RGBD2: Generative Scene Synthesis via Incremental View Inpainting using RGBD Diffusion Models

Jiabao Lei[1], Jiapeng Tang[2], Kui Jia[1,3,†]

[1]South China University of Technology
[2]Technical University of Munich, [3]Peng Cheng Laboratory

Abstract

We address the challenge of recovering an underlying scene geometry and colors from a sparse set of RGBD view observations. In this work, we present a new solution termed RGBD² that sequentially generates novel RGBD views along a camera trajectory, and the scene geometry is simply the fusion result of these views. More specifically, we maintain an intermediate surface mesh used for rendering new RGBD views, which subsequently becomes complete by an inpainting network; each rendered RGBD view is later back-projected as a partial surface and is supplemented into the intermediate mesh. The use of intermediate mesh and camera projection helps solve the tough problem of multi-view inconsistency. We practically implement the RGBD inpainting network as a versatile RGBD diffusion model, which is previously used for 2D generative modeling; we make a modification to its reverse diffusion process to enable our use. We evaluate our approach on the task of 3D scene synthesis from sparse RGBD inputs; extensive experiments on the ScanNet dataset demonstrate the superiority of our approach over existing ones. Project page: https://jblei.site/proj/rgbd-diffusion.

1. Introduction

Scene synthesis is an essential requirement for many practical applications. The resulting scene representation can be readily utilized in diverse fields, such as virtual reality, augmented reality, computer graphics, and game development. Nevertheless, conventional approaches to scene synthesis usually involve reconstructing scenes (e.g., indoor scenes with varying sizes) by fitting given observations, such as multi-view images or point clouds. The increasing prevalence of RGB/RGBD scanning devices has established multi-view data as a favored input modality, driving and promoting technical advancements in the realm of scene reconstruction from multi-view images.

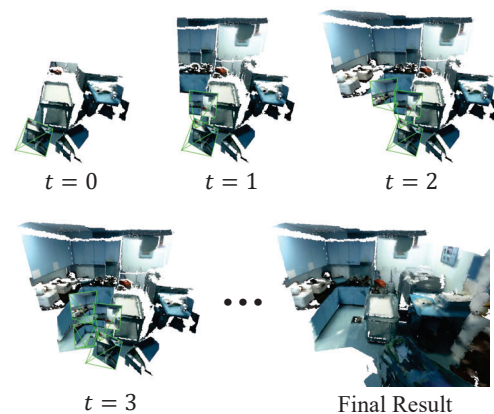Neural Radiance Fields (NeRFs) [42] have demonstrated

Figure 1. **Illustration of Our Generative Scene Synthesis.** We incrementally reconstruct the scene geometry by inpainting RGBD views as the camera moves in the scene.

potential in this regard, yet they are not exempt from limitations. NeRFs are designed to reconstruct complete scenes by fitting multi-view images, and they cannot generate or infer missing parts when the input is inevitably incomplete or missing. While recently some studies [3,5,8,56,63] have attempted to equip NeRFs with generative and extrapolation capabilities, this functionality relies on a comparatively short representation with limited elements (e.g. typically, the length of a global latent code is much shorter than that of an image: $(F = 512) \ll (H \times W = 128 \times 128 = 16,384)$) that significantly constrains their capacity to accurately capture fine-grained details in the observed data. Consequently, the effectiveness of these methods has only been established for certain categories of canonical objects, such as faces or cars [8,63], or relatively small toy scenes [5].

We introduce a novel task of generative scene synthesis from sparse RGBD views, which involves learning across multiple scenes to later enable scene synthesis from a sparse set of multi-view RGBD images. This task presents a challenging setting wherein a desired solution should simultaneously (1) preserve observed regions, hallucinate missing parts of the scene, (2) eliminate additional computational costs during inference for each individual test scene, (3) en-

sure exact 3D consistency, and (4) maintain scalability to scenes with unfixed scales.

We will elaborate on them in detail as follows. Firstly, to maximize the preservation of intricate details while simultaneously hallucinating potentially absent parts that may become more pronounced when views are exceedingly sparse, we perform straightforward reconstruction whose details come from images that can describe fine structures using a maximum of $H \times W$ elements (i.e. an image size) in a view completion manner. This is particularly compatible with diffusion models that operate at full image resolution with an inpainting mechanism. We also found that RGBD diffusion models greatly simplify the training complexity of a completion model, thanks to their versatile generative ability to inpaint missing RGBD pixels while preserving the integrity of known regions through a convenient training process solely operated on complete RGBD data. Secondly, our method employs back-projection that requires no optimization, thus eliminating the necessity for test-time training for each individual scene, ultimately leading to a significant enhancement in test-time efficiency. Thirdly, to ensure consistency among multi-view images, an intermediate mesh representation is utilized as a means of bridging the 2D domain (i.e. multi-view RGBD images) with the 3D domain (i.e. the 3D intermediate mesh) through the aid of camera projection. Fourthly, to enable our method to handle scenes of indeterminate sizes, we utilize images with freely designated poses as the input representation. Such manner naturally ensures SE(3) equivariance, and thus offers scalability due to the ease with which the range of the generated content can be controlled by simply specifying their camera extrinsic matrices.

Our proposal involves generating multi-view consistent RGBD views along a predetermined camera trajectory, using an intermediate mesh to render novel RGBD images that are subsequently inpainted using a diffusion model, and transforming each RGBD view into a 3D partial mesh via back-projection, and finally merging it with the intermediate scene mesh to produce the final output. Specifically, our proposed approach initiates by ingesting multiple posed RGBD images as input and utilizing back-projection to construct an intermediate scene mesh. This mesh encompasses color attributes that facilitate the rendering of RGBD images from the representation under arbitrarily specified camera viewpoints. Once a camera pose is selected from the test-time rendering trajectory, the intermediate mesh is rendered to generate a new RGBD image for this pose. Notably, the test-time view typically exhibits only slight overlap with the known cameras, leading to naturally partially rendered RGBD images. To fill the gaps in the incomplete view, we employ an inpainting network implemented as an RGBD diffusion model with minor modifications to its reverse sampling process. The resulting inpainted out-

put is then back-projected into 3D space, forming a partial mesh that complements the entire intermediate scene mesh. We iterate these steps until all test-time camera viewpoints are covered, and the intermediate scene mesh gradually becomes complete during this process. The final output of our pipeline is the mesh outcome acquired from the last step.

Extensive experiments on ScanNet [12] dataset demonstrate the superiority of our approach over existing solutions on the task of scene synthesis from sparse RGBD inputs.

2. Related Works

In this section, we provide a brief review of the literature related to diffusion models, 3D representations and generative manners, scene synthesis, and view synthesis.

Diffusion Models. In recent years, the field of 2D computer vision has experienced a surge of interest in diffusion-based generative models [20,65,67]. These models have prompted the development of image generative modeling approaches, such as GLIDE [45], unCLIP [53], Imagen [61], and Latent Diffusion Models [57], as well as the invention of sampling schedulers [20, 26, 36, 66] and guiding methods [15, 22]. Furthermore, these models have been applied to a broad range of image processing tasks, including image inpainting [37], image translation [4, 57, 60, 82], video generation [23, 25, 76], super-resolution [21, 32, 62], and image editing [2, 27, 40]. More recently, some researchers have adapted these techniques from 2D to the 3D domain, as demonstrated by methods such as [39, 51, 88, 92]. Our approach harnesses such versatility by adopting an iterative denoising strategy like [37], and utilizing a masked inpainting technique that operates on the projected RGBD views to synthesize image content.

3D Representations and Generative Manners. A variety of representations, including voxels [81, 85], point clouds [1, 39, 68, 87, 92], meshes [17, 44, 70, 71, 79], implicit surfaces [10, 30, 31, 41, 48, 72–74, 88, 91], multi-view images [11, 19, 34, 35, 86], and neural radiance fields [3, 8, 9, 13, 42, 43, 49, 56, 63, 80], have been proposed, each with its own unique advantages over the others. This has also motivated researchers to combine them with distinct generative approaches, such as VAEs [28], GANs [16], normalizing flows [54], auto-regressive models [24, 75], and the latest diffusion models [20], resulting in an extensive range of applications [1, 8, 39, 41, 44, 48, 63, 85, 87, 92]. However, most existing methods have limitations in their representation capability, such as cubically scaled-up memory consumption, or a fixed number of points, which makes them difficult to apply to scenes of uncertain scales, and poor equivariance, which only allows them to handle canonically-posed objects. In this paper, we address these issues by focusing on the generation of multi-view RGBD images that can capture intricate structures using $H \times W$ pixels created by a diffusion model. This approach reduces memory complexity from $\mathcal{O}(HWD)$ to $\mathcal{O}(HW)$ and increases expressive

ability from $\mathcal{O}(F)$ to $\mathcal{O}(HW)$.

Scene Synthesis. In this area, there are two primary research directions. The first pertains to learning configurations, including graphs [33, 38, 77, 90], top-down views [55, 78], and scene composition [50, 83]. The second line of research involves direct learning from the appearance of a scene [13, 42, 43, 49, 80], thereby obviating the necessity for specialized synthesized datasets required by the aforementioned methods. Our approach relies solely on RGBD scans, which are readily available from scanning devices, reducing the need for manual annotation. It accurately reconstructs a clean geometry based on the sparse-view input and can effectively hallucinate missing parts, especially when the input views are highly sparse.

View Synthesis. The arrival of NeRF [42] has significantly advanced the field of view synthesis. While a considerable amount of research has been dedicated to the view synthesis of object-level instances [7, 18, 47, 64], only a few studies have focused on simple scenes [5, 14, 58]. Additionally, some studies have explored techniques for improving performance using sparse inputs [46, 56]. In this paper, we aim to recover scene-level geometry from sparse RGBD images without relying on NeRFs.

3. Preliminary

To make our paper self-contained, we provide some preliminary knowledge about DDPM [20] and DDIM [66].

Definition. Given an image $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, the *forward* diffusion process is a Markov chain that sequentially produces noisy images $\mathbf{x}_1, ..., \mathbf{x}_T$ scheduled by a set of constants $\{\beta_t \in (0, 1)\}_{t=1}^T$ where T is the total number of iterations and $\beta_1 < \beta_2 < ... < \beta_T$:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1 - \beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I})$$

A nice property of using i.i.d. Gaussian distribution is that we can sample the noisy image $\mathbf{x}_t$ easily based on $\mathbf{x}_0$ in a closed form without computing all the intermediates:

$$q(\mathbf{x}_t \mid \mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1 - \bar{\alpha}_t)\mathbf{I}) \quad (1)$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^t \alpha_s$. The *reverse* diffusion process is another Markov chain starting from $\mathbf{x}_T \sim \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$ parameterized by a trainable θ:

$$p_\theta(\mathbf{x}_{t-1} \mid \mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t)) \quad (2)$$

where $\boldsymbol{\Sigma}_\theta(\mathbf{x}_t, t) = \sigma_t^2 \mathbf{I} = \eta \frac{1 - \bar{\alpha}_{t-1}}{1 - \bar{\alpha}_t}\beta_t \mathbf{I}$ [66] is independent of θ and $\mathbf{x}_t$, and $\mu_\theta(\mathbf{x}_t, t)$ can be derived from a noise estimator $\epsilon_\theta(\mathbf{x}_t, t)$ via:

$$\mu_\theta(\mathbf{x}_t, \mathbf{x}_o) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1 - \alpha_t}{\sqrt{1 - \bar{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\right)$$

We practically implement the noise estimator $\epsilon_\theta(\mathbf{x}_t, t)$ as a UNet [57, 59] parameterized by θ. However, in our case, it is also conditioned on $\hat{\mathbf{x}}_0$ which will be detailed in Sec. 4.2.

Training. To train the noise estimator $\epsilon_\theta(\mathbf{x}_t, t)$, we adopt the simplified training objective [20]:

$$\min_\theta \mathbb{E}_{t, \epsilon_t, \mathbf{x}_0}\left[\|\epsilon_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon_t, t) - \epsilon_t\|^2\right]$$

where the time step $t \sim \mathcal{U}\{1, T\}$ is uniformly sampled, $\epsilon_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a standard Gaussian noise, and the image $\mathbf{x}_0 \sim q(\mathbf{x}_0)$ is randomly drawn from the data distribution.

Inference. We employ a strided DDIM [66] scheduler to progressively recover the clean image from $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ to $\mathbf{x}_0$ with a subset of S (usually $S \ll T$) steps $\{\tau_i\}_{i=1}^S$:

$$\mathbf{x}_{\tau_{i-1}} \leftarrow \sqrt{\bar{\alpha}_{\tau_{i-1}}}\left(\frac{\mathbf{x}_{\tau_i} - \sqrt{1 - \bar{\alpha}_{\tau_i}}\epsilon_\theta(\mathbf{x}_{\tau_i}, \tau_i)}{\sqrt{\bar{\alpha}_{\tau_i}}}\right) +$$
$$\sqrt{1 - \bar{\alpha}_{\tau_{i-1}} - \sigma_{\tau_i}^2} \cdot \epsilon_\theta(\mathbf{x}_{\tau_i}, \tau_i) + \sigma_{\tau_i}^2 \epsilon_{\tau_i}$$

where $\epsilon_{\tau_i} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is a standard Gaussian noise. It is interesting that setting $\eta = 0$ implies $\forall i$, $\sigma_{\tau_i} \equiv 0$, making the inference process deterministic. Nevertheless, diversity can still be achieved by choosing a different $\mathbf{x}_T$ from $\mathcal{N}(\mathbf{0}, \mathbf{I})$.

4. Approach

In this work, we introduce a new scene synthesis approach that relies on the sequential generation of RGBD frames using a trained RGBD diffusion model conditioned on views rendered under a test-time camera trajectory from a sparse set of input RGBD images. Our approach leverages the generative ability of diffusion models to synthesize missing visual appearance (color) and geometry details (depth) while rigorously preserving visible regions by image inpainting. Such an incremental view inpainting process is interleaved with the back-projection, mesh fusion, and mesh rendering procedures to achieve global 3D consistency among distinct temporal frames.

This section is organized as follow. In Sec. 4.1, we will start by introducing the overall framework of repeatedly converting the RGBD image into a partial 3D mesh and performing RGBD inpainting for rendered views in a progressive manner. Later in Sec. 4.2, we will elaborate more on the details of the conditional RGBD diffusion model for rendered view inpainting based on a trained DDPM [20].

4.1. Incremental View Inpainting Fashion

As shown in Figure 2, given a sparse set of N RGBD views $\{\mathbf{I}_i\}_{i=1}^N$ with their associated camera intrinsic $\mathbf{K}$ and extrinsic $\{\mathbf{R}_i\}_{i=1}^N$ matrices, and additionally a camera trajectory composed of M viewpoints $\{\mathbf{T}_j\}_{j=1}^M$ with the same intrinsic matrix, the essence of our method is to progressively synthesize RGBD images $\mathbf{O}_j$ at each specified novel view $\mathbf{T}_j$. The generated novel-view frames should be consistent with $\{\mathbf{I}_i\}_{i=1}^N$ in both geometry and appearance. The final output should be a consistent 3D
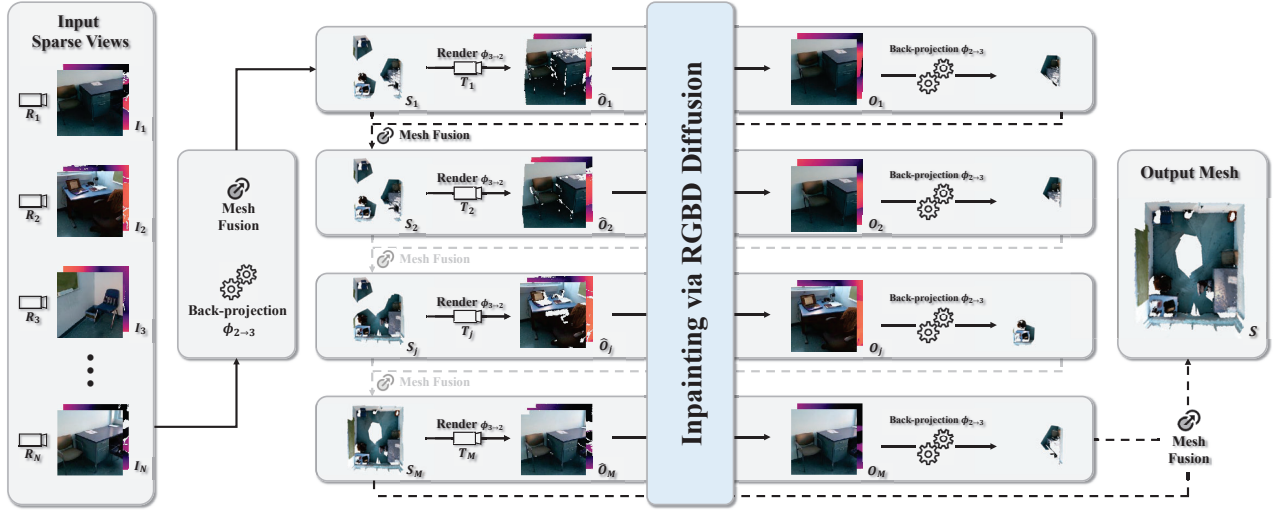
Figure 2. **3D Scene Synthesis via Incremental View Inpainting.** Given a sparse set of RGBD images $\{\mathbf{I}_i\}_{i=1}^{N}$ with camera intrinsic $\mathbf{K}$ and extrinsic $\{\mathbf{R}_i\}_{i=1}^{N}$ matrices, our goal is to generate a coherent 3D scene mesh $\mathcal{S}$ via predicting RGBD frames $\{\mathbf{O}_j\}_{j=1}^{M}$ along a novel-view trajectory $\{\mathbf{T}_j\}_{j=1}^{M}$. To achieve this goal, we firstly fuse the inputs of $\{\mathbf{I}_i\}_{i=1}^{N}$ into an initial mesh $\mathcal{S}_1$, and then render it to obtain an incomplete image $\hat{\mathbf{O}}_1$ that is later inpainted to obtain $\mathbf{O}_1$ using a RGBD diffusion model. After that, $\mathbf{O}_1$ is back-projected and integrated with $\mathcal{S}_1$ to produce a more complete scene, $\mathcal{S}_2$. By iteratively repeating this process, we can progressively obtain $\mathcal{S}_2, ..., \mathcal{S}_M$. Finaly, the fused result $\mathcal{S}_{M+1}$ is the eventual desired output $\mathcal{S}$. Gray dashed lines "- - -" denote omitted and unvisualized steps.

colored mesh $\mathcal{S}$ converted from the back-projection result $\left(\cup_{j=1}^{M}\phi_{2\to3}(\mathbf{O}_j)\right) \cup \left(\cup_{i=1}^{N}\phi_{2\to3}(\mathbf{I}_i)\right)$. Please also refer to Algorithm 1 for a detailed description of the procedure.

Algorithm 1 Incremental RGBD View Inpainting

Require: RGBD images $\{\mathbf{I}_i\}_{i=1}^{N}$ with camera extrinsics $\{\mathbf{R}_i\}_{i=1}^{N}$, a novel trajectory of M extrinsics $\{\mathbf{T}_j\}_{j=1}^{M}$.
　$\mathcal{S}_1 \leftarrow \cup_{i=1}^{N}\phi_{2\to3}(\mathbf{I}_i)$
　for $j = 1$ to M **do**
　　$\hat{\mathbf{O}}_j \leftarrow \phi_{3\to2}(\mathbf{T}_j)$　　　▷ novel-view rendering
　　$\mathbf{O}_j \leftarrow f(\hat{\mathbf{O}}_j)$　　　　　▷ RGBD inpainting
　　$\mathcal{S}_{j+1} \leftarrow \mathcal{S}_j \cup \phi_{2\to3}(\mathbf{O}_j)$　　▷ mesh fusion
　end for
　return $\mathcal{S}_{M+1}$

Rendering and Back-projection.[1] The operator $\phi_{3\to2}$ is implemented as mesh rasterization, which allows for the rendering of a partial RGBD image $\hat{\mathbf{O}}_j$ from a mesh $\mathcal{S}_j$. This approach offers the advantage of producing a clean visibility mask $\mathbf{m}_j$, which is not possible with NeRFs [42]. The back-projection operator $\phi_{2\to3}$ is responsible for the conversion from a depth map into a point cloud, where the connectivity between points is inherited from the connectivity of the 2D pixel grid. Furthermore, mesh faces that are either in close proximity to the viewpoint or exhibit slender characteristics are filtered out to ensure accuracy.

Challenges and Solutions. To circumvent the limitations

posed by potential 3D inconsistency in both geometry and appearance, as well as the challenge of solely handling specific canonically-posed scenes, we propose several strategies to address these issues. Firstly, our method for synthesizing novel views combines rendering (mesh rasterization) and inpainting techniques, and interleaves the view synthesis process with online RGBD fusion via back-projection and mesh combination. The use of perspective camera projection ensures strict adherence to 3D constraints, resulting in visually consistent and accurate synthesized views. In concrete terms, we begin by rendering the mesh $\mathcal{S}_j$ under view $\mathbf{T}_j$ using a rendering operation $\phi_{3\to2}(\mathbf{T}_j)$. This process yields an incomplete RGBD image $\hat{\mathbf{O}}_j$ with missing regions, which is subsequently inpainted using a diffusion model described in detail in Sec. 4.2, resulting in a complete image $\mathbf{O}_j$. Once image $\mathbf{O}_j$ has been generated for view $\mathbf{T}_j$, it can be fused into a 3D mesh via $\mathcal{S}_{j+1} = \phi_{2\to3}(\mathbf{O}_j) \cup \mathcal{S}_j$ using a back-projection operator $\phi_{2\to3}$. Secondly, our solution reduces the learning difficulty and can handle non-canonical scenes by decomposing the 3D scene as a Markov chain of temporal RGBD images rendered from arbitrarily specified novel viewpoints. The applicability of handling scenes with arbitrary scaling, movement, and posing is attributed to the utilization of two SE(3) equivariant operators, namely $\phi_{3\to2}$ and $\phi_{2\to3}$, as well as the independence of absolute coordinates. Moreover, the presence of redundant information in adjacent frames, combined with our suggested decomposition rule, facilitates the minimization of learning complexity in an auto-regressive manner.

---

[1] To simplify notations, we disregard certain arguments in operators $\phi_{3\to2}$ and $\phi_{2\to3}$, such as the camera intrinsic matrix $\mathbf{K}$.

In accordance with the Markov chain decomposition principle, the distribution of the scene can be expressed as the joint distribution of view frames:

$$p(\mathcal{S}) = \prod_j p(\mathbf{O}_j \mid \{\mathbf{O}_s\}_{s=1}^{j-1}, \{\mathbf{I}_s\}_{s=1}^N)$$

where the prediction of $\mathbf{O}_j$ is based on the fusion result of all the previously known frames $\{\mathbf{O}_s\}_{s=1}^{j-1} \cup \{\mathbf{I}_s\}_{s=1}^N$.

4.2. RGBD Diffusion for Rendered View Inpainting

In this section, we describe the implementation details of the way to employ diffusion models to inpaint the missing regions of the RGBD image $\hat{\mathbf{O}}_j$ with a binary mask $\mathbf{m}_j$ obtained by rendering visibility, where the value of 1 is assigned to the corresponding ray that intersects with the geometry surface, while 0 is assigned to all other cases, both of which are rendered by projecting $\mathcal{S}_j$ under a novel viewpoint $\mathbf{T}_j$. It is noteworthy that only those pixels located where $\mathbf{m}_j = 0$ in $\hat{\mathbf{O}}_j$ are considered invalid, and therefore, are entirely filled with zeros (i.e., $\hat{\mathbf{O}}_j \odot \mathbf{m}_j \equiv \hat{\mathbf{O}}_j$).
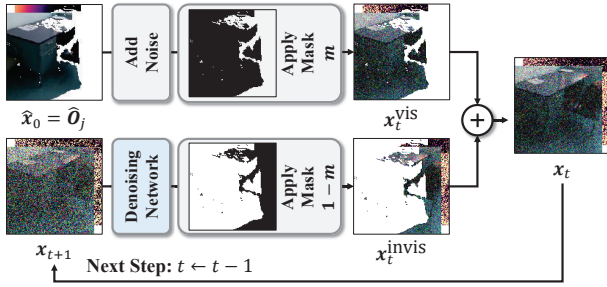

Figure 3. **RGBD Diffusion for Rendered View Inpainting.** In each iteration, we utilize a forward diffusion step to sample the visible region $\mathbf{x}_t^{\text{vis}}$ (*top*) from the incomplete input $\hat{\mathbf{x}}_0$, while employing a denoising diffusion step to obtain the complementary content $\mathbf{x}_t^{\text{invis}}$ (*bottom*) from the output $\mathbf{x}_{t+1}$ generated in the previous iteration $t+1$.

A Single Reverse Diffusion Step for $\mathbf{T}_j$ at Time t.[2] We initialize the clean diffusion sample $\hat{\mathbf{x}}_0$ at time $t = 0$ as $\hat{\mathbf{O}}_j$, and those noisy samples at time $0 < t \leq T$ are denoted as $\mathbf{x}_t$. To ensure 3D consistency, we seek to predict the invisible region $\mathbf{x}_0^{\text{invis}}$ while preserving the visible portion $\mathbf{x}_0^{\text{vis}}$ unaltered. We follow [37] and modify the standard denoising process to meet our needs. At each reverse diffusion step t, we use the forward diffusion process defined by the Eqn. (1) to sample a visible region $\mathbf{x}_t^{\text{vis}}$ masked by $\mathbf{m}$, and the reverse diffusion step defined by Eqn. (2) to sample a full part from $\mathbf{x}_{t+1}$ which is later masked by $1 - \mathbf{m}$ to form an invisible region $\mathbf{x}_t^{\text{invis}}$.

$$\mathbf{x}_t^{\text{vis}} = \sqrt{\bar{\alpha}_t}\hat{\mathbf{x}}_0 + (1 - \bar{\alpha}_t)\epsilon_t^{\text{vis}} \odot \mathbf{m}$$

---

[2] For notational simplicity, we ignore the novel camera subscript j in the following part of this Sec. 4.2 since all notations we consider here are derived from the same camera $\mathbf{T}_j$.

$$\mathbf{x}_t^{\text{invis}} = \left(\mu_\theta\left(\mathbf{x}_t, \hat{\mathbf{x}}_t, t\right) + \sigma_t \epsilon_t^{\text{invis}}\right) \odot (1 - \mathbf{m})$$

where $\epsilon_t^{\text{vis}}, \epsilon_t^{\text{invis}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. The noisy image $\mathbf{x}_t$ can be simply calculated as the sum of $\mathbf{x}_t^{\text{vis}}$ and $\mathbf{x}_t^{\text{invis}}$, as expressed by $\mathbf{x}_t = \mathbf{x}_t^{\text{vis}} + \mathbf{x}_t^{\text{invis}}$.

Diffusion Network. We implement the diffusion model as a UNet [57, 59] conditioned on the observed region $\mathbf{x}_0$. We input the concatenation of $\mathbf{x}_t$ and $\hat{\mathbf{x}}_0$ into the network.

Classifier-free Guidance. To further enhance the controllability of the generation process, we introduce a classifier-free guidance [22] mechanism. Specifically, we train a unified network ϵ_θ comprising of an unconditional model $\epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)$, where the shared variable $\mathbf{c}$ is incorporated, and a conditional model $\epsilon_\theta(\mathbf{x}_t, \hat{\mathbf{x}}_0, t)$. In this way, the predicted noise $\tilde{\epsilon}_\theta$ can be recomputed as follows:

$$\tilde{\epsilon}_\theta(\mathbf{x}_t, \hat{\mathbf{x}}_0, t) = \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t) + \beta \times [\epsilon_\theta(\mathbf{x}_t, \hat{\mathbf{x}}_0, t) - \epsilon_\theta(\mathbf{x}_t, \mathbf{c}, t)]$$

where $\beta \geq 0$ is the guidance factor, being responsible for the trade-off between sampling quality and diversity [22].

5. Experiments

Dataset. We conducted experiments on the ScanNet-V2 [12] dataset, which was pre-processed by removing redundant frames [69]. For training, we used the first $1,293$ scenes, while for metric evaluation, we randomly selected 18 scenes with over 50 views each from the remaining as our test set. We also evaluated under various sparsity settings (5%, 10%, 20%, and 50%) by uniformly downsampling views.

Comparison. We compared against the neural graphics primitive (NGP) [43], which has demonstrated impressive performance in scene modeling with high efficiency. To enhance its geometric quality, we incorporated a depth supervision (DS) loss [13] to build an improved variant called DS-NGP [13, 43]. We also compare against Neural RGBD (N-RGBD) [3], which recovers implicit surfaces from RGBD scans, and Dense Depth Prior (DDP) [56], which learns a NeRF utilizing view completion from sparse RGBD views.

Evaluation Metrics. For assessing the visual quality of RGB images, we adopted the peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), and learned perceptual image patch similarity (LPIPS) [89] that is based on the AlexNet [29] backbone. To evaluate the geometry quality, we computed the mean squared error (MSE) on depth maps, and sample $10,000$ points uniformly on meshes constructed via back-projection to evaluate the chamfer distance (CD) and completeness (Comp.) with a threshold of 0.1m. We also measured the computational time required to execute different stages of the method. For LPIPS, MSE, and CD, the lower the better; for PSNR, SSIM, and Comp., the higher the better. All reported metrics are averaged across the test scenes.

Implementation Details. Our model has 157M parameters and was trained for three days on 7 NVIDIA 3090-Ti GPUs, using a batch size of 280. The learning rate was initialized at 1×10^{-4} and reduced to 1×10^{-6} over a period of 300 epochs, utilizing a cosine annealing strategy. The image resolution is 128×128 with a rendering chunk size of 7.

5.1. Ablation Studies

We conducted ablation studies to validate the effectiveness of each component in our proposal. Results concluded here are nontrivial and a bit nuanced.

Effects of Different Ingredients. We examined the anticipated efficacy of the proposed conditioning and inpainting components in our approach. Table 2 displays the numerical results, and Figure 5-(a-c) presents the visualizations. The combined use of conditioning and inpainting yields superior visual performance compared to the ground truth. Nevertheless, when the model is conditioned, the impact of inpainting on geometric quality becomes less crucial. Moreover, as more views are provided, the stochastic generation process becomes increasingly deterministic, resulting in structures that more closely resemble the ground truth.

Effects of Guidance Scale. The guidance factor β significantly affects the conditioning effect on the results. To investigate the optimal β under various settings, we conducted experiments with β chosen from 0.0, 0.5, 1.0, 2.0, and 5.0, respectively, and evaluated their performance. Quantitative results are presented in Table 3. It is interesting that only a suitable value of β (1.0 or 2.0) yields optimal performance, while smaller or larger values of β result in underperformance. Surprisingly, for geometric recovery (MSE and CD), the optimal value of β appears to increase as the percentage gets larger. For a scene with a percentage $\geq 50\%$, the best value of β is amazingly greater than 5.0. However, such a large percentage is not optimal for visual appearance. This is because using a larger value of β at a low percentage can cause the generated results to deviate unexpectedly from the ground truth. Our visualization, presented in Figure 5-(a, c, d), indicates that using an unconditional model ($\beta = 0$) leads to undesired and bizarre geometric structures since the network fails to understand the context provided by known views. However, excessively large values of β also oversaturate the color (e.g. row "(d)" at 5%), making the visual appearance unrealistic.

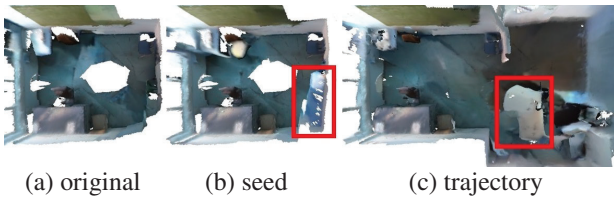**Effects of Randomness and Trajectories.** We investigated

(a) original   (b) seed   (c) trajectory

Figure 4. **Diverse results obtained by changing the (b) random seed and (c) camera trajectory based on (a) the original scene.**

the impact of employing different random seeds and camera trajectories on the results. Such obtained meshes are visualized in Figure 4. As diffusion models inherently include randomness in their reverse sampling process, by switching to a different random seed, we can obtain another outcome. Moreover, the camera trajectory can affect the rough shape of the scene. It is promising that our approach can yield controllable and editable results, as it is capable of generating various and appealing outcomes by simply altering these two factors.

5.2. Generative Scene Synthesis

We evaluated the performance of our method against other similar approaches on this task through extensive experiments under various sparsity settings. Table 1 presents the time required by different methods across various stages. Our approach stands out for its efficiency, as it eliminates the need for additional optimization on individual test scenes during inference. Quantitative results are shown in

Time	Backend	Repre.	Training	Optimization	Rendering
(DS-)NGP [13, 43]	CUDA C++	NeRF	—	∼ 3 min.	∼ 0.05 sec.
N-RGBD [3]	PyTorch	NeRF	—	∼ 2 hours	∼ 2 sec.
DDP [56]	PyTorch	NeRF	∼ 1 day	∼ 1 hour	∼ 1 sec.
Ours	PyTorch	Mesh	∼ 3 days	—	∼ 3 sec.

Table 1. **Efficiency Comparison of Different Methods.** It details their distinct implementation backends, the underlying representation (Repre.) methods, and the time required for *training* on multiple training scenes, *optimization* on a single test scene, and *rendering* a single view after per-scene optimization.

Table 4. We observe that our approach exhibits a clear advantage in visual metrics over the others when the provided views are extremely sparse (5%). Interestingly, our method consistently achieves the best performance in all the geometry-related metrics. Qualitative results are presented in Figure 6. DS-NGP [13, 43] and DDP [56] struggle to accurately recover geometry due to their inability to hallucinate and extrapolate missing regions. In comparison, N-RGBD [3] can achieve better surface completion results by learning and extrapolating neural implicit surfaces. However, their performance significantly degrades when input views are extraordinarily sparse (5%). In contrast, our method consistently exhibits plausible appearance that closely resembles the ground truth, particularly in scenarios with sparse-view inputs.

6. Discussions

Limitations. Our current implementation has several limitations that may impact its usefulness in some scenarios. Firstly, it is incapable of handling color discrepancies caused by lighting variations. Secondly, it lacks surface extrapolation capabilities that can be provided through implicit field representation. Lastly, the limited receptive field of our design is confined to the observable volume of the current camera view and may result in inconsistent and discontinuous predictions, particularly in the case of a large

Factors		Visual												Geometric												
		$\text{PSNR}_{\text{color}}$				$\text{SSIM}_{\text{color}}$				$\text{LPIPS}_{\text{color}}$				$\text{MSE}_{\text{depth}}$				CD_{mesh}				$\text{Comp.}_{\text{mesh}}@0.1m$			
Cond.	Inpa.	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%
		9.33	9.30	9.27	9.45	0.331	0.330	0.330	0.333	0.637	0.636	0.637	0.635	1.309	1.304	1.310	1.293	3061	2777	2463	1758	0.513	0.598	0.711	0.862
	✓	12.4	14.7	16.5	17.9	0.411	0.496	0.557	0.583	0.520	0.446	0.393	0.359	1.001	0.837	0.761	0.730	1934	850	443	149	0.600	0.781	0.881	0.931
✓		12.5	13.6	14.7	16.1	0.444	0.473	0.513	0.556	0.449	0.409	0.362	**0.315**	0.897	0.808	**0.662**	0.595	1163	**699**	176	**99.0**	**0.751**	0.817	0.887	0.928
✓	✓	**14.6**	**16.0**	**17.4**	**18.4**	**0.522**	**0.555**	**0.593**	**0.603**	**0.448**	**0.399**	**0.359**	0.338	**0.825**	**0.805**	0.688	0.628	**1058**	902	**156**	100	0.747	**0.839**	**0.909**	**0.936**

Table 2. **Ablation studies of the effects of conditioning (Cond., $\beta = 1$) and inpainting (Inpa.) on visual (color images) and geometric (depth maps and meshes) results on the task of scene synthesis from sparse RGBD inputs.** The combined use of conditioning and inpainting leads to superior performance in visual metrics, and inpainting plays a less significant role in terms of geometric results when the model is conditioned.
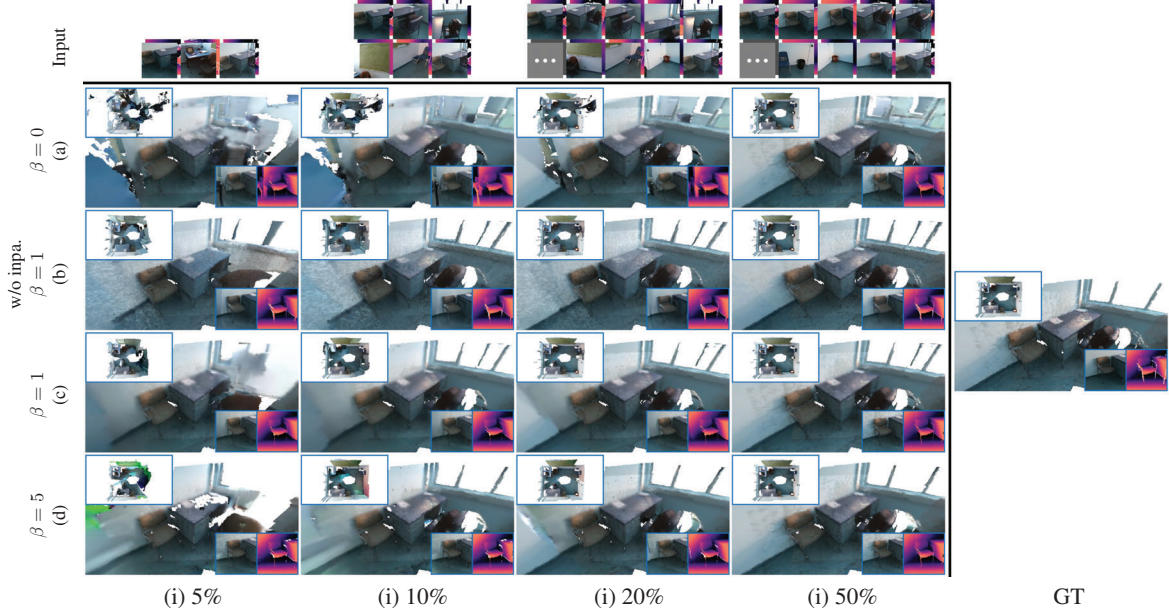

Figure 5. **Qualitative results of ablation studies on the task of scene synthesis from sparse RGBD inputs.** "(a)", "(b)" and "(c)" correspond to the second (w/o conditioning, $\beta = 0$), third (w/o inpainting, $\beta = 1$) and fourth (w/ all, $\beta = 1$) rows of Table 2, and "(d)" corresponds to the last (w/ all, $\beta = 5$) row of Table 3. Each image in the grid consists of four sub-images, with two located at the lower-right corner displaying rendered RGBD images generated from either NeRFs (others) or meshes (ours), and the other two positioned at the top-left corner and underneath, respectively, showcasing the back-projected triangular meshes captured from a close-up perspective. It is clear that the third row basically shows the most favorable appearance.

Guidance Factor β	Visual												Geometric											
	$\text{PSNR}_{\text{color}}$				$\text{SSIM}_{\text{color}}$				$\text{LPIPS}_{\text{color}}$				$\text{MSE}_{\text{depth}}$				CD_{mesh}				$\text{Comp.}_{\text{mesh}}@0.1m$			
	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%
0.0	12.4	14.7	16.5	17.9	0.411	0.496	0.557	0.583	0.520	0.446	0.393	0.359	1.001	0.837	0.761	0.730	1934	850	443	149	0.600	0.781	0.881	0.931
0.5	13.2	15.5	17.1	18.2	0.452	0.530	0.578	0.596	0.496	0.418	0.374	0.347	0.999	0.845	0.772	0.719	1980	606	223	111	0.653	0.818	0.894	0.933
1.0	**14.6**	**16.0**	17.4	18.4	0.522	0.555	0.593	0.603	0.448	0.399	0.359	0.338	**0.825**	0.805	0.688	0.628	**1058**	902	156	100	0.747	0.839	0.909	0.936
2.0	14.5	15.8	**17.5**	**18.4**	**0.532**	**0.561**	**0.598**	**0.606**	**0.439**	**0.393**	**0.352**	**0.334**	0.894	**0.800**	**0.654**	0.593	1562	**515**	**144**	87.2	**0.753**	**0.846**	**0.910**	**0.936**
5.0	13.3	14.9	17.1	18.2	0.488	0.531	0.579	0.598	0.475	0.418	0.367	0.342	0.992	0.856	0.663	**0.582**	2551	1676	175	**87.2**	0.747	0.842	0.908	0.934

Table 3. **Ablation studies of the effects of guidance factor β on visual (color images) and geometric (depth maps and meshes) results on the task of scene synthesis from sparse RGBD inputs.** A larger β ideally strengthens the conditioning effect. The visual metrics are found to be the best when β is set to 1 or 2. Interestingly, the optimal value of β for the best geometry appears to increase as views become denser. However, setting such a large value of β (e.g. $\beta = 5$) leads to sub-optimal visual results.

circular camera trajectory.

Future Works. To improve our design, potential areas of investigation include: (1) Modeling color smoothness and variation, as demonstrated by NeRF [42]. (2) Supporting advanced physical lighting effects, such as SVBRDF, as implemented in TANGO [9]. (3) Incorporating appearance/surface extrapolation by learning an implicit field, such as NeRF [42, 43] or SDF [3, 80]. (4) Exploring gener-

ative [5] or optimizable [6, 84] camera trajectories for scene synthesis. (5) Investigating reconstruction from sparse-view RGB inputs only, using depth inpainting/estimation following the reverse sampling technique proposed in [37] by utilizing a versatile RGBD diffusion model. (6) Leveraging the multi-modal [52] or generative [57] power of large-scale pre-trained models, such as the recently widespread Stable Diffusion [57].

Methods	Visual												Geometric											
	$\text{PSNR}_{\text{color}}$				$\text{SSIM}_{\text{color}}$				$\text{LPIPS}_{\text{color}}$				$\text{MSE}_{\text{depth}}$				CD_{mesh}				$\text{Comp.}_{\text{mesh}}@0.1m$			
	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%
NGP	10.4	12.4	14.4	17.4	0.293	0.377	0.437	0.498	0.582	0.476	0.415	0.376	7.01	7.39	7.37	6.52	29994	22973	16676	8066	0.289	0.446	0.646	0.850
DS-NGP	10.1	11.9	13.5	15.3	0.205	0.281	0.321	0.361	0.605	0.533	0.503	0.476	2.30	1.69	1.44	1.17	5362	1701	768	230	0.529	0.722	0.845	0.918
N-RGBD	14.1	**16.5**	**18.4**	20.0	0.401	0.513	0.595	0.652	0.490	**0.384**	**0.320**	0.287	1.51	1.23	1.21	1.13	3503	1345	1210	643	0.705	0.819	0.857	0.872
DDP	14.1	16.1	18.3	**20.9**	0.418	0.504	**0.599**	**0.702**	0.517	0.410	0.329	**0.259**	1.44	1.05	0.94	0.88	2363	1013	637	451	0.507	0.624	0.719	0.812
Ours	**14.6**	16.0	17.4	18.4	**0.522**	**0.555**	0.593	0.603	**0.448**	0.399	0.359	0.338	**0.82**	**0.80**	**0.68**	**0.62**	**1058**	**902**	**156**	**100**	**0.747**	**0.839**	**0.909**	**0.936**

Table 4. **Quantitative comparison with other approaches on visual (color images) and geometric (depth maps and meshes) results on the task of scene synthesis from sparse RGBD inputs.** Our approach outperforms other methods consistently in terms of geometric metrics as the sparsity varies. However, our method only show a definitive visual advantage across various visual metrics when the input views are highly sparse (5%).

(i) 5%   (i) 10%   (i) 20%   (i) 50%   GT

Figure 6. **Qualitative comparison results with DS-NGP [13, 43], DDP [56], N-RGBD [3] on the task of scene synthesis from sparse RGBD inputs.** Each image in the grid consists of four sub-images, with two located at the lower-right corner displaying rendered RGBD images generated from either NeRFs (others) or meshes (ours), and the other two positioned at the top-left corner and underneath, respectively, showcasing the back-projected triangular meshes captured from a close-up perspective. Our approach produces images that exhibit a close resemblance to the ground truth (GT), while comparative methods may result in inferior outcomes, especially when dealing with excessively sparse input views and outward-looking cameras.

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3D point clouds. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 40–49. PMLR, 10–15 Jul 2018. 2

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18208–18218, June 2022. 2

[3] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural rgb-d surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6290–6301, June 2022. 1, 2, 5, 6, 7, 8

[4] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2022. 2

[5] Miguel Ángel Bautista, Pengsheng Guo, Samira Abnar, Walter Talbott, Alexander T Toshev, Zhuoyuan Chen, Laurent Dinh, Shuangfei Zhai, Hanlin Goh, Daniel Ulbricht, Afshin Dehghan, and Joshua M. Susskind. GAUDI: A neural architect for immersive 3d scene generation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 1, 3, 7

[6] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior, 2022. 7

[7] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16133, June 2022. 3

[8] Eric R. Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. Pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5799–5809, June 2021. 1, 2

[9] Yongwei Chen, Rui Chen, Jiabao Lei, Yabin Zhang, and Kui Jia. TANGO: Text-driven photorealistic and robust 3d stylization via lighting decomposition. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2, 7

[10] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[11] Christopher B Choy, Danfei Xu, JunYoung Gwak, Kevin Chen, and Silvio Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2

[12] 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 (CVPR)*, July 2017. 2, 5

[13] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12882–12891, June 2022. 2, 3, 5, 6, 8

[14] 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 International Conference on Computer Vision (ICCV)*, pages 14304–14313, October 2021. 3

[15] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021. 2

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. 2

[17] Thibault Groueix, Matthew Fisher, Vladimir G. Kim, Bryan C. Russell, and Mathieu Aubry. A papier-mâché approach to learning 3d surface generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 2

[18] Jiatao Gu, Lingjie Liu, Peng Wang, and Christian Theobalt. StyleneRF: A style-based 3d aware generator for high-resolution image synthesis. In *International Conference on Learning Representations*, 2022. 3

[19] Xiaoguang Han, Zhaoxuan Zhang, Dong Du, Mingdai Yang, Jingming Yu, Pan Pan, Xin Yang, Ligang Liu, Zixiang Xiong, and Shuguang Cui. Deep reinforcement learning of volume-guided progressive view inpainting for 3d point scene completion from a single depth image. In *Proceed-*

ings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 2

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 2, 3

[21] Jonathan Ho, Chitwan Saharia, William Chan, David J. Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *Journal of Machine Learning Research*, 23(47):1–33, 2022. 2

[22] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 2, 5

[23] Jonathan Ho, Tim Salimans, Alexey A. Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models. In *ICLR Workshop on Deep Generative Models for Highly Structured Data*, 2022. 2

[24] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, nov 1997. 2

[25] Tobias Höppe, Arash Mehrjou, Stefan Bauer, Didrik Nielsen, and Andrea Dittadi. Diffusion models for video prediction and infilling. *Transactions on Machine Learning Research*, 2022. 2

[26] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2

[27] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2426–2435, June 2022. 2

[28] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. 2

[29] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 5

[30] Jiabao Lei and Kui Jia. Analytic marching: An analytic meshing solution from deep implicit surface networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5789–5798. PMLR, 13–18 Jul 2020. 2

[31] Jiabao Lei, Kui Jia, and Yi Ma. Learning and meshing from deep implicit surface networks using an efficient implementation of analytic marching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):10068–10086, Dec 2022. 2

[32] Haoying Li, Yifan Yang, Meng Chang, Shiqi Chen, Huajun Feng, Zhihai Xu, Qi Li, and Yueting Chen. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479:47–59, 2022. 2

[33] Manyi Li, Akshay Gadi Patil, Kai Xu, Siddhartha Chaudhuri, Owais Khan, Ariel Shamir, Changhe Tu, Baoquan Chen, Daniel Cohen-Or, and Hao Zhang. Grains: Generative recursive autoencoders for indoor scenes. *ACM Trans. Graph.*, 38(2), feb 2019. 3

[34] Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part I*, page 515–534, Berlin, Heidelberg, 2022. Springer-Verlag. 2

[35] Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14458–14467, October 2021. 2

[36] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. In *International Conference on Learning Representations*, 2022. 2

[37] 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. 2, 5, 7

[38] Andrew Luo, Zhoutong Zhang, Jiajun Wu, and Joshua B. Tenenbaum. End-to-end optimization of scene layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 3

[39] 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 (CVPR)*, pages 2837–2845, June 2021. 2

[40] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations*, 2022. 2

[41] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[42] 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 *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, page 405–421, Berlin, Heidelberg, 2020. Springer-Verlag. 1, 2, 3, 4, 7

[43] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4), jul 2022. 2, 3, 5, 6, 7, 8

[44] Charlie Nash, Yaroslav Ganin, S. M. Ali Eslami, and Peter Battaglia. PolyGen: An autoregressive generative model of 3D meshes. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7220–7229. PMLR, 13–18 Jul 2020. 2

[45] Alexander Quinn 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. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16784–16804. PMLR, 17–23 Jul 2022. 2

[46] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5480–5490, June 2022. 3

[47] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. Stylesdf: High-resolution 3d-consistent image and geometry generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13503–13513, June 2022. 3

[48] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[49] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM Trans. Graph.*, 40(6), dec 2021. 2, 3

[50] Despoina Paschalidou, Amlan Kar, Maria Shugrina, Karsten Kreis, Andreas Geiger, and Sanja Fidler. Atiss: Autoregressive transformers for indoor scene synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12013–12026. Curran Associates, Inc., 2021. 3

[51] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *The Eleventh International Conference on Learning Representations*, 2023. 2

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. 7

[53] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *CoRR*, abs/2204.06125, 2022. 2

[54] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1530–1538, Lille, France, 07–09 Jul 2015. PMLR. 2

[55] Daniel Ritchie, Kai Wang, and Yu-An Lin. Fast and flexible indoor scene synthesis via deep convolutional generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 3

[56] Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12892–12901, June 2022. 1, 2, 3, 5, 6, 8

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 2, 3, 5, 7

[58] Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14356–14366, October 2021. 3

[59] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells III, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III*, volume 9351 of *Lecture Notes in Computer Science*, pages 234–241. Springer, 2015. 3, 5

[60] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, SIGGRAPH '22, New York, NY, USA, 2022. Association for Computing Machinery. 2

[61] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2

[62] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(4):4713–4726, April 2023. 2

[63] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 20154–20166. Curran Associates, Inc., 2020. 1, 2

[64] Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 3

[65] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2256–2265, Lille, France, 07–09 Jul 2015. PMLR. 2

[66] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2, 3

[67] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[68] Yongyi Su, Xun Xu, and Kui Jia. Weakly supervised 3d point cloud segmentation via multi-prototype learning, 2022. 2

[69] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15598–15607, June 2021. 5

[70] Jiapeng Tang, Xiaoguang Han, Junyi Pan, Kui Jia, and Xin Tong. A skeleton-bridged deep learning approach for generating meshes of complex topologies from single rgb images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 2

[71] Jiapeng Tang, Xiaoguang Han, Mingkui Tan, Xin Tong, and Kui Jia. Skeletonnet: A topology-preserving solution for learning mesh reconstruction of object surfaces from rgb images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6454–6471, Oct 2022. 2

[72] Jiapeng Tang, Jiabao Lei, Dan Xu, Feiying Ma, Kui Jia, and Lei Zhang. Sa-convonet: Sign-agnostic optimization of convolutional occupancy networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6504–6513, October 2021. 2

[73] Jiapeng Tang, Lev Markhasin, Bi Wang, Justus Thies, and Matthias Nießner. Neural shape deformation priors. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2

[74] Jiapeng Tang, Dan Xu, Kui Jia, and Lei Zhang. Learning parallel dense correspondence from spatio-temporal descriptors for efficient and robust 4d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6022–6031, June 2021. 2

[75] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 2

[76] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. MCVD - masked conditional video diffusion for prediction, generation, and interpolation. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2

[77] Kai Wang, Yu-An Lin, Ben Weissmann, Manolis Savva, Angel X. Chang, and Daniel Ritchie. Planit: Planning and instantiating indoor scenes with relation graph and spatial prior networks. *ACM Trans. Graph.*, 38(4), jul 2019. 3

[78] Kai Wang, Manolis Savva, Angel X. Chang, and Daniel Ritchie. Deep convolutional priors for indoor scene synthesis. *ACM Trans. Graph.*, 37(4), jul 2018. 3

[79] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, September 2018. 2

[80] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 27171–27183. Curran Associates, Inc., 2021. 2, 3, 7

[81] Peng-Shuai Wang, Yang Liu, Yu-Xiao Guo, Chun-Yu Sun, and Xin Tong. O-cnn: Octree-based convolutional neural networks for 3d shape analysis. *ACM Trans. Graph.*, 36(4), jul 2017. 2

[82] Weilun Wang, Jianmin Bao, Wengang Zhou, Dongdong Chen, Dong Chen, Lu Yuan, and Houqiang Li. Semantic image synthesis via diffusion models, 2022. 2

[83] X. Wang, C. Yeshwanth, and M. Niesner. Sceneformer: Indoor scene generation with transformers. In *2021 International Conference on 3D Vision (3DV)*, pages 106–115, Los Alamitos, CA, USA, dec 2021. IEEE Computer Society. 3

[84] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters, 2021. 7

[85] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. 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. 2

[86] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In

H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. 2

[87] 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 (ICCV)*, October 2019. 2

[88] Biao Zhang, Jiapeng Tang, Matthias Niessner, and Peter Wonka. 3dshape2vecset: A 3d shape representation for neural fields and generative diffusion models, 2023. 2

[89] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018. 5

[90] Zaiwei Zhang, Zhenpei Yang, Chongyang Ma, Linjie Luo, Alexander Huth, Etienne Vouga, and Qixing Huang. Deep generative modeling for scene synthesis via hybrid representations. *ACM Trans. Graph.*, 39(2), apr 2020. 3

[91] Wenbin Zhao, Jiabao Lei, Yuxin Wen, Jianguo Zhang, and Kui Jia. Sign-agnostic implicit learning of surface self-similarities for shape modeling and reconstruction from raw point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10256–10265, June 2021. 2

[92] 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 (ICCV)*, pages 5826–5835, October 2021. 2