
PDF: Point Diffusion Implicit Function for Large-scale Scene Neural Representation

Yuhan Ding^{1*}

dingyh22@m.fudan.edu.cn

Fukun Yin^{1*}

fkyin21@m.fudan.edu.cn

Jiayuan Fan^{2†}

yfan@fudan.edu.cn

Hui Li²

lihui21@m.fudan.edu.cn

Xin Chen³

chenxin2@shanghaitech.edu.cn liuwen@shanghaitech.edu.cn

Wen Liu³

liuwen@shanghaitech.edu.cn

ChongShan Lu¹

cslu17@fudan.edu.cn

Gang Yu³

iskicy@gmail.com

Tao Chen¹

eetchen@fudan.edu.cn

¹ School of Information Science and Technology, Fudan University, China

² Academy for Engineering and Technology, Fudan University, China

³ Tencent PCG, Shanghai, China

Abstract

Recent advances in implicit neural representations have achieved impressive results by sampling and fusing individual points along sampling rays in the sampling space. However, due to the explosively growing sampling space, finely representing and synthesizing detailed textures remains a challenge for unbounded large-scale outdoor scenes. To alleviate the dilemma of using individual points to perceive the entire colossal space, we explore learning the surface distribution of the scene to provide structural priors and reduce the samplable space and propose a **Point Diffusion implicit Function, PDF**, for large-scale scene neural representation. The core of our method is a large-scale point cloud super-resolution diffusion module that enhances the sparse point cloud reconstructed from several training images into a dense point cloud as an explicit prior. Then in the rendering stage, only sampling points with prior points within the sampling radius are retained. That is, the sampling space is reduced from the unbounded space to the scene surface. Meanwhile, to fill in the background of the scene that cannot be provided by point clouds, the region sampling based on Mip-NeRF 360 is employed to model the background representation. Expensive experiments have demonstrated the effectiveness of our method for large-scale scene novel view synthesis, which outperforms relevant state-of-the-art baselines.

1 Introduction

Implicit neural representations can handle single objects or small scenes well, and they are widely used in the fields of virtual reality [5], 3D reconstruction [21, 31, 36], video generation [7] and computer animation [20, 1, 41] on tasks such as scene representation and new perspective synthesis. However, when the target scene is enlarged, such as urban-scale outdoor scenes, the performance of traditional implicit neural representation methods will be severely compromised. The problem is that

*Joint first authors.

†Corresponding author.

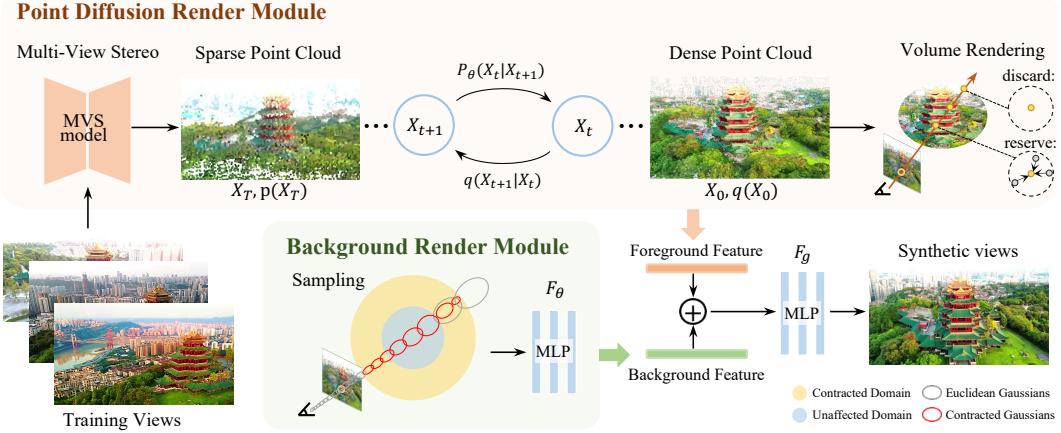


Figure 1: The pipeline of our point diffusion implicit function. Our method consists of two modules, a point diffusion rendering module and a background rendering module. The former learns the surface distribution of the scene through a diffusion-based point cloud super-resolution model and renders foreground features from the dense point cloud surface. The latter follows Mip-NeRF 360’s strategy to render background features. Finally, the foreground and background features are fused to generate photo-realistic novel views for large-scale outdoor scenes.

when the scene expands, its sampling space grows at the cubic level, so it is difficult for individual sampling points to cover the entire space.

Fortunately, some methods try to solve this problem in two ways, narrowing the sampling space and expanding the sampling area. For the first approach, the method represented by Mega-NeRF [28] decomposes the sampling space into multiple subspaces and models each subspace separately to reduce complexity. But with the scale of the scene growing, such as reaching the city level, the number of subspaces also increases cubically. For the second approach, the method represented by Mip-NeRF 360 [2] compresses the sampling space or samples an area instead of a single point so that the sampling space can be filled more easily. But this also brings a loss of precision.

Benefiting from the inspiration of geometric priors aiding vision tasks, which are widely used in the domain of 3D reconstruction and stereo vision [16, 3, 6]. We are curious whether implicit large scene representations could be made easier with explicit representations. Moreover, for outdoor unbounded large scenes, most of the sampling space is filled with air rather than buildings, cars, plants and other objects that we care about. A reasonable solution is to restrict the large-scale sampling space of the implicit neural representation to the object surface, which is provided by the scene geometry prior. That is to say, we compress a 3D sampling space to a 2D surface plane, which will greatly reduce the representation complexity. At the same time, the network will pay more attention to the foreground, which is the same as the human visual perception system. Of course, for the neglected background information that is relatively less important, we can provide a relatively less accurate expression by compressing the scene to sample the area in space.

In this paper, we propose **PDF**, a **P**oint **D**iffusion **I**mplicit **F**unction for large-scale scene neural representation, which learns a dense surface distribution via a diffusion-based point prior generative model to reduce the sampling space. To achieve this, we first explore a large-scale outdoor point cloud augmentation method based on the Point-Voxel Diffusion model [43]. Since point clouds of real outdoor scenes often lack dense ground truth, it is difficult to train a completion module through "sparse-dense" point cloud pairs. Therefore, we train a point cloud super-resolution network to downsample the point cloud twice and generate the denser one from the sparser one, which can generate a dense point cloud without ground truth. With the help of the surface point cloud, the sampling points will be retained only if there are reconstruction points within a certain radius, so the space will be greatly reduced to the scene surface. However, the reconstructed point cloud can only model the scene surface and cannot deal with the unbounded background of outdoor scenes. So we follow the idea of NeRF++ [41] and model the foreground and background separately, which can be distinguished by whether there is a sampling point to find the neighborhood point in the point cloud for each sampling ray. Finally, we use Mip-NeRF 360 [2] to sample regions in scene space to extract features as background features.

Extensive experiments show the effectiveness of our point diffusion implicit function for large-scale scene neural representation, which achieves photo-realistic rendering results and outperforms state-of-the-art methods on OMMO [15] and BlendMVS dataset [35]. We summarize the contributions as follows: **1)** Aiming at novel view synthesis for large outdoor scenes, we propose an implicit neural representation framework based on point diffusion models to provide dense surface priors to cope with the exploding sampling space. **2)** A novel point cloud super-resolution diffusion module is proposed to generate dense surface points from sparse point clouds without dense annotations. **3)** Extensive experiments demonstrate that our PDF network outperforms state-of-the-art methods, including robustness to large-scale outdoor scene representation and the capability to synthesize more photo-realistic novel views. Our code and models will be available.

2 Related Work and Background

2.1 Implicit Neural Representation

In recent years, Implicit Neural Representation (INR) has witnessed significant advancements and provides a versatile framework for representing complex functions and generating high-dimensional data [23, 17, 26, 37]. By implicitly encoding the scene’s appearance and geometry, neural radiance fields enable highly realistic rendering and novel view synthesis [25, 18, 8, 19].

Building upon this foundation, subsequent research has focused on addressing the limitations and pushing the boundaries of INR. Efforts have been made to improve the efficiency and scalability of neural radiance fields. For instance, Hanocka et al. proposed DeepSDF[22], which leverages signed distance functions to implicitly represent 3D shapes. This formulation allows for efficient ray-marching and facilitates tasks such as shape manipulation and interpolation. Furthermore, recent advancements in INR have explored differentiable rendering and differentiable volumetric rendering, enabling the incorporation of geometric and physical priors [10, 27, 4] into the representation. These methods leverage the differentiable nature of neural networks to optimize scene parameters, leading to improved realism and control over the generated content [13, 39, 9, 30]. Another significant extension to the field of INR is PixelNeRF [38]. It extends the capabilities of INR to handle images, going beyond the realm of 3D scenes. PixelNeRF introduces a new differentiable sampler to handle image-based representations, enabling efficient and accurate sampling of pixels from the neural radiance field. In addition to PixelNeRF, Semantic Neural Radiance Fields[12] proposed a method to learn scene representations that capture geometry, appearance, and semantic information, facilitating interactive virtual scene editing and content creation.

Overall, these advancements in Implicit Neural Representation have greatly expanded the capabilities of INR. These developments offer promising avenues for realistic image synthesis, shape completion, scene reconstruction, and dynamic content generation. The ongoing research in this field holds great potential for further advancements in computer graphics, computer vision, and virtual reality applications.

2.2 Large-scale Scene Representation

Large-scale scene representation is a crucial aspect of Implicit Neural Representation (INR) research, particularly in the context of computer graphics and computer vision. It involves capturing and modeling complex scenes that encompass extensive spatial extents, such as urban environments, landscapes, or virtual worlds.

One notable work in the domain of large-scale scene representation is Neural Scene Flow Fields[14]. This paper introduces a novel approach to model dynamic scenes at a large scale. The authors propose a scene flow field representation that captures both the geometry and motion of objects in the scene. By leveraging a neural network architecture, they achieve accurate and temporally consistent scene synthesis and reconstruction, even in highly complex and dynamic scenes. The Neural 3D Mesh Renderer[11] is another significant contribution in large-scale scene representation. This work addresses the challenge of representing and rendering detailed 3D meshes of large-scale scenes efficiently. The authors propose a neural network-based renderer that predicts view-dependent textures and geometric details of the scene. This approach enables real-time rendering and interaction with large-scale 3D scenes, opening up possibilities for interactive virtual reality experiences and immersive simulations. In addition to these works, Mega-NeRF [28] and Bungee-NeRF [33] are

two other notable approaches based on the neural radiance field for constructing interactive 3D environments from large-scale visual captures. They address the challenges of modeling and rendering large-scale scenes, spanning from buildings to multiple city blocks and utilizing thousands of images captured from drones. They extend the capabilities of NeRF to handle multi-scale rendering, capturing various levels of detail and enabling the interactive exploration of diverse 3D environments.

Overall, the field of large-scale scene representation within Implicit Neural Representation has witnessed significant progress. These contributions have paved the way for realistic, interactive, and semantically meaningful representations of expansive virtual environments, urban landscapes, and dynamic scenes. The ongoing research in this area holds great potential for further advancements in computer graphics, virtual reality, and immersive simulations.

3 Methodology

In this paper, we aim to develop a novel point diffusion model implicit function to reduce the sampling space and improve the ability to represent large-scale scenes (*c.f.* Fig.1). Our PDF network mainly consists of two modules, a diffusion-based point cloud super-resolution and rendering foreground module and a region-sampling-based background module. The former introduces a diffusion model network to enhance the sparse point cloud reconstructed from the input image into a dense point cloud, which provides optional points in the rendering stage to reduce the sampling space (*c.f.* Sec.3.1). The latter samples regions rather than individual points from unbounded scenes so that it is easy to fill sampled regions and complement the background for new viewpoint synthesis (*c.f.* Sec.3.2). In the final subsection, implementation details and losses are elaborated (*c.f.* Sec.3.3).

3.1 Point Upsampling Diffusion

In this section, we introduce our large-scale outdoor point cloud super-resolution module based on a denoising diffusion probabilistic model (*c.f.* Fig.2).

Point cloud pair preparation. Due to the lack of dense large-scale outdoor point cloud ground truth, we need to train a diffusion-based super-resolution network to sample a dense surface $x_d \in \mathbb{R}^{N \times 3}$ from the point cloud reconstructed by COLMAP [24] and denoted as $x_s \in \mathbb{R}^{M \times 3}$. At the same time, in order to prevent over-fitting, the point cloud reconstructed from the training views should not be used as the ground truth of the diffusion model, but the sparse point cloud $z_0 \in \mathbb{R}^{n \times 3}$ and sparser point cloud $x_0 \in \mathbb{R}^{m \times 3}$ pairs should be down-sampled on this basis as the training data, where $m < n < M < N$. More specifically, we downsample the sparse point cloud x_s reconstructed by COLMAP to get an even sparser point cloud z_0 . Then we further downsample z_0 to get the sparsest point cloud x_0 , where x_s , z_0 and x_0 have progressively sparser relationships. Our training process recovers z_0 from the sparsest x_0 . During testing, we take x_s as input to generate a denser super-resolved point cloud x_d .

Point Super-resolution Diffusion. Our point super-resolution denoising diffusion probabilistic model is a generative model, which starts with Gaussian noise and progressively denoises to generate object shapes. We record the output containing different levels of noise produced by each step as \hat{x}_T , $\hat{x}_{T-1}, \dots, \hat{x}_0$, where \hat{x}_T is sampled from Gaussian noise, and \hat{x}_0 represents the generated point cloud with dense surface. Since we already have a sparse point cloud prior z_0 , our target point cloud can be denoted as $x_0 = (z_0, \hat{x}_0)$ and the intermediate point cloud during the denoising process can be denoted as $x_t = (z_0, \hat{x}_t)$. So next we define a point super-resolution diffusion process involving a prior shape z_0 , consisting of a forward process and a backward process.

Forward process. Gaussian noise is repeatedly added to the original point cloud x_0 , resulting in a series of noisy point clouds x_1, x_2, \dots, x_T :

$$q(\hat{x}_t | \hat{x}_{t-1}, z_0) \sim \mathcal{N}(\hat{x}_t; \sqrt{1 - \beta_t} \hat{x}_{t-1}, \beta_t I) \quad (1)$$

where β_t is a pre-determined increasing sequence of Gaussian noise values and controls how much noise is added in each step.

Reverse process. Given a point cloud with more noise x_t , reverse the forward process and find the posterior distribution for a less noisy one x_{t-0} :

$$p_\theta(\hat{x}_{t-1} | \hat{x}_t, z_0) \sim \mathcal{N}(\mu_\theta(x_t, z_0, t), \sigma_t^2 I) \quad (2)$$

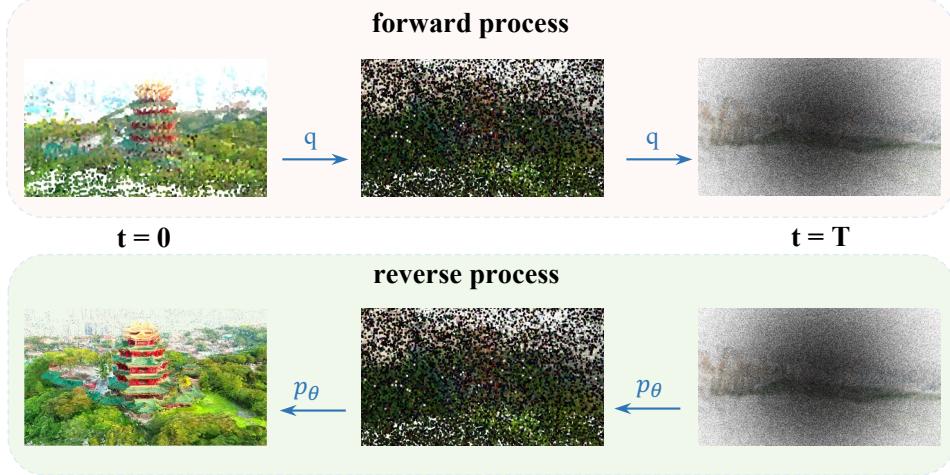


Figure 2: Our point upsampling diffusion. In the forward process, Gaussian noise is gradually added to the sparse point cloud. In the reverse process, the noise is gradually removed to obtain a dense point cloud surface.

where $\mu_\theta(x_t, z_0, t)$ is the predicted shape at $t - 1$ step.

Therefore, our point cloud upsampling diffusion model can be regarded as a noise adding and denoising process. The former gradually adds random noise to the initial point cloud x_0 through the forward process; the latter denoises sequentially through the reverse process to obtain a dense point cloud x_0 . Based on Markov transition probabilities, the whole process can be expressed as:

$$q(\hat{x}_{0:T}, z_0) = q(\hat{x}_0, z_0) \prod_{t=1}^T q(\hat{x}_t | \hat{x}_{t-1}, z_0) \quad (3)$$

$$p_\theta(\hat{x}_{0:T}, z_0) = p(\hat{x}_T, z_0) \prod_{t=1}^T p_\theta(\hat{x}_{t-1} | \hat{x}_t, z_0) \quad (4)$$

Throughout the optimization process, our prior shape z_0 is fixed, and only the missing surface point cloud is diffused. The network is typically trained with a simplified $L2$ denoising loss:

$$\mathcal{L}_D = \|\epsilon - \epsilon_\theta(\hat{x}_t, z_0, t)\|^2 \quad (5)$$

where ϵ is the added random noise and $\epsilon \sim \mathcal{N}(0, I)$, and $\epsilon_\theta(\hat{x}_t, z_0, t)$ is the prediction noise output. Since point cloud prior z_0 is fixed, it will be masked when minimizing the loss.

3.2 Volume Rendering and Implicit Function Representation

Foreground Rendering. For the foreground, we sample points p_i along the rays from the dense point cloud x_0 and render features in the neighborhood, following Point-NeRF [34]. The difference is that our point super-resolution diffusion module only generates denser point cloud coordinates without color information, so we redesign a more general per-point aggregate module. For each sampling point p_i , we query K neighboring neural points $kp_i = \{kp_i^1, kp_i^2, \dots, kp_i^K\}$ around it within a certain Euclidean distance radius R . Then we perceive the local geometric structure as a feature f_i of each sampling point p_i to equip structural information. So we use c_i and kc_i to represent the coordinates of the sampling point and the neighborhood points respectively, and encode geometric differences kg_i of neighborhood as:

$$kg_i = MLP(c_i \oplus c_j \oplus (c_i - c_j) \oplus d(c_i, c_j)) \quad (6)$$

where $d(,)$ is the Euclidean distance between two points, and \oplus is the concatenation operator. Next, the local geometric features f_i of the sampling points p_i are obtained by neighborhood points kp_i weighted summation:

$$f_i = SUM(softmax(MLP(kg_i)) \odot kg_i) \quad (7)$$

where softmax operation is performed on each dimension, and \odot is the hadamard product. Our point-based radiance field can be abstracted as a neural module that regresses the volume density σ and view-dependent radiance r from coordinates c , local geometric features f , and ray direction d according to Point-NeRF [34]:

$$(\sigma, r) = \text{Point-NeRF}(c, d, f) \quad (8)$$

Finally, the foreground feature is synthesized by each neural sampling point along the sampling ray.

Background Rendering. Since point clouds can only cover the foreground but cannot handle unbounded backgrounds, we need to extract extra background features. Benefiting from Mip-NeRF 360 [2], which contracts the scene to a bounded ball and then samples a region to meet the challenge of large scenes, we use it to extract background features as a supplement along the same sampling ray as Point-NeRF [34].

Fore-Background Fusion. Since the detail-preserving foreground features can be obtained from the dense surface points, while the bounded domain can cope with large scenes but loses details during the compression process. So we propose a foreground-background fusion module consisting of several layers of multi-layer perceptrons to preserve their respective advantages.

We adopt L_2 loss to supervise our rendered pixels r_p from ray marching with the ground truth r_g , to optimize our PDF volume render reconstruction network.

$$\mathcal{L}_{\mathcal{R}} = \|r_p - r_g\|^2 \quad (9)$$

3.3 Implementation Details

Our PDF method is a two-stage neural representation network for outdoor unbounded large-scale scenes. We optimize these two stages separately.

In the first stage, a diffusion-based point cloud super-resolution network is designed to learn a prior distribution to generate a dense point cloud surface. In the point cloud pair preparation process, we employed the random down-sampling method with a retention rate between 0.2 and 1 for both samplings. For point super-resolution diffusion, we set $T = 1000$, $\beta_0=10^{-4}$, $\beta_T=0.01$ and linearly interpolate other β 's for all experiments. We use Adam optimizer with learning rate 2×10^{-4} and train on 4 A100 GPUs for around one day.

In the second stage, the foreground and background extraction modules plus a feature fusion module are optimized. We find 8 neighbors for each sampling point and expand the dimension of neighborhood geometric features to 8. Both the foreground and the background output a 128-dimensional feature, and then they are concatenated and passed through 4 MLP layers to get the color of the rendered point. We train this stage using Adam optimizer with an initial learning rate 5×10^{-4} for 2×10^6 iterations about 20 hours on a single A100 GPU.

4 Experiments

4.1 Experimental settings

Dataset. We use two outdoor large-scale scene datasets, OMMO [15] and BlendedMVS [35], to evaluate our model. The OMMO dataset is a real fly-view large-scale outdoor multi-modal dataset, containing complex objects and scenes with calibrated images, prompt annotations and point clouds. The number of training point cloud samples in the OMMO dataset varies from 40,000 to 100,000 for different scenes, including abundant real-world urban and natural scenes with various scales, camera trajectories, and lighting conditions. More experimental results can be found in our supplementary material.

Baselines and Evaluation Metrics. We compare our method with the previous state-of-the art methods on novel view synthesis, including NeRF [20], NeRF++ [41], Mip-NeRF [1], Mip-NeRF 360 [2], Mega-NeRF [28], Ref-NeRF [29]. NeRF is the first continuous MLP-based neural network for synthesizing photo-realistic views of a scene through volume rendering. NeRF++ models large-scale unbounded scenes by separately modeling foreground and background neural representations.

Mip-NeRF reduces aliasing artifacts and better represents fine details by using anti-aliasing cone sampling. Mip-NeRF 360 models large unbounded scenes using non-linear scene parameterization, online distillation, and distortion-based regularization. Mega-NeRF uses a sparse structure and geometric clustering algorithm to decompose the scenes. Ref-NeRF improves synthesized views by restructuring radiance and regularizing normal vectors. To evaluate the performance of each method for large-scale implicit neural representation, we use three common metrics for novel view synthesis: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM [32]), and Learned Perceptual Image Patch Similarity (LPIPS [42]). Higher PSNR and SSIM indicate better performance, while lower LPIPS indicates better performance.

4.2 Performance Comparison

Quantitative Results. Quantitative comparisons on the OMMO [15] dataset are shown in Tab.1, including the mean PSNR, SSIM, and LPIPS. We outperform other methods on all average evaluation metrics, especially lpips, a perceptual metric close to the human visual system, which is significantly more sensitive to the foreground than the background. So better LPIPS metric indicates that our model can better reconstruct the foreground of the scene, benefiting from sampling the foreground from the reconstructed dense point cloud surface instead of the entire sampling space.

NeRF [20], NeRF++ [41], Mip-NeRF [1], and Ref-NeRF [29] are not specially designed for large-scale scenes, so directly applying them to large scenes will lead to performance degradation. Mip-NeRF 360 [2] and Mega-NeRF [28] have achieved the optimal performance in one or several scenes by sampling regions in the limited sampling space or subdividing the sampling space. But it is still not as good as ours in most scenes due to the loss of detail caused by compression or decomposing the sampling space.

Table 1: Quantitative results of our PDF method with the baselines on the OMMO dataset. \uparrow means the higher, the better.

Scene ID	NeRF[20]			NeRF++[41]			Mip-NeRF[1]			Mip-NeRF 360[2]			Mega-NeRF[28]			Ref-NeRF[29]			Ours		
	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow	PSNR \uparrow	SSIM \uparrow	LPIPS \downarrow
1	16.93	0.37	0.744	16.86	0.36	0.780	16.84	0.37	0.793	13.91	0.31	0.771	16.12	0.34	0.782	15.10	0.34	0.755	14.80	0.32	0.705
2	15.31	0.44	0.694	14.89	0.47	0.653	15.16	0.40	0.731	15.06	0.44	0.646	15.64	0.47	0.679	15.90	0.49	0.632	19.63	0.62	0.374
3	14.38	0.28	0.556	14.64	0.29	0.547	14.56	0.29	0.533	14.25	0.31	0.526	15.21	0.33	0.517	15.44	0.37	0.526	14.74	0.34	0.515
4	25.39	0.86	0.431	27.47	0.90	0.380	21.78	0.76	0.469	27.68	0.94	0.292	23.36	0.86	0.419	27.86	0.91	0.404	31.74	0.94	0.202
5	22.26	0.67	0.531	24.32	0.73	0.450	14.98	0.54	0.633	25.76	0.80	0.317	25.78	0.76	0.433	23.54	0.71	0.491	27.58	0.90	0.162
6	24.09	0.68	0.504	25.59	0.75	0.396	23.18	0.66	0.529	28.86	0.90	0.211	24.92	0.77	0.393	26.07	0.72	0.459	23.69	0.87	0.212
7	5.36	0.17	0.747	21.93	0.71	0.542	15.57	0.64	0.624	23.05	0.73	0.523	22.33	0.69	0.552	25.79	0.73	0.511	21.46	0.81	0.193
8	21.14	0.50	0.594	22.91	0.57	0.509	19.82	0.46	0.638	25.07	0.71	0.354	16.65	0.48	0.431	21.21	0.49	0.604	27.62	0.92	0.101
9	14.92	0.34	0.744	14.57	0.34	0.732	14.58	0.34	0.746	15.40	0.30	0.706	17.32	0.49	0.234	0.43	0.649	15.77	0.49	0.381	
10	22.26	0.55	0.626	24.37	0.60	0.578	19.80	0.53	0.643	26.68	0.72	0.420	21.78	0.62	0.558	24.23	0.58	0.597	25.74	0.83	0.136
11	22.36	0.82	0.420	24.61	0.85	0.342	22.81	0.82	0.423	27.06	0.93	0.217	24.37	0.84	0.355	30.29	0.95	0.188			
12	22.41	0.59	0.533	24.29	0.68	0.447	22.13	0.60	0.526	28.12	0.83	0.274	21.60	0.62	0.493	23.06	0.60	0.524	27.92	0.86	0.063
13	22.27	0.59	0.608	23.52	0.62	0.581	18.90	0.54	0.673	26.63	0.77	0.403	25.50	0.72	0.517	23.29	0.61	0.594	25.94	0.74	0.205
14	19.85	0.55	0.569	23.89	0.74	0.417	17.06	0.48	0.655	28.06	0.89	0.224	24.42	0.75	0.411	21.76	0.63	0.508	28.11	0.94	0.127
15	20.35	0.53	0.552	21.71	0.61	0.490	19.44	0.49	0.594	28.63	0.89	0.179	22.69	0.67	0.445	20.33	0.50	0.576	27.22	0.89	0.136
16	17.86	0.40	0.631	18.75	0.41	0.597	18.49	0.40	0.610	10.01	0.34	0.850	20.26	0.53	0.509	19.64	0.43	0.572	18.70	0.47	0.392
17	22.02	0.57	0.610	24.20	0.67	0.461	17.01	0.53	0.696	29.53	0.83	0.247	17.23	0.57	0.529	23.17	0.59	0.529	26.59	0.88	0.111
18	26.06	0.75	0.428	25.57	0.73	0.461	24.61	0.73	0.469	28.55	0.86	0.265	24.76	0.73	0.448	22.79	0.67	0.569	28.07	0.91	0.152
19	14.20	0.40	0.726	13.86	0.37	0.703	13.84	0.39	0.738	14.72	0.37	0.676	23.81	0.68	0.465	14.34	0.39	0.691	27.55	0.84	0.170
20	22.84	0.61	0.499	23.28	0.64	0.475	22.41	0.60	0.519	28.33	0.86	0.228	21.11	0.63	0.499	21.54	0.55	0.574	26.88	0.81	0.197
21	22.59	0.51	0.532	21.84	0.47	0.593	22.31	0.51	0.537	25.64	0.75	0.344	21.92	0.51	0.578	21.07	0.44	0.672	28.62	0.94	0.141
22	16.53	0.47	0.733	20.66	0.56	0.575	13.37	0.42	0.776	24.79	0.77	0.362	20.84	0.60	0.527	20.31	0.53	0.615	26.33	0.85	0.074
23	18.99	0.41	0.669	19.51	0.42	0.597	18.09	0.39	0.671	21.25	0.51	0.539	20.13	0.44	0.585	19.94	0.41	0.622	21.64	0.65	0.206
24	19.32	0.39	0.696	23.14	0.52	0.535	16.89	0.37	0.715	25.86	0.71	0.373	23.87	0.56	0.518	22.17	0.45	0.616	30.90	0.87	0.097
25	24.72	0.55	0.528	22.42	0.51	0.613	24.24	0.54	0.542	28.91	0.79	0.306	25.98	0.63	0.457	23.62	0.50	0.598	30.85	0.94	0.083
26	8.56	0.24	0.564	19.94	0.59	0.513	13.43	0.35	0.688	14.59	0.46	0.626	19.23	0.67	0.467	21.00	0.62	0.489	23.88	0.83	0.311
27	4.54	0.01	0.705	21.25	0.55	0.546	14.82	0.45	0.674	21.26	0.60	0.235	20.59	0.61	0.543	20.82	0.52	0.590	21.77	0.66	0.164
28	24.48	0.66	0.479	23.28	0.64	0.475	24.76	0.66	0.406	29.62	0.87	0.240	25.87	0.72	0.442	22.17	0.45	0.616	29.22	0.91	0.153
29	22.98	0.61	0.540	23.17	0.62	0.529	23.01	0.61	0.539	25.51	0.74	0.400	21.57	0.61	0.557	21.11	0.54	0.631	25.86	0.84	0.174
30	20.23	0.52	0.605	23.27	0.64	0.476	18.63	0.46	0.675	26.54	0.84	0.296	24.04	0.69	0.459	21.62	0.54	0.588	26.10	0.93	0.096
31	18.97	0.37	0.645	19.05	0.37	0.643	18.91	0.36	0.659	13.08	0.23	0.708	20.93	0.60	0.545	19.18	0.37	0.645	26.68	0.90	0.208
32	17.99	0.58	0.621	18.99	0.61	0.540	11.28	0.42	0.687	17.16	0.57	0.601	21.29	0.70	0.475	18.98	0.60	0.565	23.43	0.69	0.142
33	5.79	0.01	0.745	20.19	0.50	0.597	14.31	0.42	0.755	22.76	0.63	0.457	22.89	0.64	0.478	21.23	0.52	0.578	22.91	0.75	0.134
Mean	18.72	0.48	0.600	21.45	0.58	0.538	18.39	0.50	0.623	23.10	0.67	0.419	21.63	0.62	0.508	21.28	0.55	0.574	25.10	0.79	0.205

Qualitative Results. Qualitative results on the OMMO [15] dataset are shown in Fig.3. We can see that the rendering results of NeRF [20] and Mip-NeRF [1] are of the lowest quality, they use global MLPs for the entire space to reconstruct radiance fields, resulting in a trade-off in the accuracy of sampling the foreground and background, which makes them almost impossible to handle large-scale unbounded scenes. NeRF++ [41], Mega-NeRF [28] and Ref-NeRF [29] improve some limitations of NeRF by corresponding techniques, but the rendering results are often missing details, especially when the scene contains a lot of intricate details. The rendering quality of Mip-NeRF 360 [2] is relatively high, but loses some detail and edges due to its down-scaling of the scene into a limited sampling space. Our method uses the dense point cloud up-sampled by the diffusion model as a detailed foreground geometry prior combined with Mip-NeRF 360 background features, so our model can reconstruct the fine foreground texture provided by the generative model. At the same time,

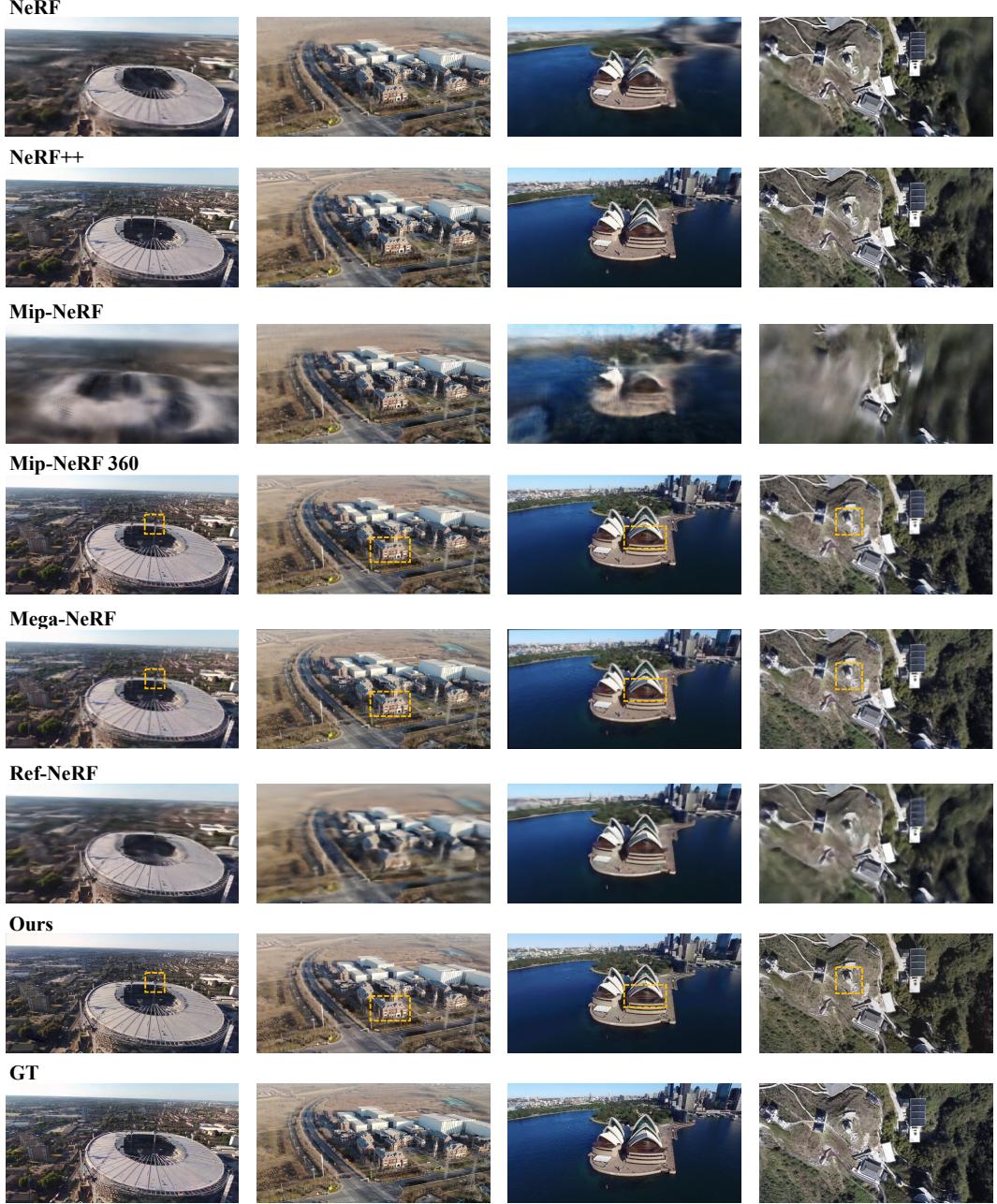


Figure 3: Qualitative results of our method with the baselines on the OMMO dataset. Our PDF method outperforms baseline methods with reliably constructed details. For Mip-NeRF and Mega-NeRF, which are also aimed at large scenes, we use yellow dashed boxes to mark some areas that are easy to distinguish the performance of details. Please zoom-in for the best of views.

compared with Mip-NeRF 360, our method is more robust to scene representation and new view generation without failing scenes (*c.f.* Fig.4).

4.3 Ablation Studies

We perform multiple ablation studies to validate the effectiveness of our proposed modules. Tab.2 shows the impact of diffusion point cloud super-resolution module and background feature fusion module on the 5-th scene (sydney opera house) from the OMMO dataset [15].

For the ablation experiment on the effectiveness of diffusion, we remove the diffusion-based point cloud up-sampling module and sample directly on the sparse point cloud reconstructed by COLMAP [24] from training views. Since the directly reconstructed point cloud is very sparse and concentrated in the central area, only a very blurry image with large missing blocks can be rendered, as shown in the first column of Fig.5. At the same time, quantitative indicators also suggest that this method is not suitable for outdoor unbounded large-scale scenes with its PSNR 9.28.

For the ablation experiment on the effectiveness of background fusion, we remove the background fusion module and render novel view images directly from the diffusion-enhanced point cloud. As shown in the second column of Fig.5, with the help of the dense point cloud produced by the diffusion module learning the scene distribution, we find that large missing patches have been filled in and produce a more refined foreground. However, limited by the characteristics of point cloud expression, the background points are very sparse, which leads to blurred background rendering results. Quantitative results, while substantially improved, still convey poor image quality.

As shown in the third column of Fig.5, using the background fusion module alone can also fill in the missing blocks of the background, but due to the sparseness of the point cloud reconstructed by COLMAP [24], it will lead to the loss of detail and blurring of the rendering result. However, our method, which combines a diffusion module and a background fusion module, achieves satisfactory quantitative and qualitative performance and surpasses existing methods.

We also perform ablation experiments to compare our method with other point cloud up-sampling methods. With the same experimental setup, we use a GAN-based method [40] for point cloud up-sampling instead of the diffusion-based up-sampling module. Tab.3 shows the quantitative results for three scenes(scan5, scan11 and scan12) in the OMMO dataset. Our method exhibits superior performance compared to the GAN-based point cloud up-sampling method, primarily due to its ability to preserve the structural and topological characteristics of point clouds while effectively handling incomplete or noisy point cloud data. In addition, Fig.6 shows the visualization results of the diffusion-based point cloud up-sampling module, and our method can not only densify the sparse point cloud reconstructed by the COLMAP, but also fill in the missing regions of the point cloud such as the background and empty space.

Table 2: Quantitative performance of ablation experiments, including removing both the diffusion-based point cloud up-sampling module and the background fusion module, removing only the diffusion-based point cloud up-sampling module, removing only the background fusion module, our PDF method.

Method	PSNR↑	SSIM↑	LPIPS↓
w/o diffusion, w/o background	9.28	0.51	0.355
w/o diffusion, w/ background	21.05	0.83	0.219
w/ diffusion, w/o background	22.93	0.78	0.235
Ours	27.58	0.90	0.162

5 Conclusions and Limitations

In this paper, we propose PDF, a point diffusion implicit function for large-scale scene neural representation, and demonstrate its robustness and fidelity on novel view synthesis tasks. The core of our method is to provide dense point cloud surface priors to reduce the huge sampling space of large-scale scenes. Therefore, a point cloud super-resolution module based on diffusion model is

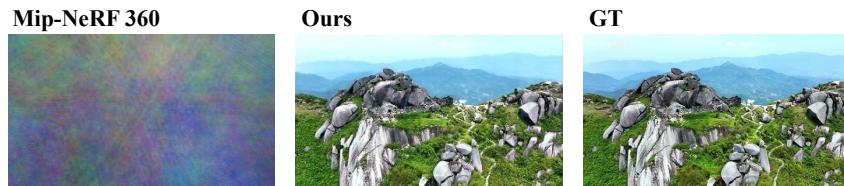


Figure 4: A failure scene representation of Mip-NeRF 360.



Figure 5: Qualitative performance of ablation experiments. From left to right: removing both the diffusion-based point cloud up-sampling module and the background fusion module, removing only the background fusion module, removing only the diffusion-based point cloud up-sampling module, our PDF method, and the groundtruth.

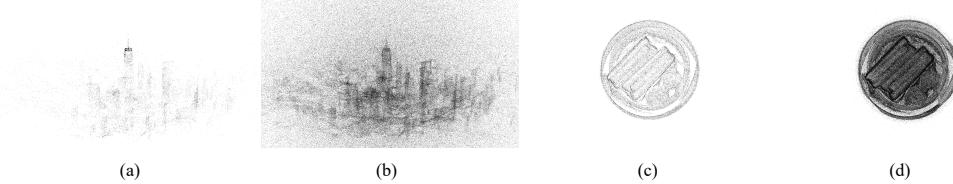


Figure 6: Qualitative evaluation of the diffusion-based point cloud super-resolution module. From left to right: (a) Point cloud of a large-scale scene reconstructed using COLMAP. (b) Point cloud of the same large-scale scene enhanced using our method. (c) Point cloud of a small-scale scene reconstructed using COLMAP. (d) Point cloud of the same small-scale scene enhanced using our method. (zoom-in for the best view).

Table 3: Quantitative results of ablation experiments, including removing the diffusion-based point cloud up-sampling module, using the GAN-based point cloud up-sampling method, our PDF method.

Method	PSNR↑	SSIM↑	LPIPS↓
w/o diffusion	21.85	0.84	0.204
GAN-based method [40]	24.83	0.86	0.161
Ours	28.60	0.90	0.137

proposed to learn from the sparse point cloud surface distribution reconstructed from training views to generate more dense point clouds. However, only constraining the sampling space to the point cloud surface does not fully solve the novel view synthesis problem since point clouds do not have background information. So Mip-NeRF 360 [2] is employed to provide background features and synthesize photo-realistic new perspectives. Extensive experiments demonstrate that our method outperforms current methods in both subjective and objective aspects. At the same time, ablation experiments also prove the effectiveness of our core module, point up-sampling diffusion.

In future work, we will attempt to explore a cross-scene point cloud up-sampling generalization diffusion model instead of training a diffusion model for each scene to improve efficiency. Even more futuristicly, it may be possible to extract representative scene representations and inject them into reconstructed point clouds to achieve cross-scene rendering, i.e., generalized point diffusion NeRF.

Acknowledgements

This work is supported by National Natural Science Foundation of China (No. 62101137, 62071127, and U1909207), Shanghai Natural Science Foundation (No.23ZR1402900), and Zhejiang Lab Project (No.2021KH0AB05).

References

- [1] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021.
- [2] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.
- [3] 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, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.
- [4] Tianlong Chen, Peihao Wang, Zhiwen Fan, and Zhangyang Wang. Aug-nerf: Training stronger neural radiance fields with triple-level physically-grounded augmentations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15191–15202, 2022.
- [5] Nianchen Deng, Zhenyi He, Jiannan Ye, Budmonde Duinkharjav, Praneeth Chakravarthula, Xubo Yang, and Qi Sun. Fov-nerf: Foveated neural radiance fields for virtual reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(11):3854–3864, 2022.
- [6] Yuan-Chen Guo, Di Kang, Linchao Bao, Yu He, and Song-Hai Zhang. Nerfren: Neural radiance fields with reflections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18409–18418, 2022.
- [7] Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5784–5794, 2021.
- [8] Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. Hdr-nerf: High dynamic range neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18398–18408, 2022.
- [9] Yi-Hua Huang, Yue He, Yu-Jie Yuan, Yu-Kun Lai, and Lin Gao. Stylizednerf: consistent 3d scene stylization as stylized nerf via 2d-3d mutual learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18342–18352, 2022.
- [10] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, 2022.
- [11] Hiroharu Kato, Yoshitaka Ushiku, and Tatsuya Harada. Neural 3d mesh renderer. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3907–3916. Computer Vision Foundation / IEEE Computer Society, 2018.
- [12] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *NeurIPS*, 2022.
- [13] Verica Lazova, Vladimir Guzov, Kyle Olszewski, Sergey Tulyakov, and Gerard Pons-Moll. Control-nerf: Editable feature volumes for scene rendering and manipulation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4340–4350, 2023.
- [14] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 6498–6508. Computer Vision Foundation / IEEE, 2021.
- [15] Chongshan Lu, Fukun Yin, Xin Chen, Tao Chen, Gang Yu, and Jiayuan Fan. A large-scale outdoor multi-modal dataset and benchmark for novel view synthesis and implicit scene reconstruction. *arXiv preprint arXiv:2301.06782*, 2023.
- [16] Baorui Ma, Yu-Shen Liu, and Zhizhong Han. Reconstructing surfaces for sparse point clouds with on-surface priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6315–6325, 2022.

- [17] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM 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*, pages 7210–7219, 2021.
- [18] Ben Mildenhall. *Neural Scene Representations for View Synthesis*. PhD thesis, University of California, Berkeley, USA, 2020.
- [19] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16190–16199, 2022.
- [20] 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. *Communications of the ACM*, 65(1):99–106, 2021.
- [21] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5589–5599, October 2021.
- [22] Jeong Joon Park, Peter R. Florence, Julian Straub, Richard A. Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 165–174. Computer Vision Foundation / IEEE, 2019.
- [23] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021.
- [24] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.
- [25] Shih-Yang Su, Frank Yu, Michael Zollhöfer, and Helge Rhodin. A-nerf: Articulated neural radiance fields for learning human shape, appearance, and pose. *Advances in Neural Information Processing Systems*, 34:12278–12291, 2021.
- [26] Matthew Tancik, Vincent Casser, Xincheng Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P 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*, pages 8248–8258, 2022.
- [27] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul P. Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Nießner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhöfer, and Vladislav Golyanik. Advances in neural rendering. *Comput. Graph. Forum*, 41(2):703–735, 2022.
- [28] Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs, 2022.
- [29] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022.
- [30] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, 2022.
- [31] 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. *arXiv preprint arXiv:2106.10689*, 2021.
- [32] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

- [33] Yuanbo Xiangli, Lining Xu, Xingang Pan, Nanxuan Zhao, Anyi Rao, Christian Theobalt, Bo Dai, and Dahua Lin. Bungeenerf: Progressive neural radiance field for extreme multi-scale scene rendering. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 106–122. Springer, 2022.
- [34] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5438–5448, 2022.
- [35] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020.
- [36] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021.
- [37] Fukun Yin, Wen Liu, Zilong Huang, Pei Cheng, Tao Chen, and Gang YU. Coordinates are NOT lonely - codebook prior helps implicit neural 3d representations. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.
- [38] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19–25, 2021*, pages 4578–4587. Computer Vision Foundation / IEEE, 2021.
- [39] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. Nerf-editing: geometry editing of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18353–18364, 2022.
- [40] Junzhe Zhang, Xinyi Chen, Zhongang Cai, Liang Pan, Haiyu Zhao, Shuai Yi, Chai Kiat Yeo, Bo Dai, and Chen Change Loy. Unsupervised 3d shape completion through gan inversion. In *CVPR*, 2021.
- [41] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.
- [42] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.
- [43] 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.