

# DiffRF: Rendering-Guided 3D Radiance Field Diffusion

Norman Müller<sup>1,2</sup> Yawar Siddiqui<sup>1,2</sup> Lorenzo Porzi<sup>2</sup> Samuel Rota Bulò<sup>2</sup>  
Peter Kortscheder<sup>2</sup> Matthias Nießner<sup>1</sup>

Technical University of Munich<sup>1</sup> Meta Reality Labs Zurich<sup>2</sup>

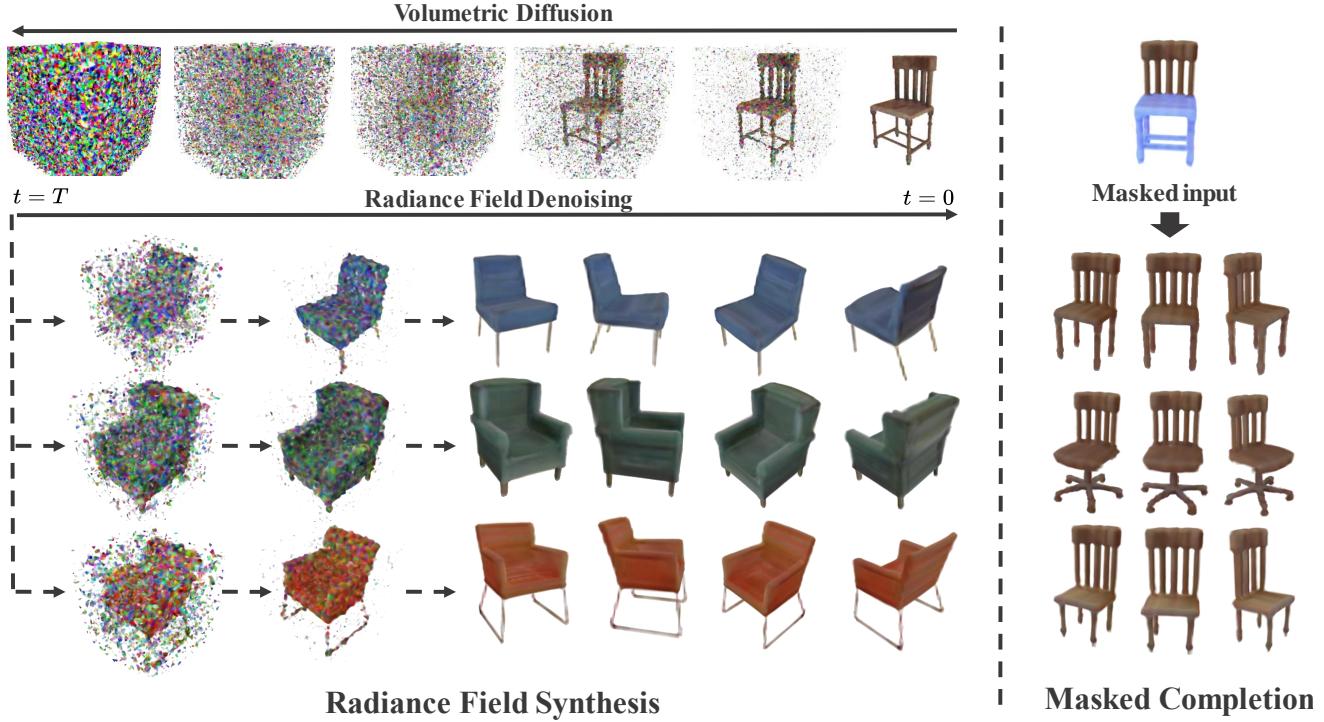


Figure 1. Our method performs denoising of a probabilistic diffusion process applied to 3D radiance fields. Guided by 3D supervision and volumetric rendering, our model enables unconditional synthesis of high-fidelity 3D assets (left). We further introduce the novel application of *masked completion* (right), i.e., the task of recovering shape and appearance from incomplete objects (highlighted in light-blue on the top right chair), solved by our model as conditional inference without task-specific training.

## Abstract

We introduce DiffRF, a novel approach for 3D radiance field synthesis based on denoising diffusion probabilistic models. While existing diffusion-based methods operate on images, latent codes, or point cloud data, we are the first to directly generate volumetric radiance fields. To this end, we propose a 3D denoising model which directly operates on an explicit voxel grid representation. However, as radiance fields generated from a set of posed images can be ambiguous and contain artifacts, obtaining ground truth radiance field samples is non-trivial. We address this challenge by pairing the denoising formulation with a rendering loss, enabling our model to learn a deviated prior that favours good image quality instead of trying to replicate fitting er-

rors like floating artifacts. In contrast to 2D-diffusion models, our model learns multi-view consistent priors, enabling free-view synthesis and accurate shape generation. Compared to 3D GANs, our diffusion-based approach naturally enables conditional generation such as masked completion or single-view 3D synthesis at inference time.

## 1. Introduction

In recent years, Neural Radiance Fields (NeRFs) [34] have emerged as a powerful representation for fitting individual 3D scenes from posed 2D input images. The ability to photo-realistically synthesize novel views from arbitrary viewpoints while respecting the underlying 3D scene geom-

try has the potential to disrupt and transform applications like AR/VR, gaming, mapping, navigation, *etc.* A number of recent works have introduced extensions for making NeRFs more sophisticated, by *e.g.*, showing how to incorporate scene semantics [15, 27], training models from heterogeneous data sources [32], or scaling them up to represent large-scale scenes [57, 59]. These advances are testament to the versatility of ML-based scene representations; however, they still fit to specific, individual scenes rather than generalizing beyond their input training data.

In contrast, neural field representations that generalize to multiple object categories or learn priors for scenes across datasets appear much more limited to date, despite enabling applications like single-image 3D object generation [5, 36, 46, 60, 65] and unconstrained scene exploration [13]. These methods explore ways to disentangle object priors into shape and appearance-based components, or to decompose radiance fields into several small and locally-conditioned radiance fields to improve scene generation quality; however, their results still leave significant gaps w.r.t. photorealism and geometric accuracy.

Directions involving generative adversarial networks (GANs) that have been extended from the 2D domain to 3D-aware neural fields generation are demonstrating impressive synthesis results [6]. Like regular 2D GANs, the training objective is based on discriminating 2D images, which are obtained by rendering synthesized 3D radiance fields.

At the same time, diffusion-based models [48] have recently taken the computer vision research community by storm, performing on-par or even surpassing GANs on multiple 2D benchmarks, and are producing photo-realistic images that are almost indistinguishable from real photographs. For multi-modal or conditional settings such as text-to-image synthesis, we currently observe unprecedented output quality and diversity from diffusion-based approaches. While several works address purely geometric representations [30, 68], lifting the denoising-diffusion formulation directly to 3D volumetric radiance fields remains challenging. The main reason lies in the nature of diffusion models, which require a one-to-one mapping between the noise vector and the corresponding ground truth data samples. In the context of radiance fields, such volumetric ground truth data is practically infeasible to obtain, since even running a costly per-sample NeRF optimization results in incomplete and imperfect radiance field reconstructions.

In this work, we present the first diffusion-based generative model that directly synthesizes 3D radiance fields, thus unlocking high-quality 3D asset generation for both shape and appearance. Our goal is to learn such a generative model trained across objects, where each sample is given by a set of posed RGB images.

To this end, we propose a 3D denoising model directly operating on an explicit voxel grid representation (Fig. 1,

left) producing high-frequency noise estimates. To address the ambiguous and imperfect radiance field representation for each training sample, we propose to bias the noise prediction formulation from Denoising Diffusion Probabilistic Models (DDPMs) towards synthesizing higher image quality by an additional volumetric rendering loss on the estimates. This enables our method to learn radiance field priors less prone to fitting artifacts or noise accumulation during the sampling process. We show that our formulation leads to diverse and geometrically-accurate radiance field synthesis producing efficient, realistic, and view-consistent renderings. Our learned diffusion prior can be applied in an unconditional setting where 3D object synthesis is obtained in a multi-view consistent way, generating highly-accurate 3D shapes and allowing for free-view synthesis. We further introduce the new task of *conditional masked completion* – analog to shape completion – for radiance field completion at inference time. In this setting, we allow for realistic 3D completion of partially-masked objects without the need for task-specific model adaptation or training (see Fig. 1, right).

We summarize our contributions as follows:

- To the best of our knowledge, we introduce the first diffusion model to operate directly on 3D radiance fields, enabling high-quality, truthful 3D geometry and image synthesis.
- We introduce the novel application of 3D radiance field masked completion, which can be interpreted as a natural extension of image inpainting to the volumetric domain.
- We show compelling results in unconditional and conditional settings, *e.g.*, by improving over GAN-based approaches on image quality (from 27.03 to 25.64 in FID) and geometry synthesis (improving MMD from 5.86 to 4.26), on the challenging PhotoShape Chairs dataset [41].

## 2. Related work

**Diffusion models.** Since the seminal work by Sohl-Dickstein *et al.* [48] on generative diffusion modeling, two classes of generative models have been proposed that perform inversion of a diffusion process: Denoising Score Matching (DSM) [49, 52, 53] and Denoising Diffusion Probabilistic Models (DDPMs) [20]. Both approaches have shown to be flavours of a single framework, coined Score SDE, in the work of Song *et al.* [54]. The term “Diffusion models” is now being used as an all-encompassing name for this constellation of methods. Several main directions being studied include devising different sampling schemes [25, 50] and noising models [12, 22], exploring alternative formulations and training algorithms [23, 38, 51], and improving efficiency [44]. An overview of current research results is given by Karras *et al.* [25].

Diffusion models have been used to obtain state-of-the-art results in many domains such as text-to-image and guided synthesis [33, 37, 42, 44], 3D shape generation [4, 30, 67, 68], molecule prediction [31, 58, 62], and video generation [21, 64]. Interestingly, diffusion models have shown to outperform Generative Adversarial Networks (GANs) in high-resolution image generation tasks [14, 44], achieving unprecedented results in conditional image generation [43]. Furthermore, compared to GANs, which are often prone to divergence and mode collapse [3, 35], diffusion models have been observed to be much easier to train, although training time is still relatively long.

**3D generation.** Initially developed for 2D image synthesis, adversarial approaches have also found success in 3D, *e.g.*, for generating meshes [16, 61], 3D textures [47], voxelized representations [8, 18], or Neural Radiance Fields (NeRFs) [5, 6, 17, 39, 46, 69]. In particular, methods in this last category have received much attention in recent years, as they can be trained purely from collections of 2D images, without any form of 3D supervision, and enable for the first time photo-realistic novel view synthesis of the generated 3D objects. Pi-GAN [5] and GRAF [46] propose similar approaches where a standard NeRF [34] model is cast in a GAN setting by adding a form of stochastic conditioning, trained with an adversarial loss. These approaches are partly limited by the high training-time memory cost of NeRF-style volumetric rendering, forcing them to use low-resolution image patches. CIPS-3D [69] and GIRAFFE [39] solve this issue by letting the volume rendering component output a low-resolution 2D feature map, which is then upsampled by an efficient convolutional network to produce the final image. This approach drastically improves the quality and resolution of the rendered images, but also introduces 3D inconsistencies, as the convolutional stage can process different views of the same object in arbitrarily different ways. StyleNeRF [17] partially addresses this problem by carefully designing the convolutional stage to minimize inconsistencies, while EG3D [6] further improves on training efficiency by replacing the MLP-based NeRF with a light-weight tri-plane volumetric model generated by a convolutional network. In contrast to our method, these GAN-based approaches do not naturally support conditional synthesis or completion.

Compared to GANs, diffusion models are relatively under-explored as a tool for 3D synthesis, but a few works have emerged in the past two years. Some diffusion based-generators have been proposed for 3D point clouds [4, 30, 67, 68], showing promising results for conditional synthesis, completion, and other related tasks [67, 68]. DreamFusion [42] and GAUDI [2], like our work, employ diffusion models in conjunction with radiance fields, with applications in both conditional (on text and images) and unconditional 3D generation. DreamFusion [42] presents an algo-

rithm to generate NeRFs, augmented with an illumination component, by optimizing a loss defined by a pre-trained 2D text-conditional diffusion model. GAUDI [2] builds a 3D scene generator by first training a conditional NeRF to reconstruct a set of indoor videos given scene-specific latents, and then fitting a diffusion model to capture the learned latent space. In contrast, our diffusion model operates directly in the space of radiance fields, learning strong volumetric priors which can be used for 3D-conditional tasks like shape completion (see Sec. 4.2).

### 3. Method

Our method consists of a generative model for 3D objects that builds on recent state-of-the-art diffusion probabilistic models [20]. It is trained to revert a process that gradually corrupts 3D objects by injecting noise at different scales. In our case, 3D objects are represented as radiance fields [34], so the learned denoising process allows our method to generate object radiance fields from noise.

Since we target the generation of 3D objects as radiance fields, we begin with a brief overview of this representation before delving into the details of our method.

#### 3.1. Radiance Fields

A radiance field  $(\sigma, \xi)$  is an implicit, volumetric representation of a 3D object that is given in terms of a density field  $\sigma : \mathcal{X} \rightarrow \mathbb{R}_+$  and an RGB color field  $\xi : \mathcal{X} \rightarrow \mathbb{R}^3$  defined over a 3D domain  $\mathcal{X} \subset \mathbb{R}^3$ .<sup>1</sup> The density field gives information about the presence of an object in a specific point in space, whereas the color field provides the eventual corresponding RGB color.

Following a ray casting logic, the radiance field can be rendered along a given ray  $r$  yielding an RGB color  $c_r$  with the following equation [34]

$$c_r := \int_0^\infty \tau_s(r) \sigma(r_s) \xi(r_s) ds, \quad (1)$$

where a ray  $r$  is a linear curve parametrized by  $s$  with unit velocity,  $r_s \in \mathcal{X}$  denotes the point along the ray at  $s$ , and  $\tau_s(r)$  is the transmittance probability at  $s$ , which is given by

$$\tau_s(r) := \exp \left( - \int_0^s \sigma(s') ds' \right). \quad (2)$$

Using the ray rendering equation above, we can render the radiance field from any given camera, yielding an image of the novel view. It is sufficient to turn the camera into a proper set of rays expressed in world coordinates.

There exist various ways of implementing a radiance field, ranging from neural networks [34] to explicit voxel

<sup>1</sup>Typically the color field spans also the viewing direction, but we omit it in this work.

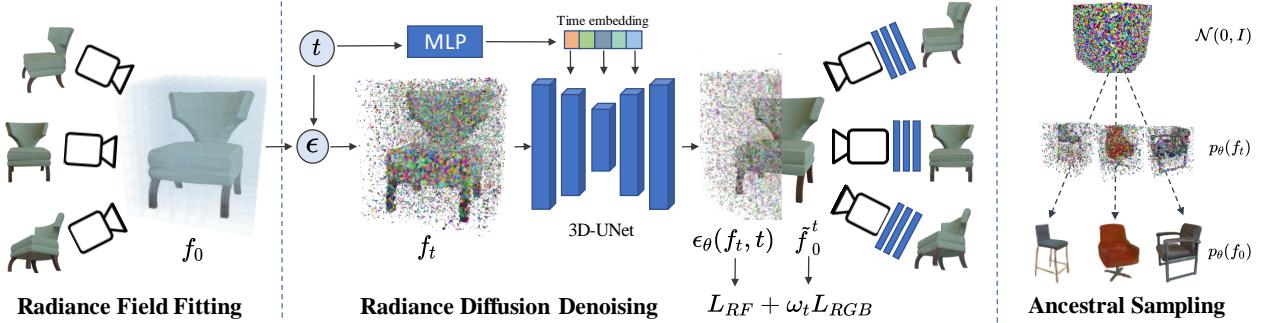


Figure 2. For a time step  $t$  uniformly sampled from  $1, \dots, T$ , we first diffuse an initial radiance field  $f_0$  according to a fixed noising schedule. The resulting  $f_t$  is passed through a time-conditioned 3D-UNet, giving an estimate of the applied noise  $\epsilon$ . We guide the model by the noise prediction loss  $L_{RF}$  as well as a rendering loss  $L_{RGB}$  on the predicted denoising  $\tilde{f}_0$ .

grids [24, 55]. In this work, we opt for the latter, since it enables good rendering quality along with faster training and inference. The explicit grid can be queried at continuous positions via bilinear interpolation of the voxel vertices.

Under the explicit representation, radiance fields become 4D tensors, where the first three dimensions index a grid spanning  $\mathcal{X}$ , whereas the last dimension indexes the density and color channels.

### 3.2. Generating Radiance Fields

Following recent advancements in the context of denoising-based generative methods [20], we formulate our generative model of radiance fields as a denoising diffusion probabilistic model.

**Generation process.** The generation (*a.k.a.* denoising) process is governed by a discrete-time Markov chain defined on the state space  $\mathcal{F}$  of all possible pre-activated radiance fields expressed as flattened 4D tensors of fixed size.<sup>2</sup> The chain has a finite number of time steps  $\{0, \dots, T\}$ . The denoising process starts by sampling a state  $f_T$  from a standard, multivariate normal distribution  $p(f_T) := \mathcal{N}(f_T | 0, I)$ , and generates states  $f_{t-1}$  from  $f_t$  by leveraging reversed transition probabilities  $p_\theta(f_{t-1} | f_t)$  that are Gaussian with learned parameters  $\theta$ . Specifically we have

$$p_\theta(f_{t-1} | f_t) := \mathcal{N}(f_{t-1} | \mu_\theta(f_t, t), \Sigma_t). \quad (3)$$

The generation process iterates up to the final state  $f_0$ , which represents the radiance field of a 3D object generated by our method. The mean of the Gaussian in (3) can be directly modeled with a neural network. However, as we will see later, it is more convenient to consider the following reparametrization of it

$$\mu_\theta(f_t, t) := a_t(f_t - b_t \epsilon_\theta(f_t, t)), \quad (4)$$

<sup>2</sup>We consider pre-activated radiance fields, where both density and RGB color channels span a linear space and we assume a proper activation function will be applied at the time of rendering. This is required to have additive noise, while preserving a valid radiance field representation.

where  $\epsilon_\theta(f_t, t)$  is the noise that has been used to corrupt  $f_{t-1}$  predicted by *e.g.* a neural network, whereas  $a_t$  and  $b_t$  are pre-defined coefficients. Also the covariance  $\Sigma_t$  takes a pre-defined value, although it could be data-dependent. Additional details about the value that the pre-defined variables take are given in Sec. 3.3.

**Diffusion process.** While the generation process works by iteratively denoising a completely random radiance field, the diffusion process works the other way around and iteratively corrupts samples from the distribution of 3D objects we want to model. We introduce it because it plays a fundamental role in the training scheme of the generation process. The diffusion process is governed by a discrete-time Markov chain with the same state space and time bounds mentioned in the generation process but with Gaussian transition probabilities that are pre-defined and given by

$$q(f_t | f_{t-1}) := \mathcal{N}(f_t | \sqrt{\alpha_t} f_{t-1}, \beta_t I), \quad (5)$$

where  $\alpha_t := 1 - \beta_t$  and  $0 \leq \beta_t \leq 1$  are predefined coefficients implementing a schedule for the injected noise variance. The process starts by picking  $f_0$  from the distribution  $q(f_0)$  of 3D object radiance fields we want to model, iteratively samples  $f_t$  given  $f_{t-1}$  yielding a scaled and noise-corrupted version of the latter, and stops with  $f_T$  being typically close to completely random depending on the implemented noise-variance schedule. By exploiting properties of the Gaussian distribution, we can conveniently express the distribution of  $f_t$  conditioned on  $f_0$  directly as a Gaussian distribution, yielding

$$q(f_t | f_0) = \mathcal{N}(f_t | \sqrt{\bar{\alpha}_t} f_0, (1 - \bar{\alpha}_t) I), \quad (6)$$

where  $\bar{\alpha}_t := \prod_{i=1}^t \alpha_i$ . This relation will be useful to quickly generate diffused data points at arbitrary time steps.

### 3.3. Training Objective

Our training objective comprises two complementary losses: i) A loss  $L_{RF}$  that penalizes the generation of radiance fields that do not fit the data distribution, and ii) an

RGB loss  $L_{\text{RGB}}$  geared towards improving the quality of renderings from generated radiance fields.

**Radiance field generation loss.** Following [20], we derive the training objective for our model starting from a variational upper-bound on the Negative Log-Likelihood (NLL). This upper-bound requires specifying a surrogate distribution that we refer to as  $q$  because it indeed corresponds to the distribution  $q$  governing the diffusion process, establishing the anticipated fundamental link with the generation process. We provide here some key steps of the derivation of the bound and refer to [20] for more details about the intermediate ones. By Jensen inequality, the NLL of a data point  $f_0 \in \mathcal{F}$  can be upper-bounded by leveraging  $q$  as follows:

$$-\log p_\theta(f_0) \leq \mathbb{E}_q \left[ -\log \frac{p_\theta(f_{0:T})}{q(f_{1:T}|f_0)} \right] := L_{\text{RF}}(f_0|\theta), \quad (7)$$

where  $f_{t_1:t_2}$  stands for  $(f_{t_1}, \dots, f_{t_2})$ . The loss  $L_{\text{RF}}(f_0|\theta)$  bounding the NLL can be further decomposed into the following sum, up to a constant independent from  $\theta$

$$L_{\text{RF}}(f_0|\theta) = \sum_{t=1}^T L_{\text{RF}}^t(f_0|\theta) + \text{const.} \quad (8)$$

Here,  $L_{\text{RF}}^t(f_0|\theta)$  takes a simple and intuitive form if we set  $a_t := \frac{1}{\sqrt{\alpha_t}}$  and  $b_t := \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}$  in (4), and pick  $\Sigma_t := \frac{\beta_t^2}{2\alpha_t(1-\bar{\alpha}_t)} I$ . Indeed, it yields

$$\begin{aligned} L_{\text{RF}}^t(f_0|\theta) &:= \mathbb{E}_q \left[ \|\epsilon - \epsilon_\theta(f_t, t)\|^2 \right] \\ &= \mathbb{E}_\phi \left[ \|\epsilon - \epsilon_\theta(\sqrt{\bar{\alpha}_t}f_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t)\|^2 \right], \end{aligned}$$

where  $\phi(\epsilon) := \mathcal{N}(\epsilon|0, I)$  is the probability distribution of  $\epsilon$ , namely a normal multivariate, and the last equality follows from (6).

**Radiance field rendering loss.** We complement the previous loss with an additional RGB loss  $L_{\text{RGB}}(f_0|\theta)$ , aimed at improving the quality of renderings from generated radiance fields. Indeed, the Euclidean metric on the representation that is implicitly used in the previous loss to assess the quality of generated radiance fields does not necessarily ensure the absence of artifacts once we try to render the radiance field. We define  $L_{\text{RGB}}(f_0|\theta)$  as a sum of time-specific terms  $L_{\text{RGB}}^t(f_0|\theta)$  similar to (8), yielding

$$L_{\text{RGB}}(f_0|\theta) := \sum_{t=1}^T L_{\text{RGB}}^t(f_0|\theta). \quad (9)$$

Given a radiance field  $f \in \mathcal{F}$  and a viewpoint  $v$ , we denote by  $R(v, f)$  the image obtained after rendering  $f$  from viewpoint  $v$  using equation (1). We also denote by  $\ell_v(f, f')$  the Euclidean distance between the rendered images from viewpoint  $v$  using radiance fields  $f$  and  $f'$ , i.e.

$$\ell_v(f, f') := \|R(v, f) - R(v, f')\|^2. \quad (10)$$

The idea is to compare the rendering of a given radiance field  $f_0$  sampled from the data distribution against the same radiance field corrupted with  $t$  diffusion steps and then fully denoised. In theory, this implies sampling first  $f_t$  from  $q(f_t|f_0)$  and then sampling back  $f_0$  from  $p_\theta(f_0|f_t)$ . This is, however, computationally demanding and we resort to a simpler approximation. From the definition of  $L_{\text{RF}}^t$ , the loss pushes towards having  $\epsilon \approx \epsilon_\theta(f_t, t)$  from which we can derive the approximation  $\tilde{f}_0^t(\epsilon, \theta) := f_0 + \frac{\sqrt{1-\bar{\alpha}_t}}{\sqrt{\alpha_t}} (\epsilon - \epsilon_\theta(f_t, t))$ . We can then define our rendering loss as

$$L_{\text{RGB}}^t(f_0|\theta) := \omega_t \mathbb{E}_{\phi, \psi} \left[ \ell_v(f_0, \tilde{f}_0^t(\epsilon, \theta)) \right], \quad (11)$$

where the expectation is taken with respect to a prior distribution  $\psi$  for the viewpoint  $v$  and  $\epsilon \sim \phi(\epsilon)$ . Since the approximation is reasonable only with steps  $t$  close to zero, we introduce a weight  $w_t$  that decays as step values increase (e.g. we use  $\omega_t := \bar{\alpha}_t^2$ ). We provide evidence in the experimental section that despite being an approximation, the proposed loss contributes to significantly improving the results.

**Final loss.** To summarize, the final training loss per data point  $f_0$  is given by the weighted combination of the radiance field generation and rendering losses introduced before, with a small variation that enables stochastic sampling of the step  $t$  from a uniform distribution  $\kappa(t)$ :

$$\begin{aligned} L(\theta) &:= L_{\text{RF}}(f_0|\theta) + \lambda_{\text{RGB}} L_{\text{RGB}}(f_0|\theta) \\ &\propto \mathbb{E}_\kappa \left[ L_{\text{RF}}^t(f_0|\theta) + \lambda_{\text{RGB}} L_{\text{RGB}}^t(f_0|\theta) \right]. \end{aligned}$$

**Implementation details** We implement  $\epsilon_\theta$  as a 3D-UNet, which is based on the 2D-UNet architecture introduced in [14] by replacing 2D convolutions and attention layers with corresponding 3D operators. Furthermore, we deploy a small CNN to refine the rendered images, improving the image quality without impacting view consistency. For training, we uniformly sample timesteps  $t = 1, \dots, T = 1000$  for all experiments with variances of the diffusion process linearly increasing from  $\beta_1 = 0.0015$  to  $\beta_T = 0.05$ , and choose to weight the rendering loss  $L_{\text{RGB}}^t$  with  $w_t = \bar{\alpha}_t^2$ . We refer to the supplementary material for additional details.

## 4. Experiments

In this section, we evaluate the performance of our method on both unconditional and conditional radiance field generation.

**Datasets.** We run experiments on the PhotoShape Chairs [41] and on the Amazon Berkeley Objects (ABO) Tables dataset [9]. For PhotoShape Chairs, we render the provided 15,576 chairs using Blender Cycles [10] from 200 views on an Archimedean spiral. For ABO Tables, we use the provided 91 renderings with 2-3 different environment map settings per object, resulting in 1676 tables. Since both

datasets do not provide radiance field representations of 3D objects, we generate them using a voxel-based approach at a resolution of  $32^3$  from the multi-view renderings.

**Metrics.** We evaluate image quality using the Fréchet Inception Distance [19] (FID) and Inception Score [45] (IS) using [40]. For comparison of the geometrical quality, we follow [1] and compute the Coverage Score (COV) and Minimum Matching Distance (MMD) using Chamfer Distance (CD). While the Coverage Score measures the diversity of the generated samples, MMD assesses the quality of the generated samples. All metrics are evaluated at a resolution of  $128 \times 128$ .

#### 4.1. Unconditional Radiance Field Synthesis

**Comparison against state of the art.** We quantitatively evaluate our approach on the task of unconditional 3D synthesis on PhotoShape [41] in Tab. 1 and ABO Tables [9] in Tab. 2. We compare against leading methods for 3D-aware image synthesis:  $\pi$ -GAN [5] and EG3D [6]. Both our method and GAN-based approaches use the same set of rendered images for training. While we pre-process the rendered images to create a radiance field representation for each of the shape samples, the GAN-based methods are trained directly on the rendered images.

Compared to these approaches, our method yields overall better image quality while achieving significant improvements in geometrical quality and diversity. Fig. 3 and Fig. 4 show a qualitative comparison of our method with  $\pi$ -GAN and EG3D. While EG3D achieves good image quality, it tends to produce inaccurate shapes and view-dependent image artifacts, like adding or removing armrests or changing the supporting structure. As the training objective (see Sec. 3.3) is to invert the diffusion process to denoise towards detailed volumetric representations, we observe that DiffRF reliably generates radiance fields with fine photo-metric and geometrical details.

**Contribution of the rendering loss.** Tab. 1 and Tab. 2 show ablation results where we evaluate the effects of 2D supervision on the radiance synthesis. Removing 2D supervision (“DiffRF w/o 2D”, line 3 in the tables), as expected, has a large effect on FID, which increases by  $\approx 8.8$  for PhotoShape and by  $\approx 8.5$  for ABO Tables. This shows that biasing the noise prediction formulation from DDPM by a volumetric rendering loss leads to higher image quality. We notice a decrease in the Coverage Score that we explain by the fact that the rendering loss guides the denoising model towards learning radiance fields with less artifacts, thus reducing the amount of diverse but spurious shapes.

**Contribution of the refinement CNN.** The removal of CNN-based image refinement (“DiffRF w/o refine”, line 4 in the tables) has a similar, but less pronounced effect: decreased image and shape quality, increased shape diversity.

Method	FID $\downarrow$	IS $\uparrow$	COV $\uparrow$	MMD $\downarrow$
$\pi$ -GAN [5]	68.16	2.428	35.18	8.176
EG3D [6]	27.03	3.641	46.37	5.864
DiffRF w/o 2D	34.42	3.274	<b>53.34</b>	4.305
DiffRF w/o refine	27.39	3.674	52.72	4.272
DiffRF	<b>25.64</b>	<b>3.768</b>	52.41	<b>4.264</b>

Table 1. Quantitative comparison of unconditional generation on the PhotoShape Chairs [41] dataset. Our method achieves a better image and geometric quality than state-of-the-art GAN-based approaches. Both the 2D rendering loss and the CNN refinement are crucial to our method, as indicated by the drop in quality without them. MMD scores are multiplied by  $10^3$ .

Method	FID $\downarrow$	IS $\uparrow$	COV $\uparrow$	MMD $\downarrow$
$\pi$ -GAN [5]	65.40	2.851	44.11	8.393
EG3D [6]	52.71	<b>3.332</b>	52.94	6.417
DiffRF w/o 2D	60.15	2.918	<b>63.73</b>	4.921
DiffRF w/o refine	53.84	3.254	62.75	4.874
DiffRF	<b>51.67</b>	3.324	61.76	<b>4.863</b>

Table 2. Quantitative comparison of unconditional generation on the ABO Tables [9] dataset. Our method achieves a better image and geometric quality than state-of-the-art GAN-based approaches, with both 2D rendering loss and CNN refinement being important. MMD scores are multiplied by  $10^3$ .

Masking:	20%	40%	60%	80%	Avg
mPSNR $\uparrow$	EG3D	22.03	23.36	23.89	24.49
	DiffRF	26.81	26.84	26.78	26.79
FID $\downarrow$	EG3D	41.52	37.68	38.12	36.08
	DiffRF	33.61	32.94	34.61	32.47

Table 3. Quantitative evaluation on the task of radiance field completion at different levels of masking. EG3D struggles to faithfully reconstruct non-masked regions of the original sample, while our method maintains the non-masked regions and synthesises coherent completions for the masked regions.

#### 4.2. Conditional Generation

GANs need to be trained in order to be conditioned on a particular task, while diffusion models can be effectively conditioned at test-time [14, 29, 54]. We leverage this property for the novel task of masked radiance field completion.

**Masked Radiance Field Completion.** Shape completion and image inpainting are well-studied tasks [56, 63, 66, 68] aiming to fill missing regions within a geometrical representation or in an image, respectively. We propose to combine both in the novel task of masked radiance field completion: Given a radiance field and a 3D mask, synthesize a completion of the masked region that harmonizes with the non-



Figure 3. Qualitative comparison between  $\pi$ -GAN [5], EG3D [6], and our method on PhotoShape Chairs [41]. Our approach leads to diverse, geometrically accurate models that allow for high-quality renderings.



Figure 4. Qualitative comparison between  $\pi$ -GAN [5], EG3D [6], and our method on ABO Tables [9]. Our approach generates high quality and diverse samples with accurate geometry.

masked region. Inspired by RePaint [29], we perform conditional completion by gradually guiding the unconditional sampling process in the known region to the input  $f^{in}$

$$f_0^{t-1} = \sqrt{\bar{\alpha}_t}(m \odot \tilde{f}_0^t + (1 - m) \odot f^{in}) \\ f_{t-1} \sim \mathcal{N}(f_0^{t-1}, (1 - \bar{\alpha}_t)I),$$

where  $m$  is a binary mask applied to the input (light blue in Fig. 6) and  $\odot$  denotes element-wise multiplication on the voxel grid.

A quantitative analysis of the masking performance is shown in Tab. 3, where we compare our method against EG3D at different levels of masking. At each level, we randomly mask 200 samples and evaluate the completion performance in terms of FID (by rendering from 10 random views), as well as the photo-metric accuracy of unmasked regions (mPSNR). For EG3D, we use masked GAN inversion, where we perform global latent optimization (GLO) to minimize the photo-metric error on the re-projected, un-



Figure 5. Qualitative completion of masked chairs from PhotoShape [41]. DiffRF shows more diverse proposals compared to EG3D, while also maintaining the original non-masked regions.

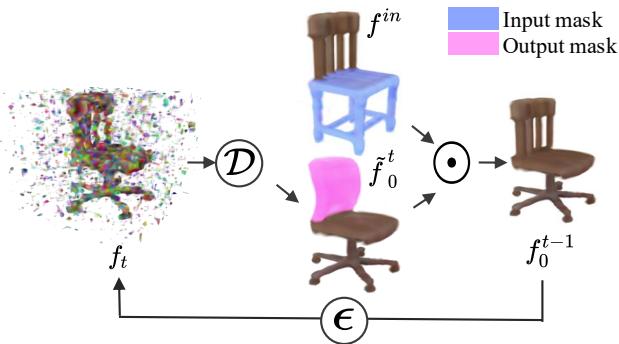


Figure 6. Masked radiance field completion. For a trained denoiser  $\mathcal{D}$  and a given a radiance field with a 3D masked region (shown in blue), completion can be obtained by iteratively fusing the non-masked input region with the estimated denoisings  $\tilde{f}_0^t$  and applying the diffusion process on the resulting  $f_0^{t-1}$ .

masked regions. We notice that, due to the single latent code representation, EG3D struggles to faithfully reconstruct the unmasked region of the input sample, and regularization is needed to not corrupt the overall representation. Fig. 5 further shows a qualitative comparison of the masking performance against EG3D.

**Image-to-Volume Synthesis.** DiffRF can be used to obtain 3D radiance fields from single view images by steering the sampling process using volumetric rendering. For this, we adopt the Classifier Guidance formulation from [14] to guide the denoising process towards minimizing the rendering error against a posed RGB image with corresponding object mask (obtained, for example, with off-the-shelf segmentation networks). Fig. 7 shows qualitative results for this single-image reconstruction task on chairs from ScanNet [11]. With our detailed prior, we can reconstruct reasonable radiance fields even in this ill-posed setting.

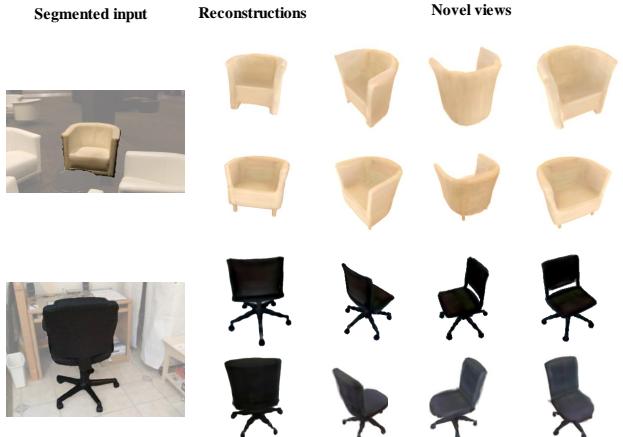


Figure 7. Qualitative synthesis results of single-view reconstruction on ScanNet [11]. Given a single posed image and its foreground mask, our method can recover meaningful proposals for the radiance fields describing the objects (shown in separate rows).

### 4.3. Limitations

While our method shows promising results on the tasks of conditional and unconditional radiance field synthesis, several limitations remain. Compared to GAN-based approaches, sampling times are significantly longer. In this context, it would be interesting to explore leveraging faster sampling methods [26]. Finally, our model is constrained in the maximum grid resolution by training-time memory limitations. These could be addressed by exploring factorized neural fields representations, such as the one in [7].

### 5. Conclusions

We introduce DiffRF – a novel approach for 3D radiance field synthesis based on denoising diffusion probabilistic models. To the best of our knowledge, DiffRF is the first generative diffusion-based method to operate directly on volumetric radiance fields. Our model learns multi-view consistent priors from collections of posed images, enabling free-view image synthesis and accurate shape generation. We evaluated DiffRF on several object classes, comparing its performance against state-of-the-art GAN-based approaches, and demonstrating its effectiveness in both conditional and unconditional 3D generation tasks.

### Acknowledgements

This work was done during Norman’s and Yawar’s internships at Meta Reality Labs Zurich as well as at TUM, funded by a Meta SRA. Matthias Nießner was also supported by the ERC Starting Grant Scan2CAD (804724).

## References

- [1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International conference on machine learning*, pages 40–49. PMLR, 2018. 6
- [2] Miguel Angel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander Toshev, Zhiyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, et al. Gaudi: A neural architect for immersive 3d scene generation. *arXiv preprint arXiv:2207.13751*, 2022. 3
- [3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 3
- [4] Ruojin Cai, Guandao Yang, Hadar Averbuch-Elor, Zekun Hao, Serge J. Belongie, Noah Snavely, and Bharath Hariharan. Learning gradient fields for shape generation. *CoRR*, abs/2008.06520, 2020. 3
- [5] Eric Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *arXiv*, 2020. 2, 3, 6, 7, 12
- [6] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3D generative adversarial networks. In *CVPR*, 2022. 2, 3, 6, 7, 12
- [7] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. *arXiv preprint arXiv:2203.09517*, 2022. 8
- [8] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. Text2shape: Generating shapes from natural language by learning joint embeddings. In *Asian conference on computer vision*, pages 100–116. Springer, 2018. 3
- [9] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21126–21136, 2022. 5, 6, 7, 12, 14, 15
- [10] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 5
- [11] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 8
- [12] Giannis Daras, Mauricio Delbracio, Hossein Talebi, Alexandros G Dimakis, and Peyman Milanfar. Soft diffusion: Score matching for general corruptions. *arXiv preprint arXiv:2209.05442*, 2022. 2
- [13] Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. Unconstrained scene generation with locally conditioned radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (ICCV)*, 2021. 2
- [14] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 3, 5, 6, 8, 12, 13
- [15] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *International Conference on 3D Vision (3DV)*, 2022. 2
- [16] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daqing Li, Or Litany, Zan Gajcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *arXiv preprint arXiv:2209.11163*, 2022. 3
- [17] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. Stylenerf: A style-based 3d-aware generator for high-resolution image synthesis. *arXiv preprint arXiv:2110.08985*, 2021. 3
- [18] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Escaping plato’s cave: 3d shape from adversarial rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9984–9993, 2019. 3
- [19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6
- [20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020. 2, 3, 4, 5
- [21] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *arXiv preprint arXiv:2204.03458*, 2022. 3
- [22] Emiel Hoogeboom and Tim Salimans. Blurring diffusion models. *arXiv preprint arXiv:2209.05557*, 2022. 2
- [23] Chin-Wei Huang, Jae Hyun Lim, and Aaron C Courville. A variational perspective on diffusion-based generative models and score matching. *Advances in Neural Information Processing Systems*, 34:22863–22876, 2021. 2
- [24] Animesh Karnewar, Tobias Ritschel, Oliver Wang, and Niloy Mitra. Relu fields: The little non-linearity that could. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH ’22, New York, NY, USA, 2022. Association for Computing Machinery. 4
- [25] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *Proc. NeurIPS*, 2022. 2
- [26] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. *arXiv preprint arXiv:2106.00132*, 2021. 8
- [27] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 2

- [28] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 12
- [29] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11461–11471, June 2022. 6, 7
- [30] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2837–2845, 2021. 2, 3
- [31] Shitong Luo, Chence Shi, Minkai Xu, and Jian Tang. Predicting molecular conformation via dynamic graph score matching. *Advances in Neural Information Processing Systems*, 34:19784–19795, 2021. 3
- [32] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [33] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiangujun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 3
- [34] 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*, pages 405–421. Springer, 2020. 1, 3
- [35] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 3
- [36] Norman Müller, Andrea Simonelli, Lorenzo Porzi, Samuel Rota Bulò, Matthias Nießner, and Peter Kontschieder. Autorf: Learning 3d object radiance fields from single view observations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [37] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. GLIDE: towards photorealistic image generation and editing with text-guided diffusion models. *CoRR*, abs/2112.10741, 2021. 3
- [38] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021. 2
- [39] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 3
- [40] Anton Obukhov, Maximilian Seitzer, Po-Wei Wu, Semen Zhydenko, Jonathan Kyl, and Elvis Yu-Jing Lin. High-fidelity performance metrics for generative models in pytorch, 2020. Version: 0.3.0, DOI: 10.5281/zenodo.4957738. 6
- [41] Keunhong Park, Konstantinos Rematas, Ali Farhadi, and Steven M. Seitz. Photoshape: Photorealistic materials for large-scale shape collections. *ACM Trans. Graph.*, 37(6), Nov. 2018. 2, 5, 6, 7, 8, 12, 13, 15
- [42] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022. 3
- [43] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3
- [44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 3
- [45] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved techniques for training gans. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. 6
- [46] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2, 3
- [47] Yawar Siddiqui, Justus Thies, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Texturify: Generating textures on 3d shape surfaces. *arXiv preprint arXiv:2204.02411*, 2022. 3
- [48] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2
- [49] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2
- [50] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2
- [51] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34:1415–1428, 2021. 2
- [52] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *CoRR*, abs/1907.05600, 2019. 2
- [53] Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in neural information processing systems*, 33:12438–12448, 2020. 2
- [54] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2, 6

- [55] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, 2022. 4
- [56] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. *arXiv preprint arXiv:2109.07161*, 2021. 6
- [57] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-NeRF: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [58] Brian L Trippe, Jason Yim, Doug Tischer, Tamara Broderick, David Baker, Regina Barzilay, and Tommi Jaakkola. Diffusion probabilistic modeling of protein backbones in 3d for the motif-scaffolding problem. *arXiv preprint arXiv:2206.04119*, 2022. 3
- [59] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [60] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2
- [61] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. *Advances in neural information processing systems*, 29, 2016. 3
- [62] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. Geodiff: A geometric diffusion model for molecular conformation generation. *arXiv preprint arXiv:2203.02923*, 2022. 3
- [63] Guandao Yang, Xun Huang, Zekun Hao, Ming-Yu Liu, Serge Belongie, and Bharath Hariharan. Pointflow: 3d point cloud generation with continuous normalizing flows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4541–4550, 2019. 6, 12
- [64] Ruihan Yang, Prakhar Srivastava, and Stephan Mandt. Diffusion probabilistic modeling for video generation. *arXiv preprint arXiv:2203.09481*, 2022. 3
- [65] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2
- [66] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. *arXiv preprint arXiv:1801.07892*, 2018. 6
- [67] Xiaohui Zeng, Arash Vahdat, Francis Williams, Zan Gojcic, Or Litany, Sanja Fidler, and Karsten Kreis. Lion: Latent point diffusion models for 3d shape generation. *arXiv preprint arXiv:2210.06978*, 2022. 3
- [68] Linqi Zhou, Yilun Du, and Jiajun Wu. 3d shape generation and completion through point-voxel diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5826–5835, 2021. 2, 3, 6
- [69] Peng Zhou, Lingxi Xie, Bingbing Ni, and Qi Tian. Cips-3d: A 3d-aware generator of gans based on conditionally-independent pixel synthesis. *arXiv preprint arXiv:2110.09788*, 2021. 3

## Appendix

In this supplementary document, we discuss additional details about our method, the data used for training and evaluation, and show further qualitative results. We also refer to the attached supplementary video for a comprehensive overview with further qualitative results.

### A. Additional qualitative results

We provide additional qualitative results on PhotoShape Chairs [41] in Fig. 8 as well as on ABO Tables [9] in Fig. 9. Moreover, we show additional visual comparisons between  $\pi$ -GAN [5], EG3D [6], and our method on both datasets in Fig. 10 and Fig. 11.

### B. Implementation detail

**Architecture** We base our architecture on 2D U-Net structure from [14]: For this, we replace the 2D convolutions in the ResNet and attention blocks with corresponding 3D convolutions, preserving kernel sizes and strides. Furthermore, we use 3D average pooling layers instead of 2D in the downsampling steps. Our U-Net consists of 4 scaling blocks with two ResNet blocks per scale, where we linearly increase the initial feature channel dimension of 64 to 256. We use skip attention blocks at the scaling factors 2, 4, and 8 with 32 channels per head. For the refinement CNN, we use a small CNN consisting of two ResNet blocks with a hidden dimension of 64.

**Training details** We train all models with a batch size of 8 and use the Adam optimizer with an initial learning of  $10^{-4}$ . We apply a linear beta scheduling from 0.0015 to 0.05 at 1000 timesteps. From 4 random training views at a resolution of  $128 \times 128$ , we sample 8192 random pixels for the rendering supervision (with 92 z-steps for volumetric rendering) and weight the rendering loss with  $\omega_t = \bar{\alpha}_t^2$ . We train for 3.0m iterations with a decaying LR scheduling for  $10^{-4}$  to  $10^{-6}$  at a voxel grid resolution of 32 on 2 GPUs on every data set.

**Sampling time** We perform DDPM sampling for 1000 iterations leading to a run time of 48.6s per sample on an NVIDIA RTX 2080 TI. Once synthesized, our explicit representation enables rendering at  $128 \times 128$  resolution with over 380 FPS.

### C. Data

**Radiance Field Generation** For PhotoShape Chairs, we render the provided 15,576 chairs using Blender Cycles from 200 views on an Archimedean spiral at a fixed radius of 2.5 units with pitch starting from  $-20^\circ$  to  $60^\circ$ . For ABO

Tables, we use the provided 91 renderings with 2-3 different environment map settings per object, resulting in 1676 tables. On both data sets, we hold out 10% of the samples for testing and fit explicit voxel grids at a resolution of  $32^3$  using volumetric rendering with spherical harmonics of degree 2 for an initial fit. We then fine-tune our representations for spherical harmonics of degree 0, which we found to lead to sharper geometry compared to directly optimizing density and color features. We furthermore bound the feature space to  $[-1, 1]$  which we found to stabilize the sampling process noticeably affecting the rendering quality.

**Evaluation** For image quality evaluation, we calculate FID and IS by sampling 10k views by rendering 1000 samples from 10 random views at a resolution of  $128 \times 128$ . We follow [63] and evaluate the geometric quality by computing the Coverage Score (COV) and Minimum Matching Distance (MMD) using Chamfer Distance (CD)

$$\begin{aligned} \text{CD}(X, Y) &= \sum_{x \in X} \min_{y \in Y} \|x - y\|_2^2 + \sum_{y \in Y} \min_{x \in X} \|x - y\|_2^2, \\ \text{COV}(S_g, S_r) &= \frac{|\{\arg \min_{Y \in S_r} \text{CD}(X, Y) | X \in S_g\}|}{|S_r|}, \\ \text{MMD}(S_g, S_r) &= \frac{1}{|S_r|} \sum_{Y \in S_r} \min_{X \in S_g} \text{CD}(X, Y), \end{aligned}$$

on a reference set  $S_r$  (the test samples) and a generated set  $S_g$  twice as large as the reference set. We extract meshes using marching cubes [28] and sample 2048 points on the faces. To account for potentially different scaling of the samples produced by the 3D-aware GAN models, we normalize all point clouds by centering in the origin and anisotropic scaling of the extent to  $[-1, 1]$ .

For evaluation of the masked radiance field completion, we additionally compute a masked peak signal-to-noise ratio (mPSNR): Given a binary mask  $m$  of the input radiance field  $f^{in}$ , we compute for each corresponding input image the non-masked area by depth-based projection into the image plane using depth estimated from the input radiance field. We then compute the mPSNR by averaging the PSNR on the non-masked pixels for all evaluation views (we choose 10 views randomly).

### D. Conditional sampling

**Masked completion** Since the generator of EG3D [6] is trained via 2D discriminator guidance, we perform 3D masked completion via GAN inversion. For this, we start from a random initial latent code and repeat the following steps for 200 iterations on each masked sample: We render the current synthesized sample from 8 views and project the 3D input mask onto the synthesized views using the predicted depths. On the remaining non-masked regions, we



Figure 8. Additional qualitative sampling results on PhotoShape Chairs [41].

compute the photometric error with the input images. We use the Adam optimizer with a learning rate of  $10^{-2}$  with a small  $L_2$  regularization term on the code (weighted with  $5 \times 10^{-2}$ ) in order to update the latent code.

**Image-to-Volume Synthesis** Given a posed and segmented image, we condition our trained radiance field diffusion model by steering the sampling processing similar to the Classifier Guidance formulation from [14]: During sampling time, for each time step  $t$ , we gradually update the predicted denoised field  $\tilde{f}_0^t$  towards minimizing the photometric error obtained from comparing the rendering  $\tilde{I}_t$  from a given pose with the foreground-masked target image  $I$ . For this, we compute the gradient  $\nabla_{\tilde{f}_0^t}(\tilde{I}_t, I)$  on the current denoising estimate by volumetric rendering and steer the

sampling process by  $\tilde{f}_0^t \leftarrow \tilde{f}_0^t - \lambda \nabla_{\tilde{f}_0^t}(\tilde{I}_t, I)$  with a small guidance weight  $\lambda$ .



Figure 9. Additional qualitative sampling results on ABO Tables [9].



Figure 10. Additional qualitative comparisons on PhotoShape Chairs [41].



Figure 11. Additional qualitative comparisons on ABO Tables [9].