Deceptive-NeRF/3DGS: Diffusion-Generated Pseudo-Observations for High-Quality Sparse-View Reconstruction

Xinhang Liu[1], Jiaben Chen[2], Shiu-Hong Kao[1], Yu-Wing Tai[3], and Chi-Keung Tang[1]

[1] The Hong Kong University of Science and Technology
[2] University of California, San Diego
[3] Dartmouth College
{xliufe,skao}@connect.ust.hk, jic088@ucsd.edu, yu-wing.tai@dartmouth.edu, cktang@cs.ust.hk

Abstract. Novel view synthesis via Neural Radiance Fields (NeRFs) or 3D Gaussian Splatting (3DGS) typically necessitates dense observations with hundreds of input images to circumvent artifacts. We introduce Deceptive-NeRF/3DGS[1] to enhance sparse-view reconstruction with only a limited set of input images, by leveraging a diffusion model pre-trained from multiview datasets. Different from using diffusion priors to regularize representation optimization, our method directly uses diffusion-generated images to train NeRF/3DGS as if they were real input views. Specifically, we propose a *deceptive diffusion model* turning noisy images rendered from few-view reconstructions into high-quality photorealistic *pseudo-observations*. To resolve consistency among pseudo-observations and real input views, we develop an uncertainty measure to guide the diffusion model's generation. Our system progressively incorporates diffusion-generated pseudo-observations into the training image sets, ultimately densifying the sparse input observations by 5 to 10 times. Extensive experiments across diverse and challenging datasets validate that our approach outperforms existing state-of-the-art methods and is capable of synthesizing novel views with super-resolution in the few-view setting. Project page: https://xinhangliu.com/deceptive-nerf-3dgs.

Keywords: NeRF · 3D Gaussian Splatting · Few-view reconstruction

[1] In harmonic progression, a Deceptive Cadence may disrupt expectations of chord progression but enriches the emotional expression of the music. Our Deceptive-X, where "X" can be NeRF, 3DGS, or a pertinent 3D reconstruction framework—counters overfitting to sparse input views by densely synthesizing consistent pseudo-observations, enriching the original sparse inputs by fivefold to tenfold.

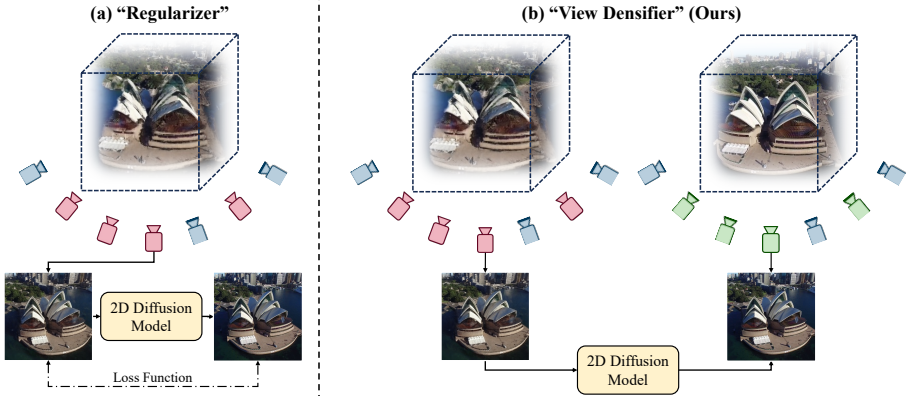

Fig. 1: Different approaches to applying 2D diffusion priors in few-view 3D reconstruction. (a) With only a few input images , an intuition is to utilize the 2D diffusion model as a "scorer" for synthesized novel views , regularizing NeRF/3DGS training. **(b)** Instead, our approach densifies input views by generating dense pseudo-observations that are consistent with the given inputs to progressively enhance reconstruction as shown. This approach avoids the need to infer the diffusion model at every training step, thereby offering the advantage of nearly tenfold faster training speed.

1 Introduction

Recent novel view synthesis approaches like Neural Radiance Fields (NeRFs) [40] and 3D Gaussian Splatting [29] have achieved unprecedented results. However, without exception, all these pipelines [5,8,41] require a large number of training views to produce visually pleasing results and are prone to generating severe artifacts when dealing with sparse observations. This issue can hamper their further consumer-level usage, where casual data collection by lay users is often made up of only sparse image capture using mobile devices.

To adapt NeRF to the few-view setting, some methods leverage large-scale datasets comprising various scenes to pre-train their models and inject prior knowledge [9,12,26,27,73]. A larger portion of methods [51,52] employs a range of regularizations derived from depth supervision [15,23,48], appearance [42,69], semantic consistency [25], visibility [32,56], or frequency patterns [71]. Although these techniques have improved the reconstruction quality, undesirable artifacts remain in the synthesized novel views. While 3DGS [29] has improved rendering quality and efficiency using differentiable splatting, it still struggles with the few-view setting, a direction that has not been sufficiently explored.

Diffusion models [24,50,54,57,76] have illuminated potential pathways for addressing the few-view NeRF/3DGS reconstruction challenge. By lifting the comprehensive visual priors learned from vast 2D datasets into 3D, there have been

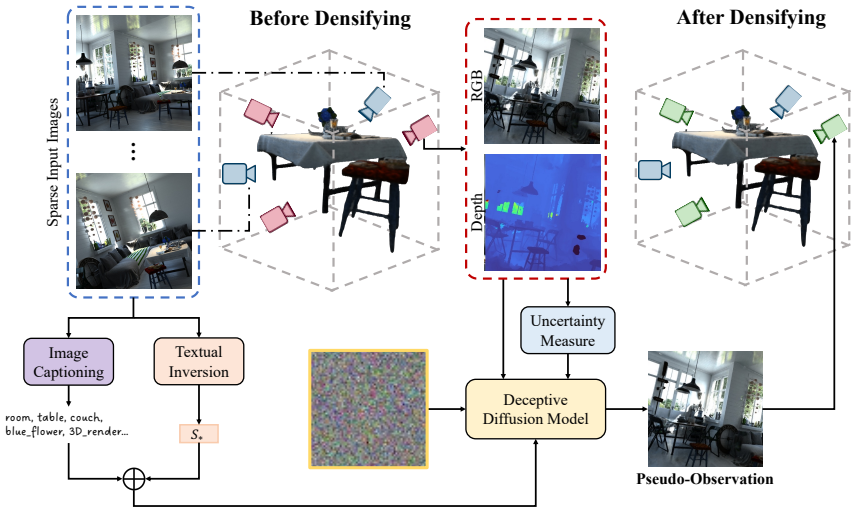

Fig. 2: Overview of Deceptive-NeRF/3DGS. 1) Given a sparse set of input images with camera poses, we train NeRF/3DGS to render coarse novel view images and depth maps. 2) Our deceptive diffusion model enhances RGB-D images from coarse reconstruction, along with a novel uncertainty measure, to generate pseudo-observations from corresponding viewpoints. 3) We continue training NeRF/3DGS using both input images (real) and pseudo-observations (fake) and repeat the aforementioned process to get our final reconstruction.

significant advancements in 3D content generation[2] [7, 14, 22, 34, 38, 65, 70, 78]. Given this backdrop, an intuitive and investigated approach to utilizing 2D diffusion models for sparse view reconstruction is to employ them as a "scorer" evaluating the quality of rendered images and thus a regularizer for NeRF training similar to score distillation sampling (SDS) [44]. This approach [68, 69] however necessitates a large diffusion model be inferred at each training step of the radiance field, a computationally intensive process.

In this paper, we propose Deceptive-NeRF/3DGS to efficiently leverage large diffusion models for sparse view reconstruction. Instead of using diffusion models only as a means to regularize the NeRF/3DGS training process, we produce *pseudo-observations* by a diffusion model, directly applying them to the training pipeline as if they were real input views, as shown in Figure 1. Specifically, we propose a *deceptive diffusion model* that turns coarsely rendered noisy images from few-view reconstructions into high-quality photorealistic pseudo-observations. To achieve seamless integration of real input views and synthetic

[2] Note that 3D content generation from images differs fundamentally from few-view reconstruction, due to the need to satisfy the multi-view geometry constraint. This work tackles the latter, where the goal is "reconstruction" rather than "generation".

pseudo-observations in the training process, we develop a reconstruction uncertainty measure to guide the diffusion model's generation.

With a trained deceptive diffusion model, our method alternates between the two steps (Figure 2): 1) feeding rendered images into a diffusion model for generating pseudo-observations and 2) optimizing the representation using both the real input images and the newly generated pseudo-observations. Consequently, this novel approach tackles the issue of sparsity by "densifying" observations by 5 to 10 times, while not demanding excessive time or computation, thanks to the one-time usage of diffusion models. As a by-product of such densification of input observations, our method acquires an additional feature at no extra cost—rendering novel views with super-resolution. This enables our approach to perform high-quality novel view synthesis even under the stringent conditions of few-view, low-resolution settings.

In summary, our contributions include the following:

- We propose a novel approach for few-view reconstruction, agnostic to the underlying scene representation.
- Our approach leverages large diffusion models to generate pseudo-observations and densify input observations fivefold to tenfold. Compared to using diffusion models as a "scorer" to regularize the training process, this manner largely reduces the computational cost, with nearly ten times the training speed.
- We propose a deceptive diffusion model to generate high-quality photorealistic pseudo-observations preserving scene semantics and view consistency.
- Our method can render novel views with super-resolution, surpassing the limitations posed by low-quality inputs.

2 Related Work

Neural 3D Representations. Using continuous 3D fields modeled by MLPs and volumetric rendering, Neural Radiance Fields (NeRFs) [40] have enabled a new and effective approach for novel view synthesis and reconstruction. Follow-up works have since emerged to enhance NeRFs and expand their applications [60], such as improving rendering quality [2,3,5], modeling dynamic scenes [1, 6,17,37,43,45,61,66,75], acceleration [18,72], and 3D scene editing [26,31,35,36, 63,75]. Recent work has shown that replacing the deep MLPs with a feature voxel grid network can significantly improve training and inference speed [8,19,41,58]. More recently, 3D Gaussian Splatting [29] further improves visual quality, rendering time, and performance. Despite significant progress, NeRF, 3DGS, and their variants still struggle to reconstruct high-quality 3D models when there are only a limited number of input views.

Few-view Reconstruction. Several studies have been conducted to enhance NeRF when provided with only sparse observations. Some approaches pre-train their models utilizing prior knowledge from extensive datasets of 3D scenes to generate novel views from the given sparse observations [9,12,26,27,73]. On the other hand, more attempts [23,49,51,55,56,69] employ a range of regularizations

to the NeRF training pipeline. Among them, [15, 48] use the estimated depth information from Structure-from-Motion (SfM) or depth estimation networks as supplementary supervision. FreeNeRF [71] regularizes the visible frequency range of NeRF's inputs to avoid overfitting. [25, 42, 69] impose regularization on rendered patches from semantic consistency, geometry, and appearance. Other attempts include the use of cross-view pixel matching [62], cross-view feature matching [11, 16], ray-entropy regularization [30], and visibility priors [32, 56]. Yet, existing approaches still fall short of the requirement of consumer-level usage and necessitate scene-specific heuristic adjustments. Our approach achieves improvements over the previously mentioned few-view NeRF techniques and, for the first time, enhances 3DGS performance in the few-view setting.

Diffusion models for 3D. By capitalizing on powerful 2D diffusion models, many works have advanced the frontier of 3D computer vision tasks, such as 3D content generation [10, 14, 22, 28, 33, 38, 44, 78, 78]. To achieve this, [34] uses a diffusion model trained on synthetic data as geometric priors to synthesize novel views given one single image. Closer to our work, DiffusioNeRF [69] and ReconFusion [68] regularize few-view NeRFs with priors from diffusion models. Unlike these methods that utilize diffusion models in a 3D setting, our approach does not employ them as a "scorer" for regularization. Instead, we use the images generated by the diffusion model as auxiliary pseudo-observations directly for NeRF training, treating them as if they were real input views. As a result, our method avoids inferring the diffusion model at every training step, reducing computational costs and achieving nearly ten times the training speed.

3 Preliminaries

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) [40] 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. While existing NeRF variants employ explicit voxel grids [8, 18, 72] 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. 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 $\hat{T}(t_k) = \exp\left(-\sum_{k'=1}^{k-1} \sigma(t_k)\delta(t_k)\right)$, $\alpha(x) = 1 - \exp(-x)$, 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):

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

3D Gaussian Splatting. 3DGS [29] represents a 3D scene explicitly through a collection of 3D Gaussians. Each of the Gaussians is parameterized with a position vector $\boldsymbol{\mu} \in \mathbb{R}^3$ and a covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{3\times 3}$. Each Gaussian influences a point $\mathbf{x}$ in 3D space following:

$$G(\mathbf{x}) = \frac{1}{(2\pi)^{3/2}\,|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}. \tag{4}$$

Each Gaussian stores an opacity logit $o \in \mathbb{R}$ and the appearance feature is represented by spherical harmonic (SH) coefficients. To render the RGB color of a pixel, one orders all the Gaussians that contribute to it and blends the ordered Gaussians using:

$$\hat{\mathbf{C}} = \sum_{i=1}^{n} \mathbf{c}_i \alpha_i \prod_{j=1}^{i-1}(1-\alpha_j), \tag{5}$$

where $\mathbf{c}_i$ is the color computed from the SH coefficients. α_i is given by evaluating a 2D Gaussian multiplied by the opacity logit o.

4 Method

Given only W observations of a scene, i.e., input images $\mathcal{I} = \{I_1, I_2, \ldots, I_W\}$ with their calibrated camera poses, our approach enhances the NeRF reconstruction through employing diffusion models to "densify" the inputs, as illustrated in Figure 2. In Section 4.1, we describe our pipeline, which continuously densifies the input observations through synthetic pseudo-observations generated by diffusion models during the NeRF training process. In Section 4.2, we propose an uncertainty measure that accompanies rendered RGB images to enforce 3D consistency on the generated pseudo-observations. In Section 4.3, we introduce our proposed deceptive diffusion model, that can generate high-quality photorealistic pseudo-observations conditioned on coarse-rendered images and uncertainty maps.

4.1 Pipeline Overview

We start by training the NeRF/3DGS representation using the few-view input images. After a set number of epochs, we obtain a coarse reconstruction, from which we can generate RGB images, depth predictions, as well as uncertainty maps at novel views.

To synthesize pseudo-observation, we randomly sample novel views within a bounding box defined by the outermost input views. At novel viewpoint A,

we render RGB image $\hat{I}_A$ and depth map D_A and further compute an uncertainty map U_A. We defer describing the process of obtaining D_A and U_A until Section 4.2. We generate a pseudo-observation at viewpoint A, denoted as P_A, using our proposed deceptive diffusion model. Thanks to the natural image prior from the latent diffusion model and multi-view uncertainty measures, the pseudo-observations could eliminate the artifacts in $\hat{I}_A$. After accumulating a sufficient number of pseudo-observations, we continue training the NeRF representation, using the combination of both the original input images (real) and pseudo-observations (synthesized). We discard 20% of the pseudo-observations with the lowest perceptual similarity to input images, quantified through the LPIPS metric. This helps avoid generating content that is far from the original inputs.

We iterate the above processes, each time sampling new viewpoints to render images and generating corresponding pseudo-observations, enhancing the training image set until we accumulate more than a sufficient number of observations. In other words, we alternate between the two steps to get a final reconstruction: 1) feeding NeRF-rendered images into a diffusion model for generating pseudo-observations and 2) optimizing the NeRF representation using both the real input images and the newly generated pseudo-observations.

4.2 Uncertainty Measure for View-Consistency

To ensure consistency between the synthesized pseudo-observations and real input views, we further guide the diffusion model with an uncertainty measure. While there are approaches to quantify the uncertainty from feature space perspectives through ensemble learning [59], variational inference [53] or spatial perturbations [21], we instead acquire uncertainty maps through correspondence matching for rendered RGB images to explicitly enforce view consistency. Compared to [21, 53, 59], our proposed measure of uncertainty does not involve modifying the underlying representation and only requires reprojecting real input observations to novel views, thus minimizing additional computation while enforcing epipolar constraints. Given a rendered RGB image at a novel view, we leverage its associated rendered depth map to warp the image at its nearest input view to the unobserved viewpoint. Similar to Equation (2), we compute the depth value for each ray as a weighted composition of distances traveled from the origin as:

$$\hat{\mathbf{D}} = \sum_{k=1}^{K} \hat{T}(t_k)\alpha(\sigma(t_k)\delta_k). \tag{6}$$

At an arbitrary novel view A, after getting the rendered image $\hat{I}_A$, we use its associated rendered depth map D_A to warp the ground truth image I_{B} towards A, where B is its nearest input viewpoint. This results in a warped image

$$I_{B\to A} = \psi(I_{\mathrm{B}}; D_A, R_{B\to A}), \tag{7}$$

where $R_{B\to A}$ is the viewpoint difference between A and B. More specifically, each pixel location p in $\hat{I}_A$ is transformed to q in I_B as:

$$q = KR_{B\to A}D_A(p)K^{-1}p, \tag{8}$$

where K is the camera parameter matrix. We define the uncertainty map for the viewpoint A as the squares of the distance between $I_{B\to A}$ and $\hat{I}_A$:

$$U_A = (I_{B\to A} - \hat{I}_A)^2. \tag{9}$$

Our proposed deceptive diffusion model will condition on U_A to generate pseudo-observations with view-consistency guidance.

4.3 Deceptive Diffusion Model

We propose a 2D diffusion model G that conditions on a NeRF-rendered RGB image $\hat{I}_A$ along with its corresponding depth prediction D_A and uncertainty map U_A to generate a refined natural image as the pseudo-observation P_A:

$$P_A = \mathrm{G}(\hat{I}_A, D_A, U_A), \tag{10}$$

where G in essential rectifies renderings from NeRF/3DGS and is thus termed the deceptive diffusion model.

Our approach capitalizes on latent diffusion models [50], which leverages natural image priors derived from internet-scale data to help ameliorate unnaturalness caused by sparse input observations. Artifacts generated by NeRFs often float in empty space and are therefore highly conspicuous in depth prediction. Conditioning the diffusion model also on the depth prediction takes advantage of this. On the other hand, the uncertainty map introduced in Section 4.2 provides cross-view constraints for the generation of pseudo-observations.

To this end, given a dataset of quartets $\{(I_{\text{fine}}, I_{\text{coarse}}, D, U)\}$, we fine-tune a pre-trained diffusion model, consisting of 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:

$$\min_\theta \mathbb{E}_{z\sim\mathcal{E},t,\epsilon\sim\mathcal{N}(0,1)}\|\epsilon - \epsilon_\theta(z_t, t, c(I_{\text{coarse}}, D, U, s))\|_2^2, \tag{11}$$

where the diffusion time step $t \sim [1, 1000]$ and $c(I_{\text{coarse}}, D, U, s)$ is the embedding of the coarse RGB image, depth estimation, uncertainty map, and a text embedding s of the coarse image. To enable diffusion models to learn such specific input conditions without disrupting their prior for natural images, we leverage ControlNet [76] to efficiently implement the training paradigm discussed below while preserving the production-ready weights of pre-trained 2D diffusion models.

Text embedding. To derive text embedding s from the coarse rendered image, we first generate a text prompt s_0 using a pre-trained image captioning network. While image captioning reliably provides descriptive textual representations for most coarse images, its efficacy can diminish for images of lower

quality or those with pronounced artifacts. To counteract this, we adopt the textual inversion [20]. We optimize a shared latent text embedding s_* shared by all the input observations and coarse images. By concatenating the embeddings we formulate a composite feature $s = [s_0, s_*]$ that encapsulates both the semantic and visual attributes of the input image. This combined strategy not only ameliorates the shortcomings of image captioning but also ensures the stylistic congruence of the generated pseudo-observations with the input images.

Data augmentation. To train the deceptive diffusion model to generate pseudo-observations from the same viewpoint's rendered RGB image, depth map, and uncertainty map, we need to construct a dataset of quartets $(I_{\text{fine}}, I_{\text{coarse}}, D, U)$. I_{fine} is a fine image with minimal artifacts, closely resembling a real image or a perfectly rendered NeRF image. I_{coarse} is a coarse image from the same viewpoint, likely presenting obvious artifacts, while D and U represent the associated depth map and uncertainty map, respectively. We can construct such a dataset 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 a small subset of input images. By rendering from the same viewpoint, such a coarse-fine NeRF duo can render paired training data samples, as well as depth and uncertainty maps. However, due to limited computational resources, we cannot afford to conduct NeRF duos training across a plethora of scenarios. Therefore, we introduce a data augmentation paradigm to mitigate the computational cost associated with preparing training data. Rather than exclusively relying on image pairs derived from NeRF duos, we exploit a more straightforward data source during the initial phase of training. We add random Gaussian noise to RGB images, utilizing these noisy images and accompanying depth maps as training inputs, while retaining the original RGB images as the training objectives. In this manner, we can readily obtain training samples by simply pairing RGB-depth data. Despite the gap between the data obtained in this manner and the data previously discussed from NeRF duos, we find that they effectively enable diffusion models to better utilize depth maps to understand imperfect RGB images. (When training with such additional data, we do not feed uncertainty maps into the diffusion model.)

5 Experimental Results

In this section, we assess the performance of our proposed Deceptive-NeRF for the task of few-shot novel view synthesis across a range of challenging datasets, comparing it with state-of-the-art methods. We introduce the setup of our experiments in Section 5.1. To demonstrate our method's capability for super-resolution on novel views, we showcase experimental results under the challenging setting of few views and low resolution in Section 5.3. To elucidate the effectiveness of our design, we conduct a series of evaluations of our approach, including analyses of the input densification scale and generated pseudo-observations, as well as an ablation study of other designs and components in Section 5.4.

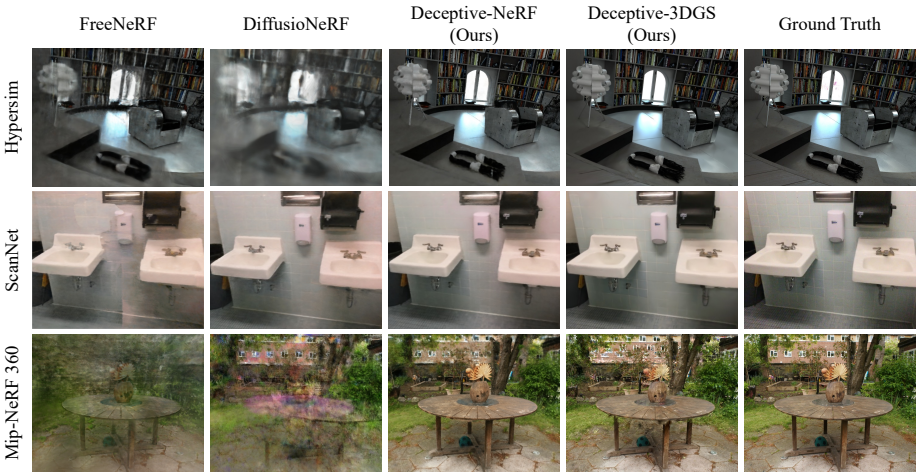

Fig. 3: Qualitative comparisons of few view reconstruction. Scenes are from the Hypersim [47], Scannet [13] and mip-NeRF 360 [4] datasets, with 10 input views.

Table 1: Quantitative comparison on Hypersim. best second-best third-best

Method	PSNR(↑)			SSIM(↑)			LPIPS(↓)		
	5-view	10-view	20-view	5-view	10-view	20-view	5-view	10-view	20-view
PixelNeRF	7.76	8.31	10.90	0.221	0.380	0.374	0.542	0.571	0.503
MVSNeRF	11.58	12.00	14.42	0.271	0.274	0.315	0.563	0.512	0.457
DS-NeRF	13.79	13.66	18.80	0.388	0.431	0.488	0.515	0.511	0.481
DietNeRF	13.01	13.51	18.62	0.417	0.479	0.481	0.541	0.527	0.472
RegNeRF	15.65	18.59	19.26	0.491	0.501	0.519	0.516	0.451	0.362
DiffusioNeRF	16.40	17.22	19.88	0.451	0.470	0.656	0.432	0.404	0.416
FlipNeRF	15.43	17.47	19.36	0.456	0.569	0.585	0.350	0.415	0.312
FreeNeRF	17.20	18.06	20.20	0.599	0.671	0.706	0.431	0.286	0.237
Deceptive-NeRF (Ours)	18.91	19.88	21.23	0.652	0.726	0.761	0.322	0.270	0.222
Deceptive-3DGS (Ours)	19.31	21.45	21.61	0.728	0.774	0.788	0.265	0.241	0.228

5.1 Experimental Setup.

Implementation Details. We train our deceptive diffusion model on a mixture of Hypersim [47] and CO3D [46]. For training data augmentation, we corrupt 60,000 images from Hypersim [47] by adding additive Gaussian noise with a standard deviation of 0.3. We use these noisy images and their depth maps as training input and the original images as training targets, without running NeRF reconstruction. For the major section of the training dataset for our deceptive diffusion model, we train coarse and fine NeRF duos for the same scenes, where coarse NeRFs are trained only with one-fifth of input images. Coarse NeRFs render RGB images and depth maps as training inputs while fine NeRFs render fine RGB images from the same viewpoints as training targets. We use a total of 200 scenes and approximately 20,000 images for the data generation of this stage, with Hypersim [47] and CO3D [46] each contributing half of them. We fine-tune a pre-trained Stable Diffusion model with ControlNet [76] into our deceptive diffusion model. We set five (3 for coarse RGB, 1 for depth, and 1 for uncertainty) control map channels to match the inputs and use all default

Table 2: Quantitative comparison on LLFF. best second-best third-best

Method	PSNR(↑)			SSIM(↑)			LPIPS(↓)		
	3-view	6-view	9-view	3-view	6-view	9-view	3-view	6-view	9-view
SRF	17.07	16.75	17.39	0.436	0.438	0.465	0.529	0.521	0.503
PixelNeRF	16.17	17.03	18.92	0.438	0.473	0.535	0.512	0.477	0.430
MVSNeRF	17.88	19.99	20.47	0.584	0.660	0.695	0.327	0.264	0.244
DietNeRF	14.94	21.75	24.28	0.370	0.717	0.801	0.496	0.248	0.183
RegNeRF	18.84	23.22	24.88	0.573	0.770	0.826	0.345	0.203	0.159
FreeNeRF	19.63	23.73	25.13	0.612	0.779	0.827	0.308	0.195	0.160
Deceptive-NeRF (Ours)	19.87	23.92	25.34	0.688	0.810	0.834	0.304	0.190	0.155
Deceptive-3DGS (Ours)	19.95	24.15	25.30	0.665	0.795	0.838	0.295	0.177	0.150

parameters for the fine-tuning task. We randomly crop all images to a resolution of 512×512.

Datasets. To validate the performance and generalizability of our method for sparse view reconstruction, we employed a series of datasets with distinct characteristics, including Hypersim [47], a synthetic indoor dataset (from which we avoided scenes already used in training the deceptive diffusion model), ScanNet [13], a real-world indoor dataset, LLFF [39], a real-world facing-forward dataset, and Mip-NeRF 360 [4], a real-world 360-degree outdoor dataset.

Evaluation Metrics. We evaluate using the following standard metrics: (i) Peak Signal-to-Noise Ratio (PSNR), (ii) Structural Similarity Index Measure (SSIM) [67] and (iii) Learned Perceptual Image Patch Similarity (LPIPS) [77], by comparing the reconstructed frames against ground truth images. These metrics are computed on the held-out view and averaged across all frames.

Baselines. We compare our method with several methods within a similar scope. PixelNeRF [73], MVSNeRF [9], and SRF [12] are representative pre-trained methods, exploiting the DTU and LLFF datasets for pre-training. We also compare our approach against diverse regularization approaches, including DS-NeRF [15], DietNeRF [25], RegNeRF [42], DiffusioNeRF [69], FlipNeRF [51], and FreeNeRF [71].

5.2 Comparisons of Few View Reconstruction

We show the qualitative results in Figure 3 and report the quantitative comparisons in Table 1 and Table 2, with more results showcased in the supplementary material. On challenging datasets such as those featuring indoor scenes facing outwards or 360-degree outdoor scenes, existing state-of-the-art methods can easily generate discernible artifacts. By utilizing pseudo-observations to densify inputs, our method avoids such artifacts produced by overfitting. Both Deceptive-NeRF and Deceptive-3DGS surpass the baseline methods across all metrics, demonstrating high-quality reconstruction under various challenging scenes and with different numbers of inputs.

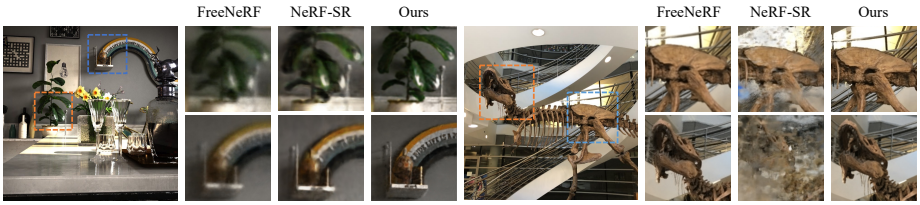

Fig. 4: Super-Resolution Capabilities. Our Deceptive-NeRF, when applied to the Hypersim and LLFF datasets with input images downsampled by a factor of 4, demonstrates exceptional super-resolution performance, capable of recovering intricate details.

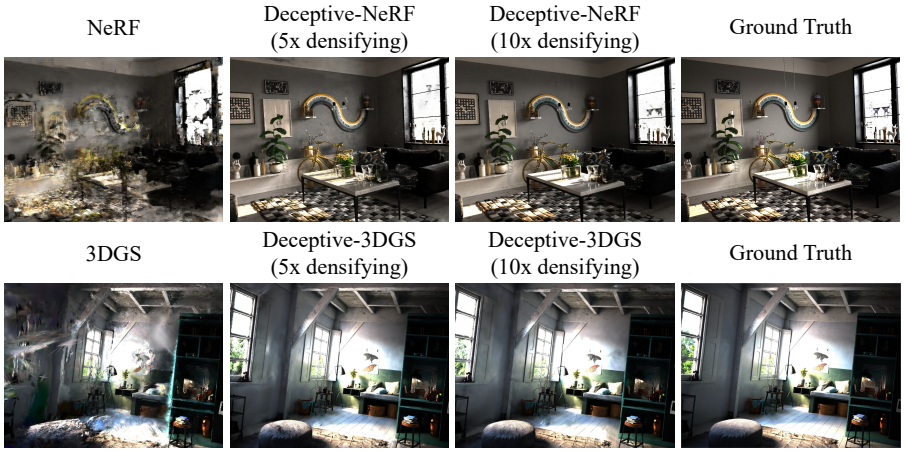

Fig. 5: Evaluation of Densification scales. We run our method with observations densified by varying scales on Hypersim [47] with 10 input views. Vanilla NeRF or 3DGS results in severe artifacts; fivefold densification reduces them but lacks detail restoration, whereas tenfold densification achieves high-quality reconstructions.

5.3 Comparisons of Novel View Super-Resolution

Due to the generation of high-quality pseudo-observations, our method is capable of super-resolving synthesized novel views under low-resolution and sparse-view conditions. We train Deceptive-NeRF on the Hypersim dataset [47] with 10 input views and the LLFF dataset [39] with 6 input views. All input images were downsampled by a factor of 4 and rendered in multiple views at the original resolution. As shown in Figure 4, our results can achieve high-quality super-resolution in novel views and recover fine details such as the glossiness of plant leaves and the texture of dinosaur skeletons. We compared Deceptive-NeRF with the few-view reconstruction approach FreeNeRF [71] and the NeRF super-resolution method NeRF-SR [64]. FreeNeRF, due to not incorporating 2D diffusion priors, fails to restore complex object details, while NeRF-SR suffers from performance degradation due to sparse observations and creates artifacts.

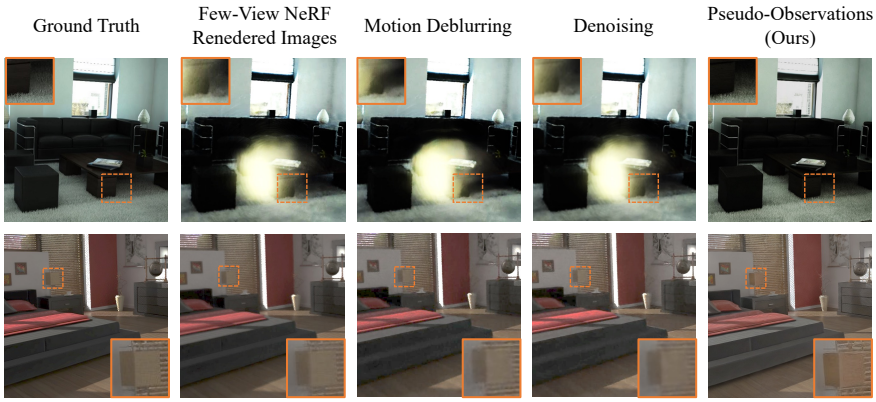

Fig. 6: Visualization of pseudo-observations. Pseudo-observations synthesized by our deceptive diffusion model remove floating artifacts and blurriness caused by sparse observation in coarse NeRFs, which cannot be achieved by image restoration models.

5.4 Evaluations

Densification scales. To study the impact of the densification scale on the results, we conducted experiments with observations densified by varying scales. We carry out these experiments using the Hypersim dataset [47] with 10 input views. As illustrated in Figure 5, not performing densification (vanilla NeRF or 3DGS) leads to severe artifacts while densifying fivefold significantly reduces artifacts but still falls short of high-quality detail restoration. Densifying tenfold, however, results in high-quality reconstructions. Please refer to the appendix for related quantitative results.

Pseudo-observations. To better understand our generated pseudo-observations and validate the effectiveness of our deceptive diffusion model, we compare them with images processed using an image restoration model [74]. Figure 6 shows pseudo-observations synthesized by our deceptive diffusion model effectively remove floating artifacts caused by sparse observation in coarse NeRFs and also mitigate blurriness as well. In contrast, image restoration models (motion deblurring and denoising) fail to convert coarse NeRF images into reasonable pseudo-observations. This is primarily because they are designed and trained for specific image restoration tasks, not for generating pseudo-observations.

Ablation studies. We conduct ablation studies on the following design choices using the Hypersim dataset [47] under the 20-view setting:

1. **Depth Conditioning.** Our deceptive diffusion model generates pseudo-observations conditioned on rendered depth maps. To gauge the significance of this choice, we train a variant that solely conditions on raw RGB images for generating pseudo-observations.
2. **Data Augmentation.** We evaluate the impact of our data augmentation procedure when training our deceptive diffusion model. Specifically, we train the model without the initial stage and rely solely on coarse-fine NeRF pairs to generate training samples.

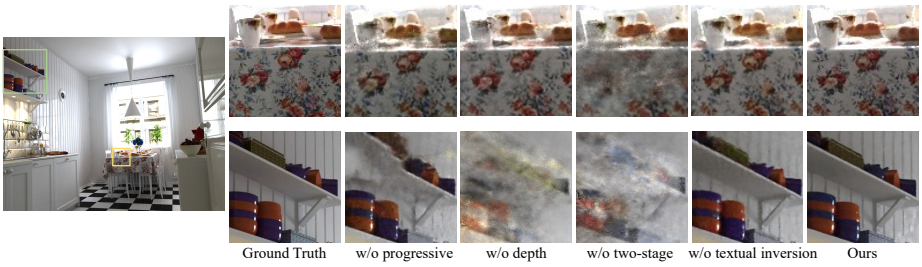

Fig. 7: Qualitative ablation studies. Our full model synthesizes novel views with the highest quality, while variants tend to produce floating artifacts and blurry details.

Table 3: Quantitative ablation studies. We ablate our design choices on the Hypersim dataset [47] with 20 input views. best second-best third-best

Progressive	Depth	Two-stage	Captioning	Inversion	PSNR (↑)	SSIM (↑)	LPIPS (↓)
	✓	✓	✓	✓	19.90	0.555	0.358
✓		✓	✓	✓	18.79	0.489	0.352
✓	✓		✓	✓	20.49	0.619	0.290
✓	✓	✓		✓	21.59	0.758	0.236
✓	✓	✓	✓		20.58	0.744	0.239
✓	✓	✓	✓	✓	22.41	0.812	0.202

3. **Text Embedding.** Our approach integrates both image captioning and textual inversion to address severely artifacted images while ensuring stylistic consistency. We test two variants of our model, one without image captioning and the other without textual inversion.

As illustrated in Figure 7 and Table 3, our complete model synthesizes the most photorealistic novel views and outperforms other methods in all quantitative metrics.

6 Conclusion

We introduce Deceptive-NeRF and Deceptive-3DGS, which utilize 2D diffusion models to generate plausible and consistent pseudo-observations, thereby enhancing sparse view reconstruction by densifying input observations. This fundamentally differs from recent methods that employ 2D diffusion priors to introduce regularization to NeRF training for sparse inputs, by significantly reducing computational costs. Our proposed deceptive diffusion model, guided by an uncertainty measure, can generate high-quality, photorealistic pseudo-observations. Experimental evidence demonstrates that our method can generate pseudo-observations of high quality for synthesizing novel views, surpassing current state-of-the-art approaches. Moreover, our method is capable of rendering novel views with super-resolution, effectively overcoming the challenges posed by low-quality inputs.

Acknowledgements This work was supported in part by Dartmouth College A&S Startup fund.

References

1. Attal, B., Huang, J.B., Richardt, C., Zollhoefer, M., Kopf, J., O'Toole, M., Kim, C.: HyperReel: High-fidelity 6-DoF video with ray-conditioned sampling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition(2023)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: IEEE/CVF International Conference on Computer Vision. pp. 5855–5864 (2021)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5470–5479 (2022)
4. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. IEEE/CVF Conference on Computer Vision and Pattern Recognition(2022)
5. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Zip-nerf: Anti-aliased grid-based neural radiance fields. IEEE/CVF International Conference on Computer Vision(2023)
6. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 130–141 (2023)
7. Chan, E.R., Nagano, K., Chan, M.A., Bergman, A.W., Park, J.J., Levy, A., Aittala, M., Mello, S.D., Karras, T., Wetzstein, G.: GeNVS: Generative novel view synthesis with 3D-aware diffusion models. In: arXiv (2023)
8. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: European Conference on Computer Vision. pp. 333–350. Springer (2022)
9. Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., Su, H.: Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In: IEEE/CVF International Conference on Computer Vision. pp. 14124–14133 (2021)
10. Chen, H., Gu, J., Chen, A., Tian, W., Tu, Z., Liu, L., Su, H.: Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In: arXiv preprint arXiv:2304.06714 (2023)
11. Chen, Y., Xu, H., Wu, Q., Zheng, C., Cham, T.J., Cai, J.: Explicit correspondence matching for generalizable neural radiance fields. In: arXiv preprint arXiv:2304.12294 (2023)
12. Chibane, J., Bansal, A., Lazova, V., Pons-Moll, G.: Stereo radiance fields (srf): Learning view synthesis from sparse views of novel scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE (jun 2021)
13. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5828–5839 (2017)
14. Deng, C., Jiang, C., Qi, C.R., Yan, X., Zhou, Y., Guibas, L., Anguelov, D., et al.: Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In: arXiv preprint arXiv:2212.03267 (2022)
15. Deng, K., Liu, A., Zhu, J.Y., Ramanan, D.: Depth-supervised nerf: Fewer views and faster training for free. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12882–12891 (2022)
16. Du, Y., Smith, C., Tewari, A., Sitzmann, V.: Learning to render novel views from wide-baseline stereo pairs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition(2023)

17. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12479–12488 (2023)
18. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: IEEE/CVF International Conference on Computer Vision. pp. 5501–5510 (2022)
19. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5501–5510 (2022)
20. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: arXiv preprint arXiv:2208.01618 (2022)
21. Goli, L., Reading, C., Selllán, S., Jacobson, A., Tagliasacchi, A.: Bayes' rays: Uncertainty quantification for neural radiance fields. arXiv preprint arXiv:2309.03185 (2023)
22. Gu, J., Trevithick, A., Lin, K.E., Susskind, J., Theobalt, C., Liu, L., Ramamoorthi, R.: Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In: International Conference on Machine Learning(2023)
23. Guangcong, Chen, Z., Loy, C.C., Liu, Z.: Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. IEEE/CVF International Conference on Computer Vision(2023)
24. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
25. Jain, A., Tancik, M., Abbeel, P.: Putting nerf on a diet: Semantically consistent few-shot view synthesis. In: IEEE/CVF International Conference on Computer Vision. pp. 5885–5894 (2021)
26. Jang, W., Agapito, L.: Codenerf: Disentangled neural radiance fields for object categories. In: IEEE/CVF International Conference on Computer Vision. pp. 12949–12958 (2021)
27. Johari, M.M., Lepoittevin, Y., Fleuret, F.: Geonerf: Generalizing nerf with geometry priors. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18365–18375 (2022)
28. Karnewar, A., Vedaldi, A., Novotny, D., Mitra, N.: Holodiffusion: Training a 3D diffusion model using 2D images. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (2023)
29. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics**42**(4) (2023)
30. Kim, M., Seo, S., Han, B.: Infonerf: Ray entropy minimization for few-shot neural volume rendering. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition(2022)
31. Kobayashi, S., Matsumoto, E., Sitzmann, V.: Decomposing nerf for editing via feature field distillation. In: Advances in Neural Information Processing Systems(2022)
32. Kwak, M.S., Song, J., Kim, S.: Geconerf: Few-shot neural radiance fields via geometric consistency. In: International Conference on Machine Learning(2023)
33. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition(2022)
34. Liu, R., Wu, R., Hoorick, B.V., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition(2023)

35. Liu, S., Zhang, X., Zhang, Z., Zhang, R., Zhu, J.Y., Russell, B.: Editing conditional radiance fields. In: IEEE/CVF International Conference on Computer Vision(2021)
36. Liu, X., Chen, J., Yu, H., Tai, Y.W., Tang, C.K.: Unsupervised multi-view object segmentation using radiance field propagation. Advances in Neural Information Processing Systems**35**, 17730–17743 (2022)
37. Liu, X., Tai, Y.W., Tang, C.K., Miraldo, P., Lohit, S., Chatterjee, M.: Gear-nerf: Free-viewpoint rendering and tracking with motion-aware spatio-temporal sampling. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19667–19679 (2024)
38. Melas-Kyriazi, L., Rupprecht, C., Laina, I., Vedaldi, A.: Realfusion: 360° reconstruction of any object from a single image. In: arXiv preprint arXiv:2302.10663 (2023)
39. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics**38**(4), 1–14 (2019)
40. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision(2020)
41. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Transactions on Graphics**41**(4), 1–15 (2022)
42. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5480–5490 (2022)
43. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: IEEE/CVF International Conference on Computer Vision(2021)
44. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. In: International Conference on Learning Representations(2022)
45. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021)
46. Reizenstein, J., Shapovalov, R., Henzler, P., Sbordone, L., Labatut, P., Novotny, D.: Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In: IEEE/CVF International Conference on Computer Vision. pp. 10901–10911 (2021)
47. Roberts, M., Ramapuram, J., Ranjan, A., Kumar, A., Bautista, M.A., Paczan, N., Webb, R., Susskind, J.M.: Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10912–10922 (2021)
48. Roessle, B., Barron, J.T., Mildenhall, B., Srinivasan, P.P., Nießner, M.: Dense depth priors for neural radiance fields from sparse input views. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition(June 2022)
49. Roessle, B., Müller, N., Porzi, L., Bulò, S.R., Kontschieder, P., Nießner, M.: Ganerf: Leveraging discriminators to optimize neural radiance fields. ACM Transactions on Graphics(2023)
50. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10684–10695 (June 2022)

51. Seo, S., Chang, Y., Kwak, N.: Flipnerf: Flipped reflection rays for few-shot novel view synthesis. IEEE/CVF International Conference on Computer Vision(2023)
52. Seo, S., Han, D., Chang, Y., Kwak, N.: Mixnerf: Modeling a ray with mixture density for novel view synthesis from sparse inputs. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20659–20668 (June 2023)
53. Shen, J., Agudo, A., Moreno-Noguer, F., Ruiz, A.: Conditional-flow nerf: Accurate 3d modelling with reliable uncertainty quantification. In: European Conference on Computer Vision. pp. 540–557. Springer (2022)
54. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265 (2015)
55. Somraj, N., Karanayil, A., Soundararajan, R.: Simplenerf: Regularizing sparse input neural radiance fields with simpler solutions. In: ACM Transactions on Graphics. pp. 1–11 (2023)
56. Somraj, N., Soundararajan, R.: ViP-NeRF: Visibility prior for sparse input neural radiance fields. In: ACM Transactions on Graphics(August 2023)
57. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. International Conference on Learning Representations(2020)
58. Sun, C., Sun, M., Chen, H.T.: Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5459–5469 (2022)
59. Sünderhauf, N., Abou-Chakra, J., Miller, D.: Density-aware nerf ensembles: Quantifying predictive uncertainty in neural radiance fields. In: IEEE International Conference on Robotics and Automation. pp. 9370–9376. IEEE (2023)
60. Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Tretschk, E., Yifan, W., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S., et al.: Advances in neural rendering. In: Computer Graphics Forum. pp. 703–735. Wiley Online Library (2022)
61. Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Lassner, C., Theobalt, C.: 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. pp. 12959–12970 (2021)
62. Truong, P., Rakotosaona, M.J., Manhardt, F., Tombari, F.: SPARF: Neural radiance fields from sparse and noisy poses. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition(2023)
63. Wang, C., Chai, M., He, M., Chen, D., Liao, J.: Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3835–3844 (2022)
64. Wang, C., Wu, X., Guo, Y.C., Zhang, S.H., Tai, Y.W., Hu, S.M.: Nerf-sr: High quality neural radiance fields using supersampling. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 6445–6454 (2022)
65. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12619–12629 (2023)
66. Wang, L., Zhang, J., Liu, X., Zhao, F., Zhang, Y., Zhang, Y., Wu, M., Yu, J., Xu, L.: Fourier plenoctrees for dynamic radiance field rendering in real-time. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13524–13534 (2022)
67. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing**13**(4), 600–612 (2004)

68. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., Holynski, A.: Reconfusion: 3d reconstruction with diffusion priors. arXiv (2023)
69. Wynn, J., Turmukhambetov, D.: DiffusioNeRF: Regularizing neural radiance fields with denoising diffusion models. In: arxiv (2023)
70. Xu, D., Jiang, Y., Wang, P., Fan, Z., Wang, Y., Wang, Z.: 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(2022)
71. Yang, J., Pavone, M., Wang, Y.: FreeNeRF: Improving few-shot neural rendering with free frequency regularization. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition(2023)
72. Yu, A., Li, R., Tancik, M., Li, H., Ng, R., Kanazawa, A.: Plenoctrees for real-time rendering of neural radiance fields. In: IEEE/CVF International Conference on Computer Vision. pp. 5752–5761 (2021)
73. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelNeRF: Neural radiance fields from one or few images. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition(2021)
74. Zamir, S.W., Arora, A., Khan, S., Hayat, M., Khan, F.S., Yang, M.H.: Restormer: Efficient transformer for high-resolution image restoration. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5728–5739 (2022)
75. Zhang, J., Liu, X., Ye, X., Zhao, F., Zhang, Y., Wu, M., Zhang, Y., Xu, L., Yu, J.: Editable free-viewpoint video using a layered neural representation. ACM Transactions on Graphics**40**(4), 1–18 (2021)
76. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: arXiv preprint arXiv:2302.05543 (2023)
77. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
78. Zhou, Z., Tulsiani, S.: Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition(2022)