\mathbf{x}_0$, we set our optimization target as the previous timestep latents $\mathbf{z}_{t-1}$ of the ancestral sampling trajectory. Additionally, we use a linearly decreasing time schedule that follows ancestral sampling instead of randomly sampled timesteps t as in SDS. As a result, we incrementally build details onto our optimized image similar to ancestral sampling. We find that our RSD is able

to achieve sharper details compared to SDS and with improved LR consistency compared to ancestral sampling. In contrast to ancestral sampling that uses one noise prediction $\epsilon_\phi(\mathbf{z}_t, y, t)$ to obtain $\mathbf{z}_{t-1}$ from $\mathbf{z}_t$, our RSD guides the image under optimization (parametrized by latent residuals θ) towards $\mathbf{z}_{t-1}$ over multiple noise predictions. Our optimization objective is given by:

$$\mathcal{L}_{RSD} = \|\mathbf{z}_{t-1} - \hat{\mathbf{z}}_{t-1}\|, \tag{7}$$

where $\mathbf{z}_{t-1}$ is the current noised latent at $t-1$, and $\hat{\mathbf{z}}_{t-1}$ is the predicted denoised latent from $\mathbf{z}_t$. Unlike SDS which is defined as a loss gradient that is applied in $\mathbf{z}_0$ space, RSD is formulated as a loss function in order to backpropagate gradients through $\mathbf{z}'_{t-1}$.

Fig. 3 provides a detailed illustration of the RSD optimization process. We first interpolate a randomly sampled image patch $\mathbf{x}_0$ by $4\times$ and encode it to latent $\mathbf{z}_0$ using the pretrained VAE encoder of SD$\times$4. We then create zero-initialized learnable latent residuals $\mathbf{h}_\theta$ for each $\mathbf{z}_0$ such that

$$\mathbf{z}'_0 = \mathbf{z}_0 + \mathbf{h}_\theta, \tag{8}$$

where $\mathbf{z}'_0$ is the refined latents representing the upscaled image $\mathbf{x}'_0$. Subsequently, we apply the forward process in Eq. (1) twice on $\mathbf{z}'_0$ at timesteps t and $t-1$ to obtain two noised latents $\mathbf{z}'_t$ and $\mathbf{z}'_{t-1}$. The UNet backbone of SD$\times$4 (parameterized by ϕ) then takes $\mathbf{z}'_t$ and conditioning y (comprising text prompt, LR image, noise level) as input to predict noise residual $\epsilon_\phi(\mathbf{z}'_t, y, t)$.

We then pass $\epsilon_\phi(\mathbf{z}'_t, y, t)$ to Eq. (3) of the DDPM reverse process to construct the predicted denoised latent $\hat{\mathbf{z}}_{t-1}$. Finally, we compute the L1 error between $\mathbf{z}'_{t-1}$ and $\hat{\mathbf{z}}_{t-1}$ using Eq. (7) and backpropagate gradients through $\mathbf{z}'_{t-1}$ towards $\mathbf{h}_\theta$. After optimization, $\mathbf{h}_\theta$ contains the latent HR residuals that can added to $\mathbf{z}_0$ to produce SR latents $\mathbf{z}'_0$. We can then decode $\mathbf{z}'_0$ using the VAE decoder of SD$\times$4 to obtain SR image $\mathbf{x}'_0$ that is used for training in the NeRF synchronization stage of I3DS. In the supplementary, we provide the pseudocode for I3DS and RSD. We also relate the formulation of RSD to SDS and we show that RSD can be viewed as a *renoised* variant of SDS.

4. Experiments

In this section, we provide both qualitative and quantitative comparisons to demonstrate the effectiveness of the proposed DiSR-NeRF. We also show ablation results under with different upscaling and 3D synchronization methods.

4.1. Experimental Settings

We compare DiSR-NeRF with five baseline models, all of which utilize the Instant-NGP [31] backbone:

- **NGP.** We train NGP only on LR images as a baseline NeRF and render at HR resolution following [8].
- **SD$\times$4.** NeRF trained over images independently upscaled (via ancestral sampling) by the SD$\times$4 upscaler.
- **NeRF-SR.** We compare with NeRF-SR [44] without its HR refinement module as it requires HR reference images.
- **DreamFusion.** We adapt the SD$\times$4 upscaler to provide SDS guidance under the DreamFusion [35] framework, where SDS gradients are backpropagated through rendered image patches to the NeRF parameters.
- **IN2N.** We adapt Instruct-NeRF2NeRF [9], and replace the InstructPix2Pix editor with the SD$\times$4 upscaler. IN2N applies Iterative Dataset Update with ancestral sampling to gradually replace training images with upscaled images in each iteration.

4.2. Dataset and Metrics

We evaluate our models over the NeRF-synthetic [28] dataset containing 8 synthetic subjects and the real-world LLFF [27] dataset containing 8 real-world scenes. On the NeRF-synthetic dataset, we use LR images of 400$\times$400 resolution for training, and we render at 1600$\times$1600 resolution. On the LLFF dataset, we train over LR images of 504$\times$378 resolution, and render at 2016$\times$1512 resolution. When evaluating each scene, we use 100 test poses distributed in a circular arrangement with all cameras directed towards the scene center.

Since SR details are produced by a generative model, different SR results can be valid for the same LR NeRF. Consequently, we do not compare our SR results against the HR ground truth. Instead, we follow [8] to use the Naturalness Image Quality Evaluator (NIQE) [29] as a no-reference evaluator to assess the quality of the rendered SR views. The NIQE metric is a blind image quality analyzer that measures statistical deviations against the natural scene statistics of natural, undistorted images.

Following [12], we use the warped LPIPS metric to assess consistency across viewpoints. The warped LPIPS metric is given by:

$$E_{warp}(I_v, I_{v'}) = \text{LPIPS}(M_{v,v'} \cdot I_v, W(I'_v)), \tag{9}$$

where I_v and I'_v are renders from nearby viewpoints v and v', $W(\cdot)$ is a warping function and $M_{v,v'}$ is a warping mask. The LPIPS [57] score is then computed between the target and warped renders over the masked regions. We use the predicted depths to backproject pixels in I'_v to a point cloud in world space, and apply a point cloud rasterizer to render the warped image $W(I'_v)$ in viewpoint v. In our experiments, we select v' as the 3rd nearest test pose from v.

4.3. Qualitative Results

The qualitative comparisons are shown in Fig. 4 and Fig. 5. Firstly, we observe that DiSR-NeRF produces clearer edges

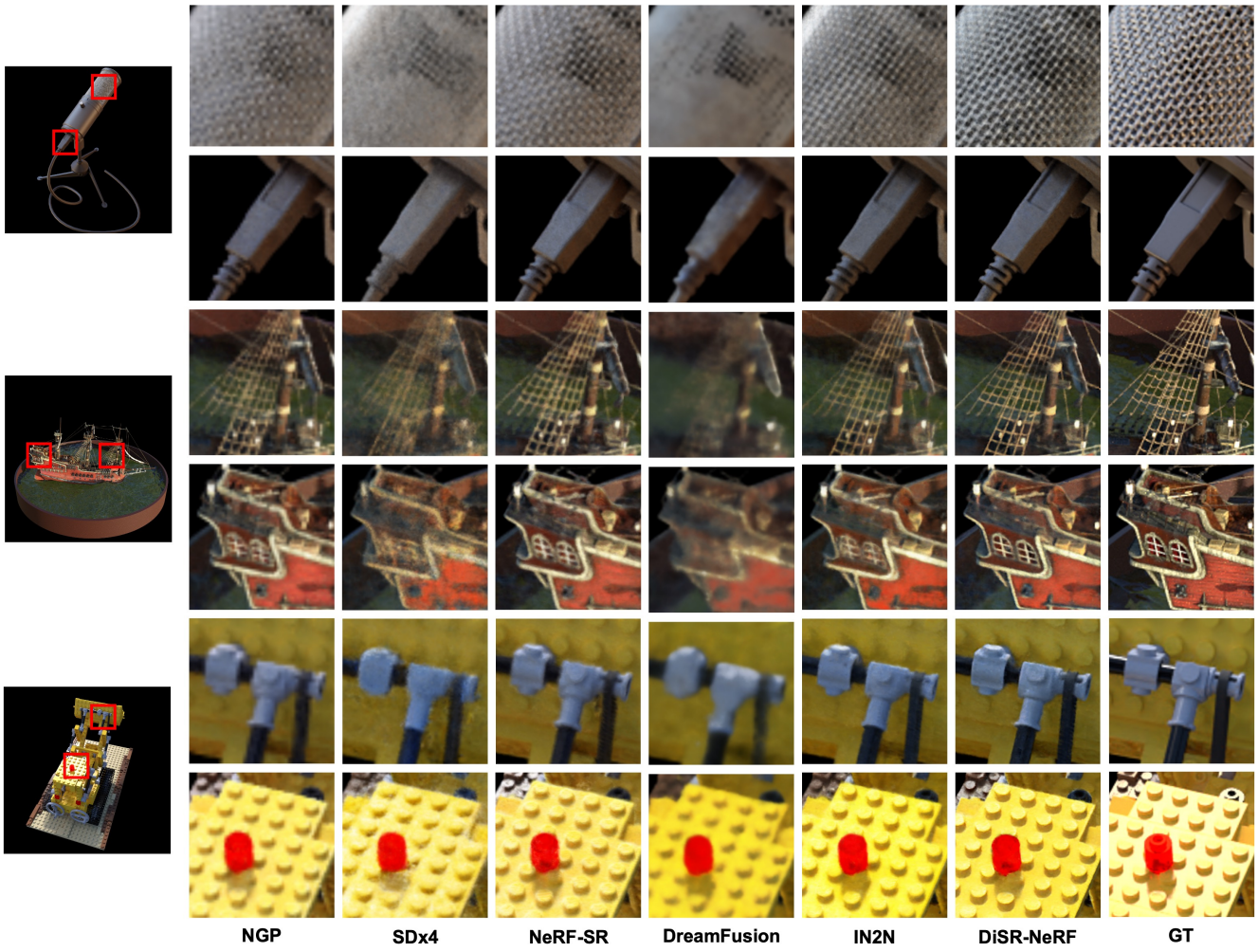

Figure 4. Qualitative Results on NeRF-Synthetic Dataset.

and sharper details compared to all baselines. For example, Fig. 4 (3rd row), our DiSR-NeRF is able to generate highly intricate details of the netting on the mast of the ship. Secondly, as discussed in Sec. 3.2, DreamFusion (4th column) fails to converge and introduces severe blurring to the rendered views. Instead, DiSR-NeRF resolves this effectively by segregating upscaling and NeRF fitting with I3DS. Lastly, we compare DiSR-NeRF against IN2N (5th column) which uses ancestral sampling for upscaling. Across both datasets, we see that our DiSR-NeRF consistently produces sharper and well-defined details over IN2N. This validates the effectiveness of our proposed RSD optimization over ancestral sampling.

4.4. Quantitative Results

We show the quantitative results in Tab. 1. DiSR-NeRF shows significantly improved NIQE scores compared to all baselines, indicating that DiSR-NeRF is able to synthesize views with greater perception quality including increased detail and sharpness. Furthermore, our DiSR-NeRF also achieves better warped LPIPS scores, validating the effectiveness of I3DS in converging towards view-consistent details. Across both synthetic and real datasets, our DiSR-

Methods	NeRF-Synthetic		LLFF	
	NIQE ↓	LPIPS ↓	NIQE ↓	LPIPS ↓
NGP	9.776	0.262	9.163	0.284
SDx4	6.286	0.189	7.461	0.201
NeRF-SR [44]	6.012	0.158	7.038	0.156
DreamFusion [35]	8.624	0.252	8.857	0.247
IN2N [9]	5.847	0.173	6.473	0.157
DiSR-NeRF (Ours)	**5.386**	**0.144**	**5.544**	**0.141**

Table 1. Quantitative comparison between DiSR-NeRF and the baselines on NIQE and warped LPIPS.

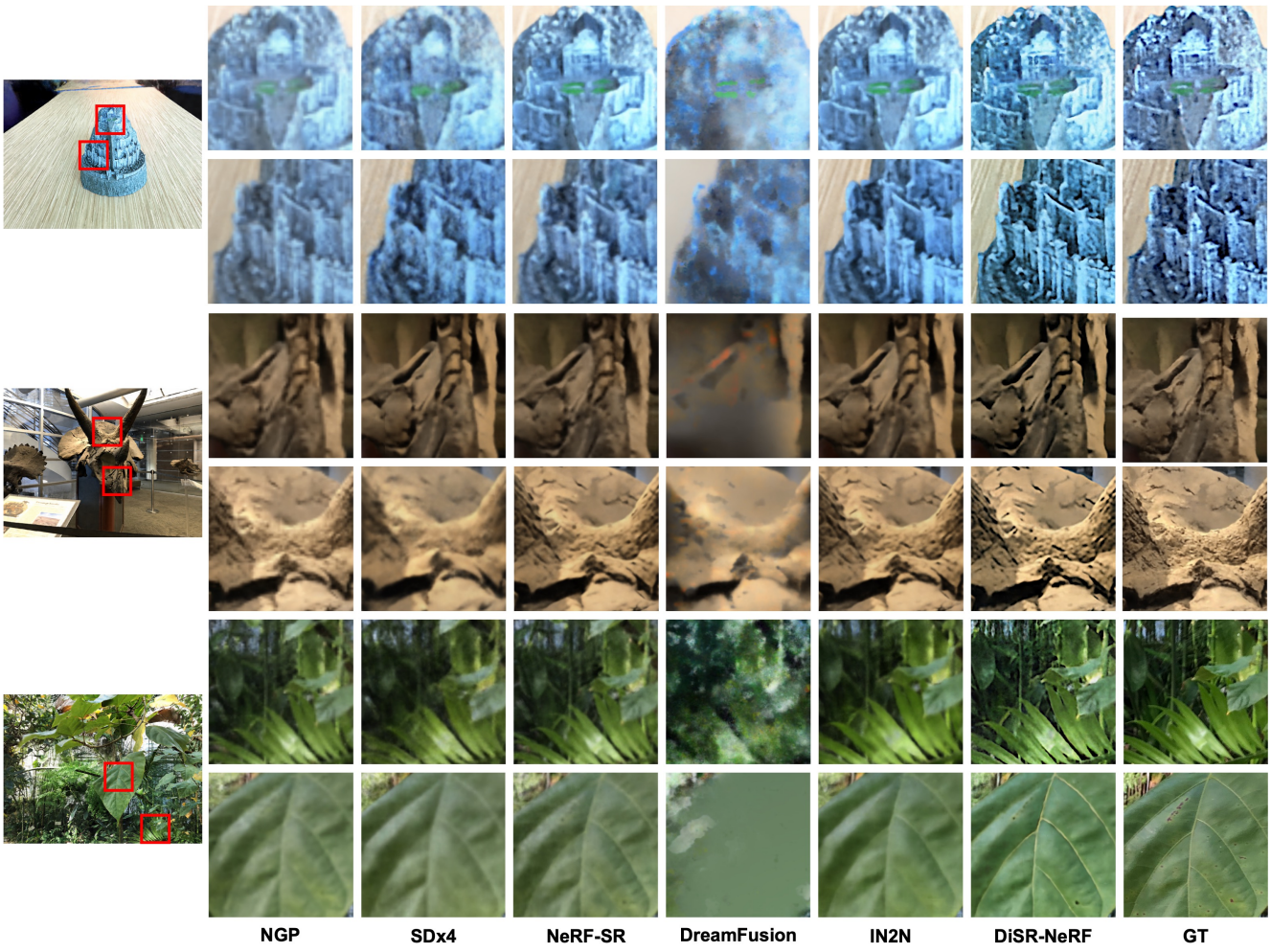

Figure 5. Qualitative Results on LLFF Dataset.

NeRF is able to effectively generate view-consistent SR details, validating its applicability to real-world scenarios.

4.5. Ablations

We also ablate the contributions of the various components in DiSR-NeRF. The quantitative results of the ablations are shown in Tab. 2 and the qualitative results in Fig. 7. Firstly, we retain I3DS and replace the upscaling method with SDS and ancestral sampling instead of RSD (Row 1-2 in Tab. 2). In both cases, replacing RSD results in an increase in NIQE and LPIPS scores, indicating poorer visual quality and lower view-consistency. This can also be observed in the visual ablations in Fig. 7. Ancestral sampling fails to produce sharp details in the SR NeRF as it tends to generate high variance details which may be inconsistent to the LR image conditioning, as shown in our 2D experiments in Fig. 6. On the other hand, SDS produces NeRF renders with increased blur compared to RSD. This effect can also be observed from the 2D experiments in Fig. 6. We hypothesize that the mode-seeking property of the SDS objective optimizes images towards modes which may be far from typical samples [32]. Thus, SDS may guide an image towards the mean of possible solutions which would result in blurred details. Unlike SDS, RSD follows an optimization trajectory similar to ancestral sampling, which guides an image towards more plausible samples.

Methods	NeRF-Synthetic		LLFF	
	NIQE ↓	LPIPS ↓	NIQE ↓	LPIPS ↓
w/o RSD (SDS)	6.273	0.175	5.903	0.164
w/o RSD (Anc.)	5.942	0.189	6.325	0.159
w/o I3DS	8.212	0.254	8.299	0.236
DiSR-NeRF	**5.386**	**0.144**	**5.544**	**0.141**

Table 2. Ablations on I3DS and RSD in DiSR-NeRF.

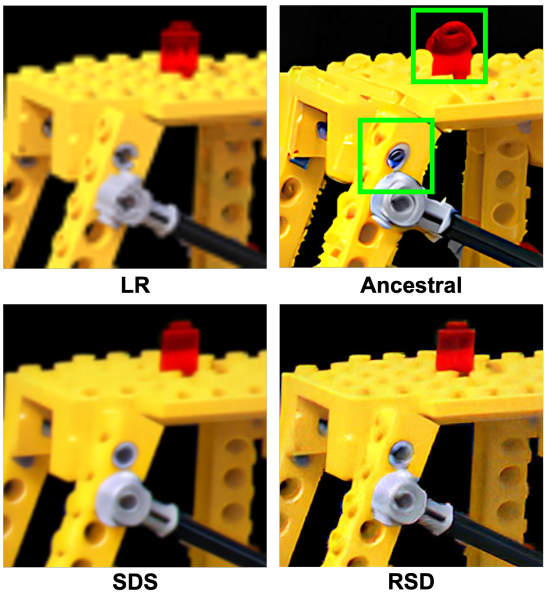

Figure 6. Comparison of 2D upscaling results. Green boxes highlight regions in ancestral sampling that deviate from LR conditioning (top left).

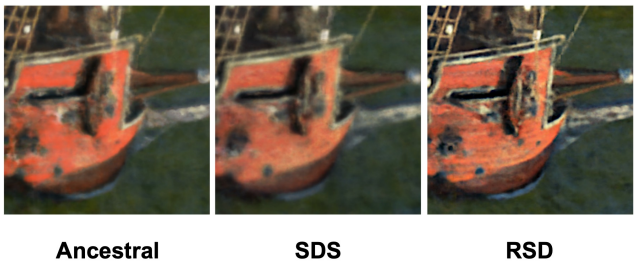

Figure 7. Quantitative ablations on upscaler method (Ancestral sampling, SDS RSD) on DiSR-NeRF.

We also ablate I3DS by replacing it with the DreamFusion framework. The results are shown in the 3rd row of Tab. 2, which shows poor performance compared to our DiSR-NeRF. This indicates that I3DS is an essential component for super-resolution NeRF. Unlike Text-to-3D models which typically provide full image SDS supervision, the SR scenario requires high-resolution SDS guidance. This means small local patches need to be rendered at high resolutions, and the NeRF can only be supervised over a small region in each training step. As a result, each training step guides the NeRF in different optimization directions and thus preventing NeRF from converging towards high quality details. The segregation of the upscaling and NeRF synchronization processes in our I3DS allows RSD guidance to be efficiently applied without the memory constraints of online rendering. Moreover, I3DS also allows the NeRF synchronization to utilize batches of randomly sampled rays across all training views which allows for more stable convergence.

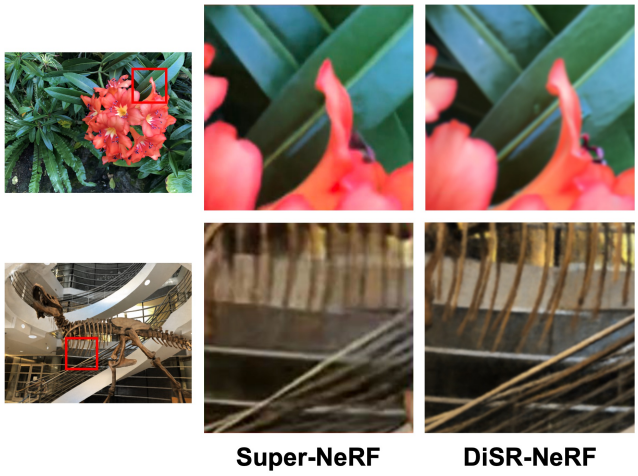

Figure 8. Qualitative comparison with Super-NeRF

4.6. Comparison with Super-NeRF

As the source code for Super-NeRF[8] is not publicly available and the manuscript contains insufficient details on the evaluation method, we are unable to provide quantitative comparisons with the published results. Nonetheless, we conduct a qualitative assessment comparing Super-NeRF and DiSR-NeRF based on available visual results in the paper. The qualitative comparison is shown in Fig. 8. Under similar settings, we observe that DiSR-NeRF can produce sharper details and clearer edges in both examples. In the LLFF-Flower scene, DiSR-NeRF is able to produce realistic textures on the leaves in the background and is able to resolve details on the stigma of the flower. In the LLFF-Trex scene, DISR-NeRF can generate clearer bone structures.

5. Conclusion

In conclusion, we propose DiSR-NeRF, a diffusion-guided super-resolution NeRF framework that distils 2D super resolution priors to the 3D domain to generate view-consistent high resolution details. DiSR-NeRF is able to achieve NeRF super-resolution without requiring high-resolution reference images or large multi-view image datasets. RSD achieves highly detailed, LR-consistent upscaling, while I3DS enables NeRFs to capture view-consistent details over successive upscaling and synchronization cycles. We believe super-resolution methods such as DiSR-NeRF will have significant practical applications especially for devices equipped with low-resolution imaging capabilities.

Acknowledgements. This research is supported by the National Research Foundation, Singapore under its AI Singapore Programmes (AISG Award No: AISG2-PhD/2021-08-012 and AISG Award No: AISG2-RP-2021-024).

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. *ICCV*, 2021. 1, 2

[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. 1, 2

[3] 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. *arXiv preprint arXiv:2103.15595*, 2021. 1

[4] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. *arXiv preprint arXiv:2203.09517*, 2022. 1

[5] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. 3

[6] Yue Chen, Xingyu Chen, Xuan Wang, Qi Zhang, Yu Guo, Ying Shan, and Fei Wang. Local-to-global registration for bundle-adjusting neural radiance fields. *arXiv preprint arXiv:2211.11505*, 2022. 1

[7] 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. 1

[8] Yuqi Han, Tao Yu, Xiaohang Yu, Yuwang Wang, and Qionghai Dai. Super-nerf: View-consistent detail generation for nerf super-resolution, 2023. 2, 5, 8

[9] Ayaan Haque, Matthew Tancik, Alexei Efros, Aleksander Holynski, and Angjoo Kanazawa. Instruct-nerf2nerf: Editing 3d scenes with instructions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 4, 5, 6

[10] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. 3

[11] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *arXiv preprint arXiv:2106.15282*, 2021. 2

[12] Hsin-Ping Huang, Hung-Yu Tseng, Saurabh Saini, Maneesh Singh, and Ming-Hsuan Yang. Learning to stylize novel views, 2021. 5

[13] X. Huang, W. Li, J. Hu, H. Chen, and Y. Wang. Refsr-nerf: Towards high fidelity and super resolution view synthesis. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8244–8253, Los Alamitos, CA, USA, 2023. IEEE Computer Society. 2

[14] Yukun Huang, Jianan Wang, Yukai Shi, Xianbiao Qi, Zheng-Jun Zha, and Lei Zhang. Dreamtime: An improved optimization strategy for text-to-3d content creation, 2023. 3

[15] Younghyun Jo, Sejong Yang, and Seon Joo Kim. Srflow-da: Super-resolution using normalizing flow with deep convolutional block. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2021. 2

[16] Adam R. Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Soňa Mokrá, and Danilo J. Rezende. Nerf-vae: A geometry aware 3d scene generative model. *https://arxiv.org/abs/2104.00587*, 2021. 1

[17] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network, 2017. 2

[18] Haoying Li, Yifan Yang, Meng Chang, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models, 2021. 2

[19] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Nemi: Unifying neural radiance fields with multiplane images for novel view synthesis. *CoRR*, abs/2103.14910, 2021. 1

[20] Jingyun Liang, Andreas Lugmayr, Kai Zhang, Martin Danelljan, Luc Van Gool, and Radu Timofte. Hierarchical conditional flow: A unified framework for image super-resolution and image rescaling. In *IEEE International Conference on Computer Vision*, 2021. 2

[21] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *IEEE International Conference on Computer Vision (ICCV)*, 2021. 1

[22] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[23] 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, 2023. 3

[24] Andreas Lugmayr, Martin Danelljan, Luc Van Gool, and Radu Timofte. Srflow: Learning the super-resolution space with normalizing flow, 2020. 2

[25] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. *arXiv preprint arXiv:2211.07600*, 2022. 3

[26] Zhenxing Mi and Dan Xu. Switch-nerf: Learning scene decomposition with mixture of experts for large-scale neural radiance fields. In *International Conference on Learning Representations (ICLR)*, 2023. 1

[27] 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

[28] 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, 4, 5

[29] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. 5

[30] Alexander Mordvintsev, Nicola Pezzotti, Ludwig Schubert, and Chris Olah. Differentiable image parameterizations. *Distill*, 2018. https://distill.pub/2018/differentiable-parameterizations. 3

[31] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 1, 5

[32] Eric Nalisnick, Akihiro Matsukawa, Yee Whye Teh, Dilan Gorur, and Balaji Lakshminarayanan. Do deep generative models know what they don't know?, 2019. 7

[33] Thu Nguyen-Phuoc, Feng Liu, and Lei Xiao. Snerf: Stylized neural implicit representations for 3d scenes, 2022. 4

[34] Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. 3

[35] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv*, 2022. 3, 5, 6

[36] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2

[37] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement, 2021. 2

[38] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 1

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. 3

[40] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020.

[41] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations, 2021. 3

[42] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. *CVPR*, 2022. 1

[43] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-NeRF: Scalable large scene neural view synthesis. *arXiv*, 2022. 1

[44] 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. *arXiv*, 2021. 2, 5, 6

[45] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A. Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation, 2022. 3

[46] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. *arXiv preprint arXiv:2102.13090*, 2021. 1

[47] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Chen Change Loy, Yu Qiao, and Xiaoou Tang. Esrgan: Enhanced super-resolution generative adversarial networks, 2018. 2

[48] Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. Real-esrgan: Training real-world blind super-resolution with pure synthetic data, 2021. 2

[49] Zhongshu Wang, Lingzhi Li, Zhen Shen, Li Shen, and Liefeng Bo. 4k-nerf: High fidelity neural radiance fields at ultra high resolutions, 2023. 1

[50] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. *arXiv preprint arXiv:2305.16213*, 2023. 3, 4

[51] 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, 2023. 2

[52] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 736–753. Springer, 2022. 1

[53] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. *arXiv preprint arXiv:2201.08845*, 2022. 1

[54] Jie-En Yao, Li-Yuan Tsao, Yi-Chen Lo, Roy Tseng, Chia-Che Chang, and Chun-Yi Lee. Local implicit normalizing flow for arbitrary-scale image super-resolution, 2023. 2

[55] Christopehr Choy Animashree Anandkumar Minsu Cho Yoonwoo Jeong, Seokjun Ahn and Jaesik Park. Self-calibrating neural radiance fields. In *ICCV*, 2021. 1

[56] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. *https://arxiv.org/abs/2012.02190*, 2020. 1

[57] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 5

[58] Junzhe Zhu and Peiye Zhuang. Hifa: High-fidelity text-to-3d with advanced diffusion guidance, 2023. 3