Deceptive-NeRF: Enhancing NeRF Reconstruction using Pseudo-Observations from Diffusion Models

Xinhang Liu HKUST
Shiu-hong Kao HKUST
Jiaben Chen UC San Diego
Yu-Wing Tai HKUST
Chi-Keung Tang HKUST

Abstract

This paper introduces Deceptive-NeRF, a new method for enhancing the quality of reconstructed NeRF models using synthetically generated pseudo-observations, capable of handling sparse input and removing floater artifacts. Our proposed method involves three key steps: 1) reconstruct a coarse NeRF model from sparse inputs; 2) generate pseudo-observations based on the coarse model; 3) refine the NeRF model using pseudo-observations to produce a high-quality reconstruction. To generate photo-realistic pseudo-observations that faithfully preserve the identity of the reconstructed scene while remaining consistent with the sparse inputs, we develop a rectification latent diffusion model that generates images conditional on a coarse RGB image and depth map, which are derived from the coarse NeRF and latent text embedding from input images. Extensive experiments show that our method is effective and can generate perceptually high-quality NeRF even with very sparse inputs.

1 Introduction

Neural Radiance Fields (NeRFs) [26] have emerged as a revolutionary 3D scene representation and achieved unprecedented results in novel view synthesis, where the goal is to render arbitrary unseen viewpoints of a scene from a given set of input images. While producing visually pleasing results, a vanilla NeRF requires a large number of training views and is prone to generating severe artifacts when observations are particularly sparse. This issue substantially limits further and more practical applications of NeRFs, such as AR/VR, autonomous driving, and robotics, considering data collection conditions where lay users can only provide casual few-view observations, for example, using their mobile devices.

To enable NeRFs to address view synthesis from sparse inputs, specifically the few-shot neural rendering problem, recent works have explored several strategies, such as transfer learning methods [51, 3], depth supervision [36, 8], patch-based regularization [15, 29, 10] and frequency regularization [49]. Although these techniques have successfully boosted the state-of-the-art performance of few-shot NeRF, many undesirable artifacts can still be observed in the synthesized novel views.

Recent progress in image synthesis using diffusion models [14, 41, 37, 53] enables a generative approach for few-shot NeRFs. Specifically, few-shot neural rendering can be conceptualized as a 3D-aware image generation task, conditioned on given input images and relative camera poses. Rather than barely extracting information from given sparse views to synthesize novel views, this paradigm [7, 48, 25, 55, 22, 1, 13] addresses the few-shot novel view synthesis problem by generating random plausible samples from the conditional distribution using diffusion models. For example, [7] aims to generate 3D-aware images by using the overall semantics and a text embedding of the given input as the conditions of the diffusion model. [22] learns a view-conditioned diffusion model that synthesizes novel views under a specified camera transformation.

Preprint. Under review.

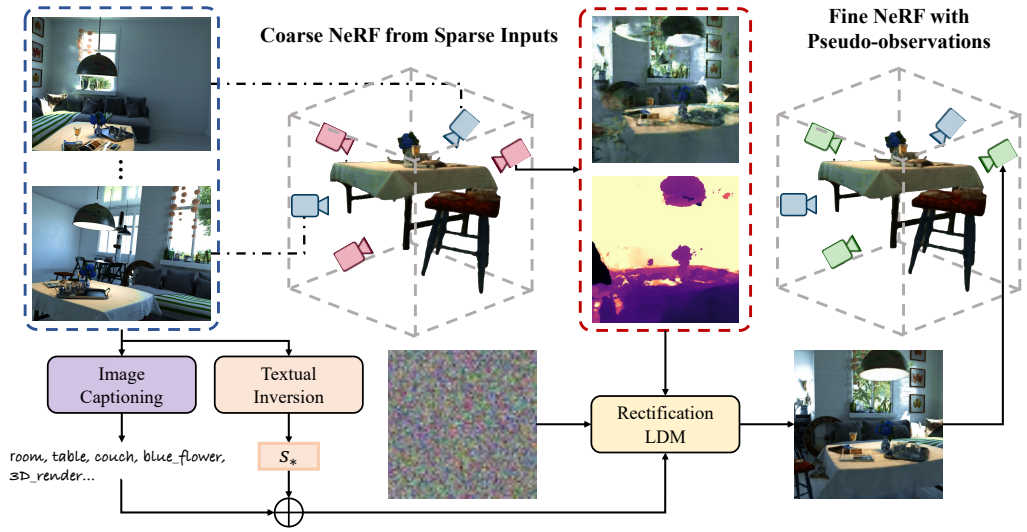

Figure 1: **Overview of Deceptive-NeRF.** 1) Given a sparse set of input images associated with their camera poses, we first train a coarse NeRF, which renders coarse novel view images and depth maps. 2) We use a rectification latent diffusion model to fine-tune RGB-D images from the coarse NeRF to synthesize pseudo-observations from corresponding viewpoints. 3) We train a fine NeRF using both input images (real) and pseudo-observations (fake) as our final reconstruction of the scene while enforcing consistency across the fake images from different viewpoints.

While these diffusion-based methods transfer the natural image prior learned from Internet-scale 2D data to 3D settings to improve the quality of 3D reconstruction from sparse inputs, they apply diffusion in a straightforward manner as a "scorer" to evaluate the quality of NeRF-rendered images, which serves as the training signal for NeRF. This approach necessitates a large diffusion model be inferred at each training step of the radiance field, which is very computationally intensive requiring a lot of time to train even a single 3D scene.

Therefore, we design a new strategy on how to wisely apply large diffusion models (Figure 1). Instead of using diffusion models only as a means to evaluate the quality of NeRF-rendered images, we directly take the images produced by diffusion models as auxiliary observations in addition to sparse inputs to train a NeRF. Specifically, our method consists of three key steps: 1) reconstruct a coarse NeRF model from given sparse views; 2) generate *pseudo-observations* based on the coarse model; 3) refine the NeRF model using pseudo-observations to produce a high-quality reconstruction. To generate plausible pseudo-observations that are geometry-consistent with while faithful to the given sparse views in scene content, we propose a *rectification latent diffusion model* conditioned on coarse RGB image and depth map, given by the coarse NeRF and latent text embedding from input images. This novel strategy not only directly addresses the sparsity issue by "densifying" with fake but consistent observations, but is also less computationally demanding than existing methods due to the one-time use of diffusion models.

In summary, our contributions include the following:

- We propose a novel approach for few-shot novel view synthesis that leverages large diffusion models to generate pseudo-observations instead of using them as a scorer to provide a training signal.
- To generate photo-realistic pseudo-observations that faithfully preserve scene identity and input view consistency, we propose a rectification latent diffusion model trained with a two-stage approach to generate images conditioned on RGB-D renderings from NeRFs.
- Extensive experiments and ablation studies validate our key design choices and demonstrate improvements over current state-of-the-art methods for both single and few-shot novel view synthesis.

2 Related Work

Novel view synthesis via NeRF. Novel view synthesis, the problem of synthesizing new viewpoints given a set of 2D images, has recently attracted considerable attention. Using continuous 3D fields and volumetric rendering, Neural Radiance Fields (NeRFs) [26] have enabled a new and effective approach for novel view synthesis. Follow-up works have since emerged to enhance NeRFs and expand their applications, such as modeling dynamic scenes [52, 30, 32, 43], acceleration [50, 11, 2, 27], and 3D scene editing [23, 52, 46, 16, 19]. Despite significant progress, NeRFs still require hundreds of input images to learn high-quality scene representations. They fail to synthesize novel views with only a few input views, limiting their potential real-world applications.

Few-shot NeRF. Several studies have been proposed to improve the rendering quality of NeRF under few-shot scenario. Among them, [4] provides an augmentation method to enrich training data and reduce overfitting risk. Transfer learning methods [51, 3] adopt large-scale multi-view datasets for pre-training to provide external priors for NeRF. Depth-supervised methods [36, 8] use the estimated depth information as a supplementary supervision for more stable optimization. Patch-based regularization methods impose regularization on rendered patches from different semantic consistency [15], geometry, and appearance [29]. Frequency regularization framework [49] regularizes the visible frequency range of NeRF's inputs to avoid overfitting when training starts. Other attempts include the use of cross-view pixel matching [44], cross-view feature matching [6, 9], ray-entropy regularization [18], and visibility priors [42]. These approaches are generally lightweight in computational complexity and can be used as plug-ins for NeRFs without requiring intricate modifications to the architecture. However, none of the approaches can serve as a panacea to handle all kinds of complex scenes, such as indoor scenes with an outward view or glossy objects with a high degree of view-dependent appearance.

Diffusion models for view synthesis. Recently, diffusion models [14, 28], a powerful class of generative models that follows a Markov process to denoise inputs, have demonstrated notable success on conditional generation [53, 37], such as text-to-image generation [34, 39, 53], image super-resolution [20, 40], and inpainting [24, 38]. By capitalizing on powerful 2D diffusion models, a number of works have advanced the frontier of 3D computer vision tasks, such as 3D content generation and few-shot novel view synthesis. DreamFusion [31] and Magic3D [21] perform text-guided 3D generation by optimizing a NeRF from scratch. Closer to our work, [5, 17, 25, 7, 55, 13] conceptualize the task of few-shot novel view synthesis as 3D-aware conditional image generation. To achieve this, [22] uses a diffusion model trained on synthetic data as geometric priors to synthesize novel views given one single image. [55] transfers 3D consistent scene representation from a view-conditioned diffusion model to improve few-shot novel view synthesis. Unlike the above synthesis approaches, our diffusion model focuses on removing artifacts of a 3D-aware image to generate dense pseudo-observations.

3 Method

As substantial portions of the complex scene may be unobserved and difficult to precisely infer from sparse views, direct prediction of 3D-consistent novel views will inevitably produce blurry outputs in regions of uncertainty. To enable plausible and 3D-consistent predictions given only sparse-view observations of a scene with known camera viewpoints, we instead take an approach outlined in Figure 1. First, we train a coarse NeRF using the given sparse-view inputs to synthesize images of novel views (Section 3.2). Then, given the resulting images from the coarse NeRF, which can be very blurry with a lot of artifacts, we propose a rectification latent diffusion model (Section 3.3) to refine these images. The refined fake images are then served as pseudo-observations to train a fine NeRF (Section 3.4).

3.1 Background

Neural Radiance Fields. A radiance field is a continuous function f mapping a 3D coordinate $\mathbf{x} \in \mathbb{R}^3$ and a viewing directional unit vector $\mathbf{d} \in \mathbb{S}^2$ to a volume density $\sigma \in [0, \infty)$ and RGB values $\mathbf{c} \in [0, 1]^3$ A neural radiance field (NeRF) [26] uses a multi-layer perceptron (MLP) to parameterize

this function:

$$f_\theta : (\mathbf{x}, \mathbf{d}) \mapsto (\sigma, \mathbf{c}) \tag{1}$$

where θ denotes MLP parameters. Some NeRF variants employ explicit voxel grids [50, 11, 2] instead of MLPs to parameterize this mapping for improved efficiency. Our proposed approach is compatible with both MLP-based NeRFs and voxel grid-based variants.

Volume Rendering. Rendering each image pixel given a neural radiance field f_θ involves casting a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ from the camera center $\mathbf{o}$ through the pixel along direction $\mathbf{d}$. The predicted color for the corresponding pixel is computed as:

$$\hat{\mathbf{C}} = \sum_{k=1}^{K} \hat{T}(t_k)\alpha(\sigma(t_k)\delta_k)\mathbf{c}(t_k), \tag{2}$$

where $\hat{T}(t_k) = \exp\left(-\sum_{k'=1}^{k-1} \sigma(t_k)\delta(t_k)\right)$, $\alpha\left(x\right) = 1 - \exp(-x)$, and $\delta_p = t_{k+1} - t_k$. A vanilla NeRF is optimized over a set of input images and their camera poses by minimizing the mean squared error (photometric loss):

$$\mathcal{L}_{\text{pho}} = \sum_{\mathbf{r}\in\mathcal{R}} \|\hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|_2^2 \tag{3}$$

3.2 Coarse NeRF from sparse inputs

Given only a few observations of a scene, i.e., input images $\{I^i_{\text{input}}\}$, we first train an initial coarse NeRF model using these sparse inputs to obtain a rough reconstruction of the scene. The goal of this coarse NeRF reconstruction is to generate initial RGB images and depth predictions at novel views, which will be used as control images feeding into the rectification latent diffusion model to generate pseudo-observations at the same viewpoints.

NeRF is known to overfit few-view observations with small photometric loss while failing to explain 3D geometry consistently across multiple views. Empirically, NeRF's over-fast convergence on high-frequency components of inputs, exacerbated in few-shot settings, impedes its ability to learn coherent 3D geometry. Therefore, following [49], we use a linearly increasing frequency mask to regulate the visible frequency spectrum based on the training time steps.

After training the coarse NeRF from sparse inputs, we sampled many more novel views than the input views and used the coarse NeRF to render RGB images and depth maps. We denote the synthesized pairs of images and depth maps as $\{(I^i_{\text{coarse}}, D^i_{\text{coarse}})\}$ Although the resulting synthesized novel views still exhibit inevitable and obvious artifacts, they provide necessary guidance for the rectification latent diffusion model to obtain refined novel view images as pseudo-observations.

3.3 Rectification Latent Diffusion Model

We propose a 2D diffusion model g that conditions on a coarse RGB image I_{coarse} and its corresponding depth prediction D_{coarse} from the coarse NeRF to synthesize a refined natural image I_{fine} from the same viewpoint:

$$\hat{I}_{\text{fine}} = g(I_{\text{coarse}}, D_{\text{coarse}}), \tag{4}$$

where g in essential rectifies images from the coarse NeRF, and is thus termed the rectification latent diffusion model. The photo-realistic natural images generated serve as pseudo-observations that would cover scarcely observed regions.

Our approach capitalizes on latent diffusion models [37], which obtain natural image priors from internet-scale data to help correct unnaturalness from few-shot NeRFs. Artifacts generated by NeRFs often manifest as "floaters" in empty space. To provide additional guidance, we design the rectification process to also be conditioned on NeRF's depth predictions.

To this end, given a dataset of triplets $\{(I_{\text{fine}}, I_{\text{coarse}}, D_{\text{coarse}})\}$, we fine-tune a pre-trained diffusion model, in which we use a latent diffusion architecture with an encoder $\mathcal{E}$, a denoiser U-Net ϵ_θ, and a decoder $\mathcal{D}$. We solve for the following objective to fine-tune the model:

$$\min_\theta \mathbb{E}_{z\sim\mathcal{E},t,\epsilon\sim\mathcal{N}(0,1)} \|\epsilon - \epsilon_\theta(z_t, t, c(I_{\text{coarse}}, D_{\text{coarse}}, s))\|_2^2, \tag{5}$$

where the diffusion time step $t \sim [1, 1000]$ and $c(I_{\text{coarse}}, D_{\text{coarse}}, s)$ is the embedding of the coarse RGB image, depth estimation, and a text embedding s of the coarse image.

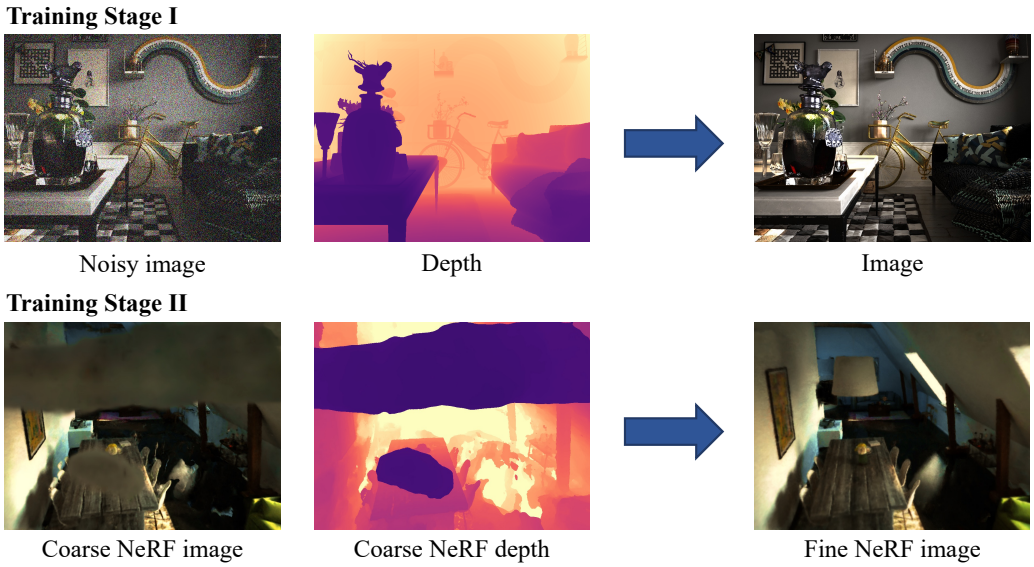

Figure 2: **Two-stage training paradigm for rectification latent diffusion model.** To train the rectification latent diffusion model, in the first stage, we use noisy RGB images and depth maps as inputs, with the denoised RGB images as training targets. In the second stage, we train coarse NeRF and fine NeRF for the same scene, and use coarse RGB images and depth maps from coarse NeRF as inputs, with the high-quality RGB images from the same viewpoint in fine NeRF as training outputs. The fine NeRF is obtained using densely sampled images to train a high-quality NeRF, and the coarse NeRF is obtained using sparse inputs randomly sampled from the dense inputs. This training strategy effectively increases the number of training samples, empowering the rectification latent diffusion model to refine images rendered by the coarse NeRF.

Text embedding. To obtain the text embedding s, we follow [7] to perform both image captioning and textual inversion [12]. On the one hand, we receive a text prompt s_0 from the input image using a pre-trained image captioning network. On the other hand, we optimize for the text embedding of all the input images of a given scene from a text-based image diffusion model for the embedding s_*. We concatenate their embeddings to produce a joint feature $s = [s_0, s_*]$ to encode both the semantic and visual characteristics of the input image.

Effective control upon diffusion models. To enable large pre-trained diffusion models (e.g., Stable Diffusion) to refine RGB-D renderings from coarse NeRFs and synthesize photo-realistic pseudo-observations, we propose fine-tuning them by conditioning on the RGB-D outputs. To enable diffusion models to learn such specific input conditions without disrupting their prior for natural images, we leverage ControlNet [53] to efficiently implement the training paradigm discussed below while preserving the production-ready weights of pre-trained 2D diffusion models.

Two-stage training paradigm. To enable the rectification latent diffusion model to generate an artifact-free image from the same viewpoint as the coarse NeRF's rendered RGB image and depth map, we need to construct a dataset of triplets $\{(I_{\text{fine}}, I_{\text{coarse}}, D_{\text{coarse}})\}$. Specifically, this is achieved by training two versions of NeRF for the same scene: a fine version of NeRF trained on all images and a coarse version of NeRF trained on only one-fifth of the images. By rendering from the same viewpoint, these two opposing NeRFs can generate paired training data samples.

However, this approach requires training two NeRFs for multiple scenes, which cannot generate enough samples within an acceptable time due to limited computing resources. Therefore, as is shown in Figure 2, we propose a two-stage training paradigm. Rather than solely utilizing image pairs from opposing NeRFs, we employ a more direct data source for the initial stage of training, where we acquire paired RGB images and depth maps from public datasets. We add random Gaussian noise to RGB images and use the noisy images and depth maps as inputs for training, with the original RGB images as the training targets. After the first stage, we return to using coarse-fine image pairs synthesized by opposing NeRFs for the second training stage. We found that this two-stage training paradigm effectively solves the problem of insufficient data.

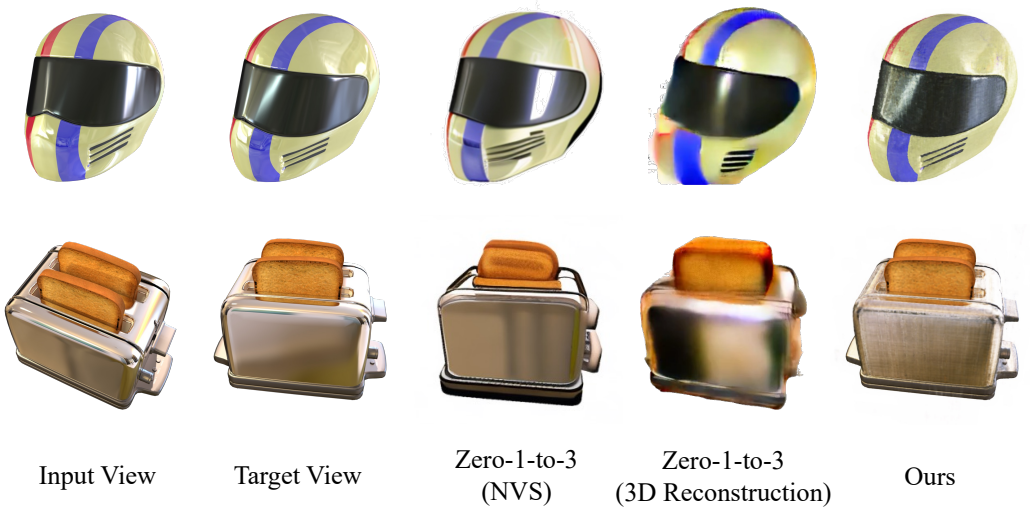

Figure 3: **Qualitative results of novel view synthesis from single images on the Shiny Blender dataset [45].** The input view shown on the left is used to synthesize randomly sampled novel views. The corresponding ground truth of target views is shown in the second column. Compared to the baseline zero-1-to-3 [22], our synthesized novel views are more consistent with the ground truth, while zero-1-to-3 exhibits a significant loss of high-frequency details. Due to zero-1-to-3's generative nature, a novel view is generated rather than reconstructing one from input information.

Table 1: Quantitative results of novel view synthesis from single images on Shiny Blender [45].

	PSNR↑	SSIM↑	LPIPS↓
zero123-N [22]	11.09	0.783	0.2218
zero123-3 [22]	9.26	0.596	0.3842
Ours	**26.94**	**0.936**	**0.0525**

Table 2: Quantitative results of few-shot novel view synthesis on Hypersim [35].

	PSNR↑	SSIM↑	LPIPS↓
FreeNeRF [49]	13.76	0.716	0.1579
DietNeRF [15]	16.73	0.607	**0.1036**
Ours	**17.77**	**0.740**	0.1057

3.4 Fine NeRF with Pseudo-observations

Using the rectification latent diffusion model, we fine-tune the RGB-D images from the coarse NeRF to obtain pseudo-observations of the scene. Thanks to the natural image prior of the latent diffusion model, the pseudo-observations eliminate the artifacts in the images rendered by the coarse NeRF. As our final 3D representation of the scene, we train a fine NeRF model by combining the original input images (real) and pseudo-observations (fake). In this way, we have alleviated the struggle of NeRF in face of sparse observations by synthesizing fake but plausible observations. It should be noted that because the rectification latent diffusion model does not constrain cross-view consistency when synthesizing images, inconsistencies may exist between the pseudo-observations and the input images. However, we found that such inconsistencies were automatically corrected during the training of the fine NeRF. The same processes of Deceptive-NeRF can also be applied again to further boost the reconstruction quality with denser pseudo-observations.

4 Experimental Results

In this section, we evaluate our proposed Deceptive-NeRF method on a variety of challenging scenarios. We present a quantitative and qualitative comparison of our model against state-of-the-art on different objects and environments. Furthermore, we conduct an in-depth analysis of the various components and architectural decisions underlying our approach. Please refer to the supplementary video for more video results.

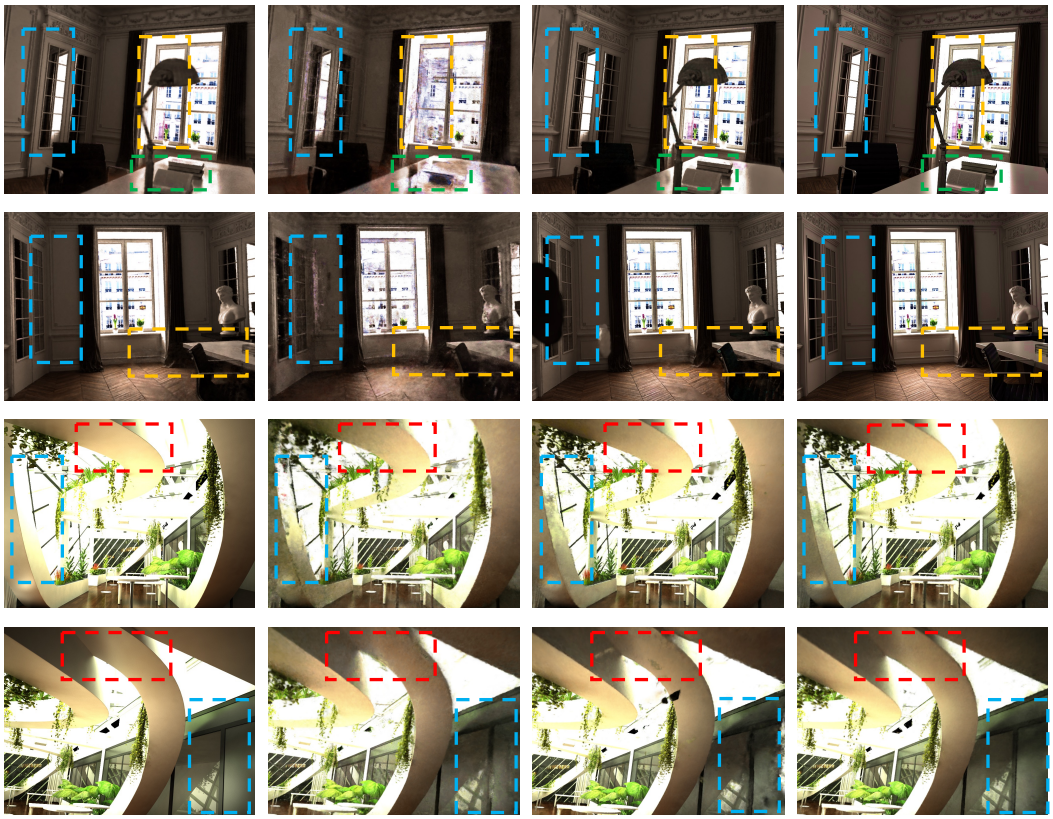

Figure 4: **Few-shot novel view synthesis on Hypersim dataset [35].** Corresponding ground truths of novel views are shown on the left. Compared to the baseline FreeNeRF [49] and DietNeRF [15], our synthesized novel views are more photo-realistic, thanks to the generated pseudo-observations. In contrast, baselines tend to generate blurry results, and even erroneously remove objects near the cameras.

Tasks. Novel view synthesis poses a longstanding three-dimensional challenge in computer vision that mandates a model to (implicitly) acquire at least adequate if not correct depth, texture, and shape of an object. Our experiments focus on taxing scenarios with an extreme paucity of input data in the form of either a single view or a few sparse views.

Baselines. We compare our method to a number of methods within a similar scope. For novel view synthesis from single images, we benchmark against the recent state-of-the-art method zero-1-to-3 [22], which also exploits the geometric priors gleaned by large-scale diffusion models regarding natural images. The method zero-1-to-3 affords two modalities of utilization: first, "novel view synthesis" to synthesize a target view image from an input view image and relative camera pose; and second, "3D reconstruction" to synthesize a 3D object model with geometry and texture from a single input view image. We benchmark our approach against zero-1-to-3 under both of these settings. For few-shot novel view synthesis, we compare against DietNeRF [15] which regularizes NeRF with a CLIP [33] image-to-image consistency loss over viewpoints; and FreeNeRF [49] which regularizes the frequency range of NeRF's inputs while penalizing the proximal camera density fields.

Benchmarks and metrics. We assess novel view synthesis conditioned on a single image using the Shiny Blender dataset [45] which contains glossy objects with complex material properties rendered in Blender under conditions akin to prototypical NeRF datasets. For few-shot novel view synthesis, we evaluate the performance of various approaches on the Hypersim dataset [35], an exacting synthetic dataset comprising of photorealistic indoor environments. We conduct an extensive quantitative analysis of our approach and these strong baselines across three metrics encapsulating distinct facets

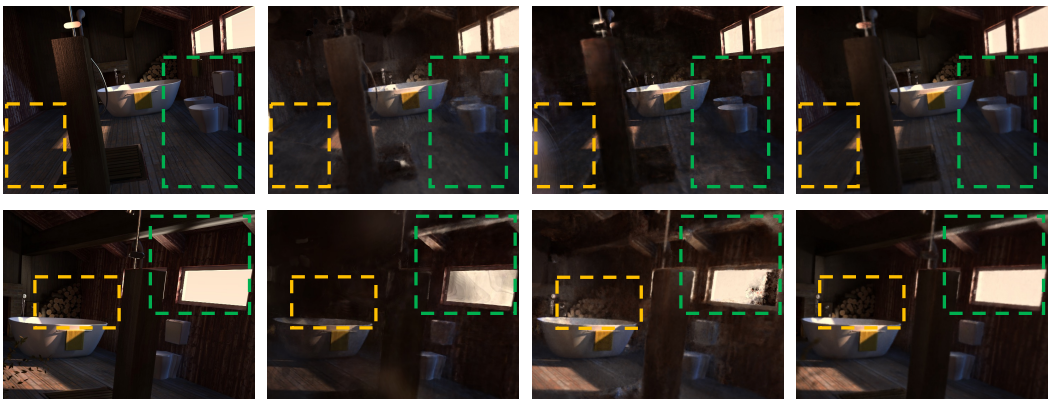

Figure 5: **Qualitative evaluation on the rectification latent diffusion model.** We conduct an ablation study on the architecture of the rectification latent diffusion model. Both variants underperform compared to the full model, demonstrating the necessity of coarse depth predictions and the two-stage training paradigm.

of image similarity, including peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM) [47], mean absolute error (MAE), and learned perceptual image patch similarity (LPIPS) [54].

4.1 Novel view synthesis from single images

We verify Deceptive-NeRF's capability of novel view synthesis from single images on Ref-NeRF dataset [45]. Figure 3 and Table 1 show that our method can generate highly photorealistic images. Due to zero-1-to-3's generative nature, it is more inclined to generate a novel view rather than reconstruct one from input views and thus cannot recover correct details. This shortcoming is even more particularly pronounced in zero-1-to-3 (3D Reconstruction), which is susceptible to generating indistinct novel views owing to 3D ambiguities. We further emphasize that unlike zero-1-to-3 and other concurrent works which require inferring the diffusion model at each training step of the NeRF, our approach is more computationally efficient due to a different way of employing the diffusion model.

4.2 Few-shot novel view synthesis

We evaluate Deceptive-NeRF and baselines on Hypersim [35], a photorealistic synthetic dataset consisting of various indoor scenes on few-shot novel view synthesis. Figure 4 and Table 2 compare the results of our method with two state-of-the-art methods as our baselines. Our method exhibits fewer artifacts in comparison. FreeNeRF can erroneously remove objects near the camera (the lamp on the first column) due to the regularization of the frequency range and near-camera density fields. DietNeRF achieves superior LPIPS scores, which may be attributed to its consistency loss based on CLIP. However, it tends to generate blurry renderings for poorly observed regions.

4.3 Ablation Studies

Rectification latent diffusion model. We conduct an ablation study on the design of the rectification latent diffusion model. Specifically, we evaluate two variants of the model: "w/o depth" denotes that the diffusion model is not conditioned on depth, and "w/o two-stage" denotes the design without the two-stage training paradigm shown in Figure 2. As shown in Figure 5 and Table 3, these two variants perform worse than the full model. This demonstrates that depth prediction from the coarse NeRF is indispensable for guiding the synthesis of pseudo-observations, while the two-stage training paradigm effectively expands the amount of training data.

Number of pseudo-observations. We conduct an additional ablation study to investigate the impact of the number of pseudo-observations. We quantify the number of pseudo-observations

Figure 6: **Qualitative evaluation on the number of pseudo-observations.** $\alpha = \frac{\#\text{pseudo-observations}}{\#\text{input views}}$. Too few pseudo-observations fail to alleviate the scarcity situation, while excessive pseudo-observations create inconsistencies making the model overfit to solutions with severe artifacts.

Table 3: Quantitative evaluation on rectification latent diffusion model.

	PSNR↑	SSIM↑	LPIPS↓
w/o depth	17.57	0.719	0.3921
w/o two-stage	18.19	0.780	0.3517
full model	**21.46**	**0.827**	**0.3297**

Table 4: Quantitative evaluation on the number of pseudo-observations.

	PSNR↑	SSIM↑	LPIPS↓
$\alpha = 0.2$	18.16	0.757	0.1198
$\alpha = 5$	17.68	0.669	0.1811
$\alpha = 1$	**23.45**	**0.821**	**0.0884**

using $\alpha = \frac{\#\text{pseudo-observations}}{\#\text{input views}}$. As illustrated in Figure 6 and Table 4, an insufficient number of pseudo-observations cannot overcome the sparsity of the original inputs, whereas too many pseudo-observations introduce conflicts causing the model to overfit to artifact-ridden solutions. We select α around 1 as our final choice.

5 Discussion

Limitations. While leveraging 2D diffusion models to enhance 3D neural representations in a novel manner, our approach faces several limitations. First, the pseudo-observations generated by the rectification latent diffusion model are not guaranteed to accurately reflect ground truth. Consequently, our results may appear deceptively realistic yet incorrect. Furthermore, as the output of a diffusion model is uncertain and depends on its denoising process, rectified images can vary. Currently, we determine which sample to use manually, but believe this could be done more judiciously. Finally, the rectification latent diffusion model can still perform poorly when handling scenes that differ substantially from its training samples. We plan to train it at a larger scale to uncover its full potential.

Conclusion. We introduce Deceptive-NeRF, which synthesizes pseudo-observations for improving NeRF reconstruction from sparse input. A coarse NeRF model is first reconstructed from the given sparse input, followed by generating pseudo-observations based on the coarse model. Finally, the NeRF model is improved using pseudo-observations to produce a high-quality reconstruction. To generate pseudo-observations that faithfully preserve the identity of the underlying scene while consistent with the sparse inputs, we develop a rectification latent diffusion model that generates images conditional on the relevant coarse RGB image and depth map, which are derived from the coarse NeRF and latent text embedding from input images. Extensive experiments and comparisons demonstrate that our method is effective and can generate perceptually high-quality NeRF reconstructions even with highly sparse inputs.

References

[1] Eric R. Chan, Koki Nagano, Matthew A. Chan, Alexander W. Bergman, Jeong Joon Park, Axel Levy, Miika Aittala, Shalini De Mello, Tero Karras, and Gordon Wetzstein. GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In *arXiv*, 2023.

[2] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, pages 333–350. Springer, 2022.

[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. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14124–14133, 2021.

[4] Di Chen, Yu Liu, Lianghua Huang, Bin Wang, and Pan Pan. Geoaug: Data augmentation for few-shot nerf with geometry constraints. In *European Conference on Computer Vision (ECCV)*, pages 322–337. Springer, 2022.

[5] Hansheng Chen, Jiatao Gu, Anpei Chen, Wei Tian, Zhuowen Tu, Lingjie Liu, and Hao Su. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *arXiv preprint arXiv:2304.06714*, 2023.

[6] Yuedong Chen, Haofei Xu, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Explicit correspondence matching for generalizable neural radiance fields. In *arXiv preprint arXiv:2304.12294*, 2023.

[7] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *arXiv preprint arXiv:2212.03267*, 2022.

[8] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12882–12891, 2022.

[9] Yilun Du, Cameron Smith, Ayush Tewari, and Vincent Sitzmann. Learning to render novel views from wide-baseline stereo pairs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[10] Thibaud Ehret, Roger Marí, and Gabriele Facciolo. Nerf, meet differential geometry! In *arXiv preprint arXiv:2206.14938*, 2022.

[11] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5501–5510, 2022.

[12] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *arXiv preprint arXiv:2208.01618*, 2022.

[13] Jiatao Gu, Alex Trevithick, Kai-En Lin, Josh Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *arXiv preprint arXiv:2302.10109*, 2023.

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

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

[16] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12949–12958, 2021.

[17] Animesh Karnewar, Andrea Vedaldi, David Novotny, and Niloy Mitra. Holodiffusion: Training a 3D diffusion model using 2D images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023.

[18] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[19] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[20] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. In *Neurocomputing*, volume 479, pages 47–59. Elsevier, 2022.

[21] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[22] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[23] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[24] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, 2022.

[25] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Realfusion: 360 {\deg} reconstruction of any object from a single image. In *arXiv preprint arXiv:2302.10663*, 2023.

[26] 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 *European Conference on Computer Vision (ECCV)*, 2020.

[27] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4):1–15, 2022.

[28] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning (ICML)*, volume 139, pages 8162–8171, 2021.

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

[30] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[31] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2022.

[32] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10318–10327, 2021.

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

[34] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. In *arXiv preprint arXiv:2204.06125*, 2022.

[35] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M Susskind. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10912–10922, 2021.

[36] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

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

[38] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM Transactions on Graphics (SIGGRAPH)*, pages 1–10, 2022.

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pages 36479–36494, 2022.

[40] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. In *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*. IEEE, 2022.

[41] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning (ICML)*, pages 2256–2265, 2015.

[42] Nagabhushan Somraj and Rajiv Soundararajan. ViP-NeRF: Visibility prior for sparse input neural radiance fields. In *ACM Transactions on Graphics (SIGGRAPH)*, August 2023.

[43] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12959–12970, 2021.

[44] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. SPARF: Neural radiance fields from sparse and noisy poses. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[45] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022.

[46] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3835–3844, 2022.

[47] 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 (TIP)*, 13(4):600–612, 2004.

[48] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Yi Wang, and Zhangyang Wang. Neurallift-360: Lifting an in-the-wild 2d photo to a 3d object with 360° views. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[49] Jiawei Yang, Marco Pavone, and Yue Wang. FreeNeRF: Improving few-shot neural rendering with free frequency regularization. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[50] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5752–5761, 2021.

[51] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[52] Jiakai Zhang, Xinhang Liu, Xinyi Ye, Fuqiang Zhao, Yanshun Zhang, Minye Wu, Yingliang Zhang, Lan Xu, and Jingyi Yu. Editable free-viewpoint video using a layered neural representation. *ACM Transactions on Graphics (TOG)*, 40(4):1–18, 2021.

[53] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *arXiv preprint arXiv:2302.05543*, 2023.

[54] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 586–595, 2018.

[55] Zhizhuo Zhou and Shubham Tulsiani. Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.