

Frequency-Modulated Point Cloud Rendering with Easy Editing

Yi Zhang^{1*} Xiaoyang Huang^{1*} Bingbing Ni^{1†} Teng Li² Wenjun Zhang¹

¹Shanghai Jiao Tong University, Shanghai 200240, China ²Anhui University

{yizhangphd, huangxiaoyang, nibingbing}@sjtu.edu.cn

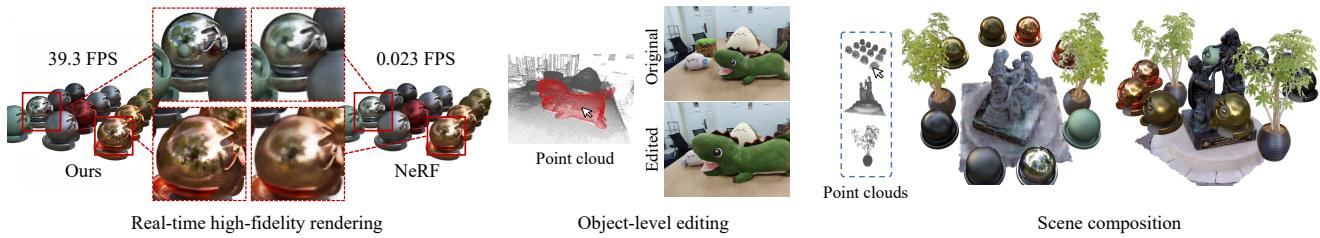


Figure 1. We propose a novel frequency-modulated point cloud rendering pipeline, which enables high fidelity local detail reconstruction, real-time rendering and user-friendly editing. Specifically, we can render sharper textures with $1700\times$ faster than NeRF [21] and support both object-level editing and scene composition.

Abstract

We develop an effective point cloud rendering pipeline for novel view synthesis, which enables high fidelity local detail reconstruction, real-time rendering and user-friendly editing. In the heart of our pipeline is an adaptive frequency modulation module called Adaptive Frequency Net (AFNet), which utilizes a hypernetwork to learn the local texture frequency encoding that is consecutively injected into adaptive frequency activation layers to modulate the implicit radiance signal. This mechanism improves the frequency expressive ability of the network with richer frequency basis support, only at a small computational budget. To further boost performance, a preprocessing module is also proposed for point cloud geometry optimization via point opacity estimation. In contrast to implicit rendering, our pipeline supports high-fidelity interactive editing based on point cloud manipulation. Extensive experimental results on NeRF-Synthetic, ScanNet, DTU and Tanks and Temples datasets demonstrate the superior performances achieved by our method in terms of PSNR, SSIM and LPIPS, in comparison to the state-of-the-art. Code is released at <https://github.com/yizhangphd/FreqPCR>.

1. Introduction

Photo-realistic rendering and editing of 3D representations is a key problem in 3D computer vision and graphics with numerous applications, such as computer games,

VR/AR, and video creation. In particular, recently introduced neural radiance field (NeRF) [21] has inspired some follow-up works aiming to editable rendering [15, 17, 20, 48, 50, 56]. Due to the deeply coupled black-box network, NeRF-based object-level editing usually requires a pre-trained segmentation model to separate the objects to be edited [17, 56]. Although some recent voxel-based variants of NeRF [48, 60] achieve multi-scene composition, they still lack the ability to extract target objects from voxels.

In contrast to implicit rendering, point cloud rendering [1, 6, 13, 18, 33, 36, 59] is a promising editable rendering model. On the one hand, explicit 3D representations are better for interactive editing. On the other hand, the geometric priors provided by point clouds can help us avoid massive sampling in volume rendering methods, which can meet the requirements of some real-time applications. As a class of representative point cloud rendering methods, NPBG and NPBG++ [1, 33] achieve real-time rendering by using point-wise features for encoding appearance information and an U-Net [35] for decoding, respectively. However, the parameter quantity increases with the size of point cloud, which may limit their application due to the excessive computational and memory complexity. Huang *et al.* [13] combine point clouds with implicit rendering, where explicit point clouds are used to estimate the geometry, and implicit radiance mapping is used to predict view-dependent appearance of surfaces. However, quantitative evaluation of its rendering results is significantly lower than that of implicit rendering methods such as NeRF [21], mainly due to the following reasons: 1) the color of each viewing ray only depends on

*Equal contribution.

†Corresponding author: Bingbing Ni.

a single surface point, thus without multiple sample color aggregation for error attenuation, surface based rendering techniques require radiance mapping to have a more precise and expressive frequency encoding ability; and 2) defects of the point cloud geometry reconstructed by MVSNet [57] cause wrong surface estimation. To this end, we introduce Adaptive Frequency Net (AFNet) to improve frequency expression ability of the radiance mapping and a preprocessing module for point cloud geometry optimization.

Radiance mapping, also known as radiance field, is a type of Implicit Neural Representation (INR). There have been some studies [2, 7, 32, 42, 46, 47, 62] on the expressive power and inductive bias of INRs. The standard Multi-layer Perceptrons (MLPs) with ReLU activation function are well known for the strong spectral bias towards reconstructing low frequency signals [32]. Some recent works [7, 42, 47] introduce strategies to enhance the high-frequency representation of MLPs from a global perspective. However, from a local perspective, the frequencies of a 3D scene are region-dependent and most real objects are composed by both weak textured regions and strong textured ones. Motivated by this, we design a novel adaptive frequency modulation mechanism based on HyperNetwork architecture [11], which learns the local texture frequency and injects it into adaptive frequency activation layers to modulate the implicit radiance signal. The proposed mechanism can predict suitable frequency without frequency supervision and modulate the radiance signal with adaptive frequency basis support to express more complex textures at negligible computational overhead.

Previous surface point-based works [1, 13, 33] could not optimize the point cloud geometry because they keep only the closest point as a surface estimation for each ray. But if we sample multiple points per ray during rendering, it will greatly reduce the rendering speed. Therefore, we use the volume rendering method as a preprocessing module to optimize the point cloud geometry. Specifically, we keep more points in the pixel buffer and learn the opacity of each point based on volume rendering. We find in our experiments that for some poorly reconstructed scenes, point cloud geometry optimization can improve the rendering PSNR by 2-4dB and avoid rendering artifacts.

For rigid object editing, we follow the deformation field construction [28–31, 49] to render the edited point cloud. Point cloud can be seen as a bridge between user editing and deformation field to achieve interactive editing and rendering. Users only need to edit the point cloud, and the deformation field between the original 3D space and the deformed space is easy to obtain by querying the corresponding transformations performed on point cloud. Moreover, to avoid cross-scene training in multi-scene composition application, we develop a masking strategy based on depth buffer to combine multiple scenes in pixel level.

We evaluate our method on NeRF-Synthetic [21], ScanNet [5], DTU [14] and Tanks and Temples [16] datasets and comprehensively compare the proposed method with other works in terms of performance (including PSNR, SSIM and LPIPS), model complexity, rendering speed, and editing ability. Our performance outperforms NeRF [21] and all surface point-based rendering methods [1, 13, 33], and is comparable to Compressible-composable NeRF (CC-NeRF), *i.e.*, the latest NeRF-based editable variant. We achieve a real-time rendering speed of 39.27 FPS on NeRF-Synthetic, which is 1700 \times faster than NeRF and 37 \times faster than CCNeRF. We also reproduce the scene editing of Object-NeRF [56] and CCNeRF [48] on ToyDesk [56] and NeRF-Synthetic [21] dataset, respectively. As shown in Fig. 1, we achieve real-time rendering with sharper details and user-friendly editing. The above results demonstrate that our method is comprehensive in terms of both rendering and editing and has great application potentials.

2. Related Work

Implicit Neural Representations. The goal of INRs is to encode a continuous target signal using a neural network, by representing the mapping between input coordinates and signal values. This technique has been used to represent various objects such as images [23, 44], shapes [4, 9, 27], scenes [21, 43], and textures [12, 25]. Tancik *et al.* [47] propose Fourier Feature Network (FFN), showing that feeding input coordinates through a Fourier feature mapping enables MLP to learn high-frequency functions in low dimensional domains. Sitzmann *et al.* [42] propose to use MLP with sinusoidal activations. Fathony *et al.* propose MFN [7], which employs a Hadamard product between linear layers and nonlinear activation functions. In [62], Yüce *et al.* prove that FFN [47], SIREN [42] and MFN [7] have the same expressive power as a structured signal dictionary.

3D Representations for View Synthesis. Implicit representations often use neural networks to encode 3D scenes [21, 24, 26, 27, 40, 51, 52, 63]. NeRF [21] uses a 5D function to represent the scene and applies volumetric rendering for view synthesis. Many recent works combine NeRF with some explicit representation to improve rendering quality or speed. NSVF [19] uses a sparse voxel octree to achieve fast rendering. Yu *et al.* propose PlenOctrees [60], a data structure derived from NeRF [21] which enables highly efficient rendering. Plenoxels [8] represents a scene as a sparse 3D grid with spherical harmonics. Xu *et al.* [53] propose Point-NeRF, which is a localized neural representation, combining volumetric radiance fields with point clouds. Instant-NGP [22] uses multi-resolution hashing for efficient encoding and also leads to high compactness. TensoRF [3] factorizes the scene tensor into lower-rank components. Huang *et al.* [13] com-

bine point cloud with implicit radiance mapping to achieve the state-of-the-art performance of point cloud rendering. However, their performance is still worse than NeRF [21]. In this work, we greatly improve the performance of point cloud rendering by introducing AFNet and a preprocessing module for geometry optimization.

Editable Rendering. Editing ability is important for 3D representations. Some recent works [10, 45, 54–56, 61] improve NeRF to editable version. Yang *et al.* [56] propose Object-NeRF for editable rendering, while it requires a 2D segmentation model to extract the target object and renders at a slow speed. Yuan *et al.* propose NeRF-Editing [61], which allows users to perform controllable shape deformation on the implicit representation. Yang *et al.* [55] present a mesh-based representation by encoding the implicit field with disentangled geometry and texture codes on mesh vertices, which facilitates a set of editing functionalities. Tang *et al.* propose CCNeRF [48], which could transform and compose several scenes into one scene with a small model.

3. Method

In this section, we first revisit the state-of-the-art point cloud rendering technique and discuss its dilemma of insufficient frequency representation, in Sec. 3.1. In Sec. 3.2, we present our proposed frequency-modulated radiance mapping technique to address above limitation, followed by the introduction to our end-to-end point cloud rendering pipeline in Sec. 3.3. We also introduce a novel point cloud geometry optimization pipeline as an optional preprocessing module in Sec. 3.4. Applications of the above pipeline on interactive object level editing and multi-scene composition are presented in Sec. 3.5.

3.1. Revisiting Point Cloud Rendering

Some traditional point cloud rendering methods focus on fast rendering of large-scale point clouds [38, 39], while they are not as photorealistic as neural rendering (*e.g.*, not smooth and lack of view-dependent appearance). Different from traditional works, our approach builds on a state-of-the-art point-based neural rendering method [13], which utilizes point cloud to determine the position of the surface by the following coordinate rectification

$$\mathbf{x} = \mathbf{o} + \frac{z}{\cos \theta} \times \mathbf{d}, \quad (1)$$

where \mathbf{o} , \mathbf{d} and θ denote camera position, normalized ray direction and angle between the ray and the z -axis pointing to the center of the scene, respectively. In practice, z value is obtained by querying the depth buffer during rasterization. As point clouds are discrete, each point is expanded to a disk of radius r during rasterization [65]. The estimated surface coordinates \mathbf{x} together with ray direction \mathbf{d} are fed

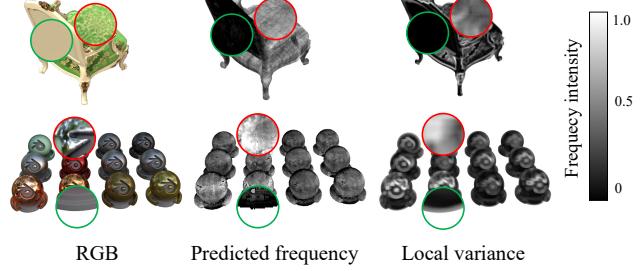


Figure 2. Original RGB images, frequency maps predicted by our hypernetwork, and local variance of original images for reference. We find that the frequencies are large for strongly textured regions (shown as gray and white color), and the frequencies are close to zeros for weakly textured regions (shown as black color).

to an MLP \mathbf{F}_Θ (Θ is the network parameter set) to obtain the features \mathbf{L} which encode the color of the surface point

$$\mathbf{L} = \mathbf{F}_\Theta(\mathbf{x}, \mathbf{d}). \quad (2)$$

Compared to volume rendering-based models [21, 53], this method avoids heavy sampling by estimating surface coordinates from point clouds when calculating ray color, and thus achieving real-time rendering. As point clouds are usually noisy and the initial depth obtained by rasterization might also be inaccurate, previous methods [1, 13, 33] often use U-Net structure [35] for further refinement.

Limitations. The limitations of the above surface point-based rendering scheme are obvious. Namely, the point cloud quality is usually poor due to sparsity and large number of outliers, which leads to lower-fidelity rendering especially in high frequency shapes, *i.e.*, highly-3D-textured regions. Simply increasing the network can not solve the problem. On the contrary, prevailing volumetric rendering techniques such as NeRF [21] do not suffer from this issue (*i.e.*, usually with higher rendering quality), since through each sampled ray there exist multiple sampling points near the surface fused together to contribute the target color, which is more stable and sufficient for dealing with high frequency shapes. Therefore, it is highly demanded to enhance the frequency expression ability of point cloud based implicit surface radiance mapping method at a relatively low computational cost overhead.

3.2. Frequency-Modulated Radiance Mapping

Motivation. To achieve radiance mapping, also known as radiance field, previous works input a spatial coordinate and infer the color or other features associated with this spatial position via standard MLP with ReLU activation layer. However, this classical neural network is well-known for its strong spectral bias towards generating lower frequency patterns [32], which has obvious difficulty in handling high frequency shape details, *e.g.*, 3D textures of objects and occupancy of edges. Recent works [7, 42, 47] analyze the expres-

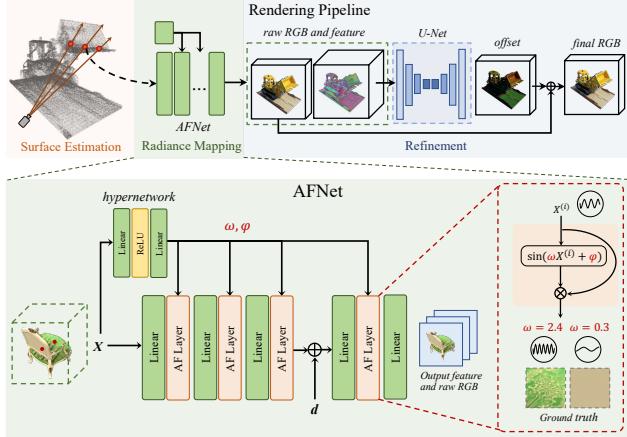


Figure 3. Overview of our rendering pipeline. In the heart of the pipeline is an adaptive frequency modulation module called AFNet, which utilizes a hypernetwork to learn local texture frequency encoding that is consecutively injected to adaptive frequency activation layers to modulate the implicit radiance signal.

sive power and inductive bias of the network and enhance the fitting ability of the network from a global perspective, while we emphasize the local perspective of the problem, *i.e.*, the frequencies of a 3D scene are region-dependent. Namely, most real objects are composed by both weak textured regions and strong textured ones. For example, in the Chair scene of NeRF-Synthetic dataset, the back of the chair belongs to weak texture areas, and the cushion with patterns belongs to strong texture areas, as shown in Fig. 2. If the signal frequency does not match the target area, the implicit radiance mapping would be poor due to misaligned frequency functional support. Thus, the above observations motivate us to propose an Adaptive Frequency Net (AFNet) for frequency-modulated radiance mapping, as shown in Fig. 3, which is capable of selecting high frequency primitives for modeling rich-3D-textured regions while uses low frequency basis for presenting non-textured regions, in order to yield overall high-fidelity rendering quality.

Adaptive Frequency Estimation and Modulation. The major working mechanism of the proposed AFNet is to first estimate the local frequency information for each point, and then this frequency encoding is injected to the proposed Adaptive Frequency Activation Layers (AF Layers) which adaptively select a proper set of frequency component, providing a more precise support for generating high-fidelity reconstruction, as shown in Fig. 3. Therefore, the core building block of the proposed AFNet includes 1) AF Layers and 2) a hypernetwork for local frequency estimation, which are introduced in detail as follows.

Adaptive Frequency Activation Layer. Assuming local frequency information is successfully estimated and encoded as scalar frequencies and phases, our AF layers first encode the input radiance signal with a nonlinear filter (we

use a sinusoidal filter) parameterized by the obtained pointwise adaptive frequency and phase. Then, a Hadamard product between the filtered feature and the input feature is taken as the activated output, which could be regarded as a frequency modulation on the input signal. Mathematically, suppose that the input feature of i -th layer is $\mathbf{X}^{(i)} \in \mathbb{R}^{M \times C}$, in which M and C are point number and feature dimension respectively, and the adaptive frequencies and phases are $\omega^{(i)}, \phi^{(i)} \in \mathbb{R}^{M \times 1}$, whose dimensions will be broadcast to $\mathbb{R}^{M \times C}$ during computation. Then the adaptive frequency activation layers can be expressed as:

$$\mathbf{Y}^{(i)} = \mathbf{X}^{(i)} \circ \sin(\omega^{(i)} \circ \mathbf{X}^{(i)} + \phi^{(i)}). \quad (3)$$

The design of Eqn. 3 has a theoretical justification. We measure the local frequency by the derivative of the color with respect to the spatial coordinate, and adaptively adjust the derivative by multiplying the feature $\mathbf{X}^{(i)}$ by the predicted frequency $\omega^{(i)}$. We then use the sine function to constrain the numerical range. The final form, $\mathbf{X}^{(i)} \circ \sin(\cdot)$, is borrowed from MFN [7].

Hypernet for Frequency Estimation. The adaptive frequencies and phases are expected to contain important information of the 3D texture variation on the surface position; however, to obtain the accurate frequency representation of each point is non-trivial. Inspired by HyperNetwork [11], an approach of using a small network to generate the weights for a larger network, we use a two-layer MLP as hypernetwork to infer the frequency and phase parameters from each spatial point. In Fig. 2, we show the predicted frequency of the 4-th adaptive frequency activation layer. As seen in the RGB images, the back of the chair and the base of the ball have weak texture, while the frequency predicted by hypernetwork is closed to 0 (shown as black color). For strongly textured regions, such as patterned cushions in chair and the metal material, the hypernetwork predicts large frequency (shown as grey and white). These observations empirically indicates that the proposed hypernetwork is capable of learning correct frequency information even there is **NO** supervision on frequency label.

3.3. End-to-end Rendering Pipeline

Our pipeline consists of three stages: surface estimation, radiance mapping and refinement, as shown in Fig. 3. In the surface estimation stage, we first rasterize the point cloud $\mathbf{P} = \{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_N\} \in \mathbb{R}^{N \times 3}$ to obtain the depth buffer $\mathbf{z} \in \mathbb{R}^{H \times W}$ at a given view. Specifically, we transform the point cloud from the world coordinate space to the normalized device coordinate sapce, in which $\mathbf{p}_k \in [-1, 1] \times [-1, 1] \times [z_{near}, z_{far}], k = 1, 2, \dots, N$. Similarly, we transform the pixel coordinates $\mathbf{I} \in \mathbb{R}^{H \times W \times 2}$ to normalized space such that $\mathbf{I}_{i,j} \in [-1, 1]^2$. Then, each point \mathbf{p}_k is expanded to a disk of radius r_k in xOy plane.

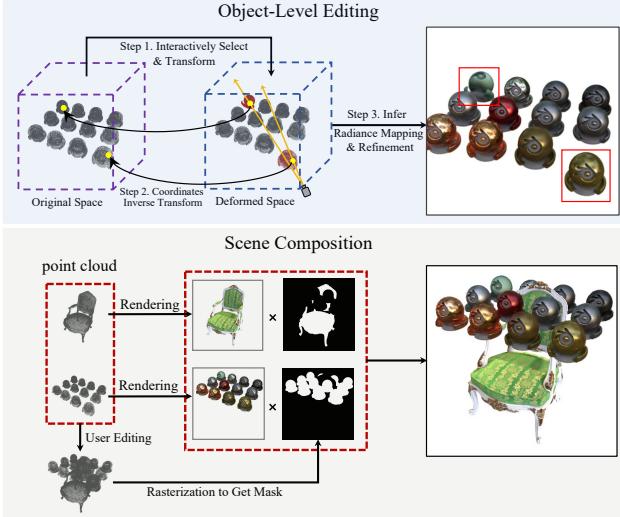


Figure 4. Object-level editing and scene composition pipelines.

The depth buffer $\mathbf{z}_{i,j}$ is the depth of the nearest disk covering pixel $\mathbf{I}_{i,j}$, that is

$$\mathbf{z}_{i,j} = \min_k \mathbf{p}_k^z, \quad s.t. (\mathbf{p}_k^x - \mathbf{I}_{i,j}^x)^2 + (\mathbf{p}_k^y - \mathbf{I}_{i,j}^y)^2 < r_k^2, \quad (4)$$

which can be seen as a special case of surface splatting [65]. Then, we use Eqn. 1 to estimate the coordinate of the surface where each ray intersects.

In radiance mapping stage, the estimated surface coordinates are fed into the proposed AFNet to predict the raw RGB values and feature vectors. In the refinement stage, the feature vectors will be fed into an U-Net to predict the offsets. The final RGBs are obtained by adding the raw RGB values and offsets. AFNet and U-Net are trained with ground truth images as supervision by using L2 loss and perceptual loss [64].

3.4. Preprocessing: Geometry Optimization

In the meantime, point clouds obtained by MVS method [57] or 3D scans inevitably contain noise or missing parts of the surface, which harms the rendering quality or even produces artifacts. In view of this, we design a point cloud geometry optimization strategy consisting of a point cloud denoising step and a completion step, considered as an optional preprocessing module in our framework, to further enhance rendering quality.

Denoising. We design a volume rendering-based pipeline specifically to optimize the point cloud geometry. In rasterization, we maintain the depth buffer of k points instead of the nearest point. The depth values are used to compute the rectified coordinates of the points using Eqn. 1. The rectified coordinates, combined with ray directions, are fed into a MLP to estimate colors. Besides, an optimizable opacity parameter α_i is assigned to each point \mathbf{p}_i . Pixel RGB values are obtained using the volume rendering equation. Inspired

by Point-NeRF [53], we train the MLP and the opacity parameters to fit the ground-truth images with the following sparse regularization:

$$\mathcal{L}_{\text{sparse}} = \frac{1}{N} \sum_{i=1}^N [\log(\alpha_i) + \log(1 - \alpha_i)]. \quad (5)$$

When the training converges, we remove those low-opacity points with a threshold μ .

Completion. For each view in training set, if a pixel in the ground truth image is occupied, but the buffer obtained by rasterization is empty, we add a set of points with opacity parameters along the pixel ray in the point clouds. To determine the depth range of the added points, we refer to the depth range of adjacent pixel buffer. Although the completion step introduces additional noise, it would be removed in the denoising stage of the next iteration.

Above denoising and completion steps run alternatively until convergence. Please see supplementary material for more details.

3.5. Scene Editing

In this section, we introduce an editing method based on the proposed rendering pipeline. At object level, we leverage point clouds to bridge the gap between user editing and deformation field. At scene level, the proposed method can combine different scenes without cross-scene training.

Object-level Editing. As an explicit 3D representation, point clouds are ideal for interactive editing. After editing target points, we can estimate the new surface coordinates of the edited point cloud by rasterization. Since radiance mapping encodes appearance of the surface in original 3D space, we need to transform the estimated surface points to original encoding space, which can be regarded as a deformation of the original space, *i.e.*, deformation field [28,31]. Specifically, we keep the indices of the points in pixel buffer during rasterization and query the transformations performed on those points. Then we perform the corresponding inverse transformation on the estimated surface coordinates to obtain the locations in original 3D space, which replaces the calculation of deformation field, as shown in Fig. 4. Finally, we feed the transformed surface points to radiance mapping module followed by refinement module to obtain the rendering result.

Scene Composition. A direct way to render a combined point cloud is to use respective AFNets for radiance mapping, and a shared U-Net for refinement. However, due to the inductive bias of U-Net, the texture of different scenes will affect each other, which would reduce the rendering quality. Moreover, cross-scene training of U-Net impairs the flexibility of editing, *i.e.*, adding a new scene to an existing composition requires retraining together. To solve the above problems, we compose the scenes in pixel level. As

Method	NeRF-Synthetic			Tanks&Temples			ScanNet			DTU		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRF [21]	31.01	0.947	0.081	25.78	0.864	0.198	25.74	0.780	0.537	26.92	0.909	0.198
CCNeRF-CP [48]	30.55	0.935	0.076	27.01	0.879	0.180	24.65	0.774	0.542	26.79	0.907	0.178
CCNeRF-HY-S [48]	31.22	0.947	0.074	27.53	0.901	0.177	25.17	0.781	0.539	27.23	0.910	0.171
NPBG [1]	28.10	0.923	0.077	25.97	0.889	0.137	25.09	0.737	0.459	26.00	0.895	0.130
NPBG++ [33]	28.12	0.928	0.076	26.04	0.892	0.130	25.27	0.772	0.448	26.08	0.895	0.131
Huang <i>et al.</i> [13]	28.96	0.932	0.061	26.35	0.893	0.130	25.88	0.794	0.415	26.22	0.900	0.132
Ours	31.24	0.950	0.049	27.79	0.902	0.125	26.66	0.803	0.400	27.27	0.908	0.129

Table 1. Quantitative evaluation on NeRF-Synthetic [21], Tanks and Temples [16], ScanNet [5] and DTU [14] datasets. Since our work focuses on surface point-based method, which supports real-time and editable rendering, we mainly compare with NPBG [1], NPBG++ [33] and Huang *et al.* [13]. We also show the performance of NeRF [21] and CCNeRF [48] for reference.

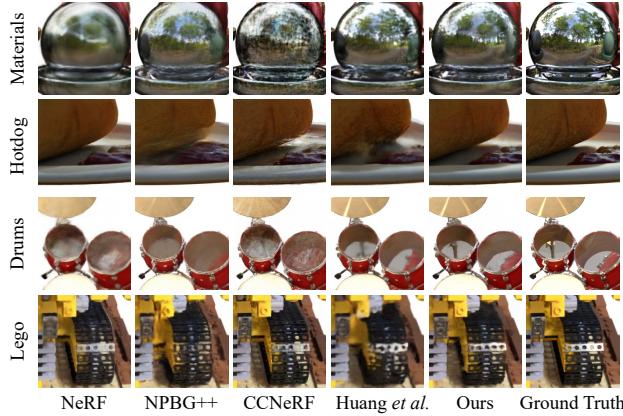


Figure 5. Qualitative evaluations on NeRF-Synthetic [21] dataset. Our method renders sharper detailed textures without artifacts.

shown in Fig. 4, we first rasterize the combined point clouds to determine which scene each pixel belongs to and generate a mask for each scene. Then we render each scene separately and multiply the masks with the rendered images. We combine the masked images to obtain the final result.

4. Experiments

4.1. Experimental Settings

Datasets and Compared Methods. Since our work focuses on surface point-based rendering methods, which enable real-time and editable rendering, we mainly compare with the following methods: 1) NPBG [1]: A famous surface point-based rendering method which enables real-time rendering and editing; 2) NPBG++ [33]: The improved version of NPBG; and 3) Huang *et al.* [13]: The state-of-the-art surface point-based rendering pipeline. We first compare our performance with the above methods on NeRF-Synthetic [21], Tanks and Temples [16], ScanNet [5] and DTU [14] datasets. To further demonstrate the superiority of our method, we conduct a comprehensive comparison with volume rendering-based methods, including NeRF [21] and its variants, on NeRF-Synthetic [21] and Tanks and Temples [16] datasets. Please see supplementary material for more details.

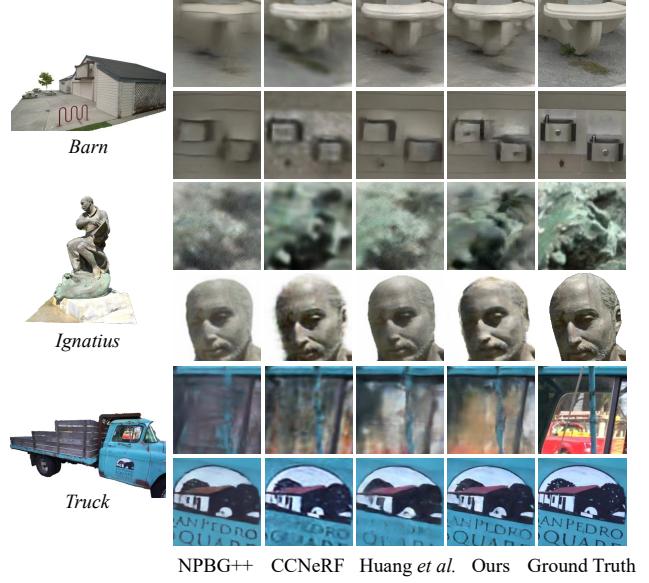


Figure 6. Qualitative evaluations on Tanks and Temples [16] dataset. Our method renders sharper detailed textures.

Implementation Details. There are two ways of rasterization: PyTorch3D [34] for headless server and OpenGL [41] for real-time rendering. Since our training is performed on a headless server, our main results are based on PyTorch3D rasterization. The widths of AFNet and hypernetwork are both 256, and camera ray directions are concated to the third layer output feature of AFNet. Moreover, we modify the positional encoding of NeRF [21] to make it more suitable for modulation, which is described in detail in supplementary material. The architecture of our U-Net is consistent with NPBG [1], *i.e.*, gated blocks are used, consisting of gated convolution and instance normalization. The U-Net down-samples the feature map for four times and the dimensions of each layer are 16, 32, 64, 128, 256. We train the AFNet and U-Net with L2 loss and perceptual loss [64] on a single NVIDIA GeForce RTX 3090 GPU. For preprocessing of point cloud optimization, we use only L2 loss and a 5-layers MLP with 256 width, where ray directions are concated to the third layer output feature. We use Adam optimizer for training, with a batch size of 2. The ini-

Method	Size(MB)	FPS	Editing Ability		NeRF-Synthetic			Tanks&Temples		
			Object	Scene	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
NeRF [21]	5.0	0.023	✗	✗	31.01	0.947	0.081	25.78	0.864	0.198
NSVF [19]	16.0	0.815	✗	✓	31.75	0.953	0.047	28.40	0.900	0.153
Object-NeRF [56]	121.2	0.1	✓	✗	31.19	0.949	0.079	25.96	0.866	0.194
PlenOctrees [60]	1976.3	168	✗	✓	31.71	0.958	0.053	27.99	0.917	0.131
Point-NeRF [53]	20.0	0.125	✗	✗	33.00	0.978	0.055	29.61	0.954	0.115
Plenoxels [8]	778.1	15	✗	✓	31.71	0.958	0.049	27.43	0.906	0.142
TensoRF [3]	71.8	1.15	✗	✓	33.14	0.963	0.056	28.56	0.920	0.118
Instant-NGP [22]	63.3	60	✗	✗	33.18	0.963	0.050	28.78	0.925	0.113
CCNeRF-CP [48]	4.4	1.05	✗	✓	30.55	0.935	0.076	27.01	0.879	0.180
CCNeRF-HY-S [48]	68.9	1.05	✗	✓	31.22	0.947	0.074	27.53	0.901	0.177
NPBG [1]	44.2	33.4	✓	✓	28.10	0.923	0.077	25.97	0.889	0.137
NPBG++ [33]	28.6	35.4	✓	✓	28.12	0.928	0.076	26.04	0.892	0.130
Huang <i>et al.</i> [13]	18.5	39.1	✓	✓	28.96	0.932	0.061	26.35	0.893	0.130
Ours	11.8	39.3	✓	✓	31.24	0.950	0.049	27.79	0.902	0.125

Table 2. Comprehensive evaluation. We report the model size (including point cloud size for point-based methods), FPS on NeRF-Synthetic [21] dataset with 800×800 resolution, editing ability of object-level transform and scene composition, and performance on NeRF-Synthetic [21] and Tanks and temples [16] datasets. The results demonstrate the excellent comprehensiveness of our method.

tial learning rates of AFNet and U-Net are $5e-4$ and $1.5e-4$ respectively.

4.2. Results of View Synthesis

We present quantitative comparison results on three performance metrics (PSNR, SSIM and LPIPS) in Tab. 1. As can be seen, we outperform the previous point-based methods with a significant margin on all datasets. Our method outperforms NeRF [21] and is comparable to the high-profile version of CCNeRF [48]. Figs. 5 and 6 present a qualitative comparison on NeRF-Synthetic [21] and Tanks and Temples [16] dataset. As can be seen, our method produces sharper details without introducing artifacts, especially for highly textured regions.

Comprehensive Analysis. To further illustrate the superiority of our method, we expand the quantitative experiments with more comparison works in Tab. 2, and comprehensively compare them from four aspects: model size, rendering speed, editing ability, and performance. It can be seen that almost all the follow-up works of NeRF are difficult to take into account the four aspects at the same time. PlenOctrees [60] achieves the state-of-the-art FPS, but its application scenarios are limited due to its huge model size. Instant-NGP [22] has high performance and real-time rendering speed, but lacks editing abilities. Object-NeRF [56] and CCNeRF [48] can achieve object-level editing and scene composition respectively, but cannot achieve real-time rendering. Our method achieves high performance and flexible editing with a small model size while rendering in real-time.

Ablation Study. The frequencies predicted by our hypernetwork are highly correlated with the true frequencies reflected by local variance, as shown in Fig. 7. Note that there

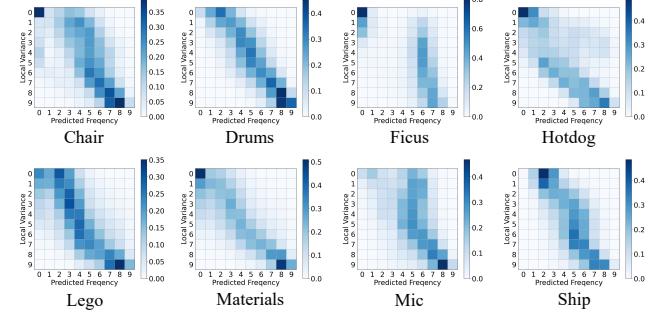


Figure 7. Confusion matrix between predicted frequency map and actual local variance on 200 test views of NeRF-Synthetic dataset. The frequency interval is divided into 10 segments, and from 0 to 9 are low frequency to high frequency. The frequencies predicted by our hypernetwork are highly correlated with the true frequencies reflected by local variance.

is a reasonable gap between our predicted 3D texture frequencies and 2D image frequencies. For example, in Ficus scene, the occlusion between leaves and background produces a high frequency region in 2D image, but the 3D texture of leaves is weak (just green). Moreover, we claim that the frequencies predicted by hypernetwork are adaptive, which can modulate the radiance signal better than actual local variance. To illustrate this, we inject the local variance of 2D images into AF layers without hypernetwork learning. The evaluation results in Tab. 3 show that the actual local variance is difficult to modulate the radiance signal. Also, point cloud geometry optimization is critical for some poorly reconstructed scenes. In Tab. 3, we show the quantitative results on PSNR and point scale. The pre-processing module improves performance while significantly reducing the size of the point cloud. In Fig. 8, we show the point clouds, depth maps and rendered images of Lego and Hotdog scenes. As can be seen, the noises near the crawler

	Chair	Drums	Ficus	Hotdog	Lego	Materi	Mic	Ship
	PSNR							
loc. var.	30.25	23.48	29.90	30.33	29.49	26.55	32.22	24.47
w/o. pre.	32.46	25.15	30.95	33.10	27.51	28.96	33.35	26.78
ours	33.06	25.95	32.19	35.82	31.56	29.69	33.64	27.97
	Number of points (Millions)							
w/o. pre.	0.69	0.54	0.40	1.02	1.07	1.19	0.50	1.52
ours	0.33	0.17	0.10	0.28	0.22	0.28	0.36	0.33

Table 3. Ablation analysis of adaptive frequency and point cloud geometry optimization. We inject the local variance extracted from the ground truth images as frequency information without using hypernetwork to predict frequency, denoted as *loc. var.*. We also show the PSNR of training without preprocessing of point cloud geometry optimization (*w/o. pre.*) and the number of points before and after preprocessing.

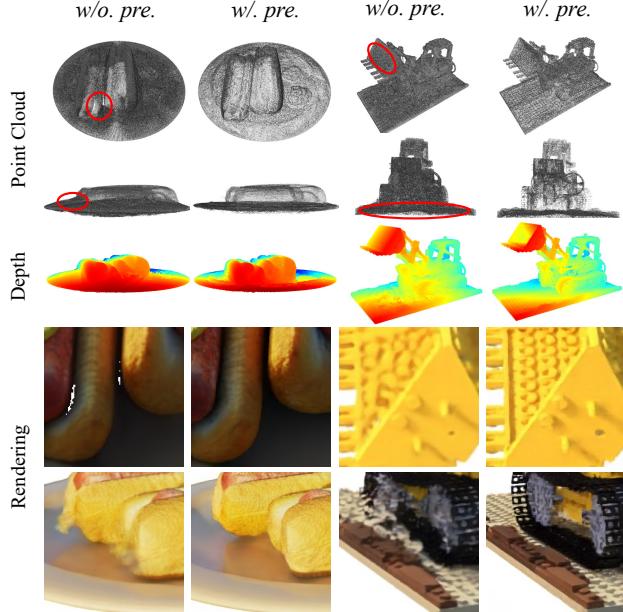


Figure 8. We show the point clouds, depth maps and rendered images with and without preprocessing of point cloud geometry optimization (*w/. pre.* and *w/o. pre.*). The noise near the crawler and the shovel in Lego scene causes artifacts and the missing part of Hotdog causes some blanks in the rendering result, while our preprocessing module optimizes the point cloud geometry and remove those artifacts.

and the shovel in Lego scene cause artifacts and the missing part of Hotdog causes some blanks in the rendering result, while our preprocessing module optimizes point cloud geometry and removes those artifacts.

4.3. Results of Editing

We also show the results of editing, including object editing and scene composition, which are shown in Figs. 9 and 10, respectively. As shown in Fig. 9, we edit objects in complex scenes, including rotation, translation and scaling.

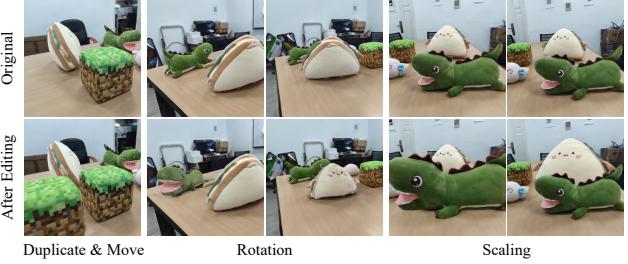


Figure 9. Object-level editing on ToyDesk [56] dataset. We can scale, duplicate, translate and rotate the objects selected by users.



Figure 10. Scene composition on NeRF-Synthetic [21] and Tanks and Temples [16] datasets. These scenes only need to be trained separately. And we can edit objects while combining the scenes.

In Fig. 10, our editing algorithm handles complex occlusion relationships. Moreover, object-level editing and scene composition are performed at the same time in the last two examples in Fig. 10, *i.e.*, the balls in Materials scene are moved and composed with other scenes. More editing results can be found in supplementary material.

5. Conclusion

We develop a novel point cloud rendering pipeline which enables high fidelity reconstruction, real-time rendering and user-friendly editing. For the novel view synthesis task, experiments on major benchmarks demonstrate the proposed method outperforms existing point cloud rendering methods and achieves the state-of-the-art.

Limitations and Future Work. Our current editing method cannot relight the edited objects (*i.e.*, the transformed objects retain their original appearance), handle non-rigid object and appearance editing. Moreover, the rendering speed of the edited scenes will be reduced. We will improve these shortcomings in the future.

Acknowledgment This work was supported by National Science Foundation of China (U20B2072, 61976137). This work was also partly supported by SJTU Medical Engineering Cross Research Grant YG2021ZD18.

References

- [1] Kara-Ali Aliiev, Artem Sevastopolsky, Maria Kolos, Dmitry Ulyanov, and Victor Lempitsky. Neural point-based graphics. In *European Conference on Computer Vision*, pages 696–712. Springer, 2020. [1](#), [2](#), [3](#), [6](#), [7](#), [14](#), [18](#), [19](#), [20](#)
- [2] Nuri Benbarka, Timon Höfer, Andreas Zell, et al. Seeing implicit neural representations as fourier series. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2041–2050, 2022. [2](#)
- [3] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. *arXiv preprint arXiv:2203.09517*, 2022. [2](#), [7](#)
- [4] 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*, pages 5939–5948, 2019. [2](#)
- [5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. [2](#), [6](#), [13](#)
- [6] Peng Dai, Yinda Zhang, Zhuwen Li, Shuaicheng Liu, and Bing Zeng. Neural point cloud rendering via multi-plane projection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7830–7839, 2020. [1](#)
- [7] Rizal Fathony, Anit Kumar Sahu, Devin Willmott, and J Zico Kolter. Multiplicative filter networks. In *International Conference on Learning Representations*, 2020. [2](#), [3](#), [4](#)
- [8] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. [2](#), [7](#)
- [9] Kyle Genova, Forrester Cole, Daniel Vlasic, Aaron Sarna, William T Freeman, and Thomas Funkhouser. Learning shape templates with structured implicit functions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7154–7164, 2019. [2](#)
- [10] Michelle Guo, Alireza Fathi, Jiajun Wu, and Thomas Funkhouser. Object-centric neural scene rendering. *arXiv preprint arXiv:2012.08503*, 2020. [3](#)
- [11] David Ha, Andrew Dai, and Quoc V Le. Hypernetworks. *arXiv preprint arXiv:1609.09106*, 2016. [2](#), [4](#)
- [12] Philipp Henzler, Niloy J Mitra, and Tobias Ritschel. Learning a neural 3d texture space from 2d exemplars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8356–8364, 2020. [2](#)
- [13] Xiaoyang Huang, Yi Zhang, Bingbing Ni, Teng Li, Kai Chen, and Wenjun Zhang. Boosting point clouds rendering via radiance mapping. *arXiv preprint arXiv:2210.15107*, 2022. [1](#), [2](#), [3](#), [6](#), [7](#), [14](#), [19](#), [20](#)
- [14] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. [2](#), [6](#), [14](#)
- [15] Kacper Kania, Kwang Moo Yi, Marek Kowalski, Tomasz Trzcinski, and Andrea Tagliasacchi. Conerf: Controllable neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18623–18632, 2022. [1](#)
- [16] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. [2](#), [6](#), [7](#), [8](#), [13](#)
- [17] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *arXiv preprint arXiv:2205.15585*, 2022. [1](#)
- [18] Zhuopeng Li, Lu Li, Zeyu Ma, Ping Zhang, Junbo Chen, and Jianke Zhu. Read: Large-scale neural scene rendering for autonomous driving. *arXiv preprint arXiv:2205.05509*, 2022. [1](#)
- [19] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. [2](#), [7](#), [13](#)
- [20] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5773–5783, 2021. [1](#)
- [21] 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. [1](#), [2](#), [3](#), [6](#), [7](#), [8](#), [12](#), [13](#), [15](#), [18](#), [19](#), [20](#)
- [22] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. [2](#), [7](#)
- [23] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015. [2](#)
- [24] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3504–3515, 2020. [2](#)
- [25] Michael Oechsle, Lars Mescheder, Michael Niemeyer, Thilo Strauss, and Andreas Geiger. Texture fields: Learning texture representations in function space. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4531–4540, 2019. [2](#)
- [26] 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*, pages 5589–5599, 2021. [2](#)
- [27] 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*, pages 165–174, 2019. 2
- [28] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2, 5
- [29] 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. *arXiv preprint arXiv:2106.13228*, 2021. 2
- [30] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Xiaowei Zhou, and Hujun Bao. Animatable neural radiance fields for modeling dynamic human bodies. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14314–14323, 2021. 2
- [31] 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. 2, 5
- [32] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International Conference on Machine Learning*, pages 5301–5310. PMLR, 2019. 2, 3
- [33] Ruslan Rakhimov, Andrei-Timotei Ardelean, Victor Lemitsky, and Evgeny Burnaev. Npbg++: Accelerating neural point-based graphics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15969–15979, 2022. 1, 2, 3, 6, 7, 13, 14, 18, 19, 20
- [34] Nikhila Ravi, Jeremy Reizenstein, David Novotny, Taylor Gordon, Wan-Yen Lo, Justin Johnson, and Georgia Gkioxari. Accelerating 3d deep learning with pytorch3d. *arXiv:2007.08501*, 2020. 6
- [35] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 1, 3
- [36] Darius Rückert, Linus Franke, and Marc Stamminger. Adop: Approximate differentiable one-pixel point rendering. *ACM Transactions on Graphics (TOG)*, 41(4):1–14, 2022. 1
- [37] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9–13 2011. IEEE. 12
- [38] Kersten Schuster, Philip Trettner, Patric Schmitz, Julian Schakib, and Leif Kobbelt. Compression and rendering of textured point clouds via sparse coding. In *High-Performance Graphics-Symposium Papers, Nikolaus Binder and Tobias Ritschel (Eds.). The Eurographics Association. https://doi.org/10.2312/hpg*, volume 20211284, 2021. 3
- [39] Markus Schütz, Bernhard Kerbl, and Michael Wimmer. Rendering point clouds with compute shaders and vertex order optimization. In *Computer Graphics Forum*, volume 40, pages 115–126. Wiley Online Library, 2021. 3
- [40] Yue Shi, Dingyi Rong, Bingbing Ni, Chang Chen, and Wenjun Zhang. Garf: Geometry-aware generalized neural radiance field. *arXiv preprint arXiv:2212.02280*, 2022. 2
- [41] Dave Shreiner, Bill The Khronos OpenGL ARB Working Group, et al. *OpenGL programming guide: the official guide to learning OpenGL, versions 3.0 and 3.1*. Pearson Education, 2009. 6
- [42] Vincent Sitzmann, Julien Martel, Alexander Bergman, David Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, 33:7462–7473, 2020. 2, 3
- [43] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, 32, 2019. 2
- [44] Kenneth O Stanley. Compositional pattern producing networks: A novel abstraction of development. *Genetic programming and evolvable machines*, 8(2):131–162, 2007. 2
- [45] Jingxiang Sun, Xuan Wang, Yong Zhang, Xiaoyu Li, Qi Zhang, Yebin Liu, and Jue Wang. Fenerf: Face editing in neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7672–7682, 2022. 3
- [46] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P Srinivasan, Jonathan T Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2846–2855, 2021. 2
- [47] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in Neural Information Processing Systems*, 33:7537–7547, 2020. 2, 3
- [48] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Compressible-composable nerf via rank-residual decomposition. *arXiv preprint arXiv:2205.14870*, 2022. 1, 2, 3, 6, 7, 15, 18, 19, 20
- [49] Edgar Treitschke, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 2
- [50] 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. 1
- [51] 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. 2
- [52] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibr-

- net: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 2
- [53] 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. 2, 3, 5, 7
- [54] Tianhan Xu and Tatsuya Harada. Deforming radiance fields with cages. In *European Conference on Computer Vision*, pages 159–175. Springer, 2022. 3
- [55] Bangbang Yang, Chong Bao, Junyi Zeng, Hujun Bao, Yinda Zhang, Zhaopeng Cui, and Guofeng Zhang. Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *European Conference on Computer Vision*, pages 597–614. Springer, 2022. 3
- [56] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13779–13788, 2021. 1, 2, 3, 7, 8, 14, 17
- [57] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 2, 5, 13
- [58] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33, 2020. 14
- [59] Wang Yifan, Felice Serena, Shihao Wu, Cengiz Öztireli, and Olga Sorkine-Hornung. Differentiable surface splatting for point-based geometry processing. *ACM Transactions on Graphics (TOG)*, 38(6):1–14, 2019. 1
- [60] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 1, 2, 7
- [61] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuwen 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. 3
- [62] Gizem Yüce, Guillermo Ortiz-Jiménez, Beril Besbinar, and Pascal Frossard. A structured dictionary perspective on implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19228–19238, 2022. 2
- [63] Jingyang Zhang, Yao Yao, and Long Quan. Learning signed distance field for multi-view surface reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6525–6534, 2021. 2
- [64] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 5, 6
- [65] Matthias Zwicker, Hanspeter Pfister, Jeroen Van Baar, and Markus Gross. Surface splatting. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 371–378, 2001. 3, 5

A. Implementation Details

A.1. End-to-end Rendering Pipeline

Rasterization Radius. During rasterization, each point is expanded to a disk to compensate for point cloud sparsity. We propose the following two heuristic rules to select disk radius: 1) the disk should cover the gaps between the points, otherwise the points from the occluded surfaces and the background can be seen through the front surface (so-called bleeding problem); 2) the radius should be as small as possible, otherwise it will lead to inaccurate depth estimation and the edges of objects will expand. In practice, the selection of the radius depends on the density of point cloud, and we can adjust the radius by observing the depth map obtained by rasterization. We also conduct an ablation study on radius, and it can be seen from Tab. 4 that the effect of the radius is not significant.

radius	3e-3	4e-3	5e-3	6e-3	7e-3
Hotdog	34.62	35.25	35.82	35.37	34.95
Mic	33.64	33.49	33.40	33.14	32.73

Table 4. PSNR of Hotdog and Mic scenes under different rasterization radius. The effect of radius is not significant.

Radiance Mapping and Refinement. We follow the position encoding form of NeRF [21], but we narrow the encoding interval to provide more fine-grained basis function support for frequency modulation, shown as follows:

$$\begin{aligned} \gamma(\mathbf{p}) = & (\sin(2^0\pi\mathbf{p}), \cos(2^0\pi\mathbf{p}), \\ & \sin(2^{0.5}\pi\mathbf{p}), \cos(2^{0.5}\pi\mathbf{p}), \\ & \dots, \\ & \sin(2^{L-1}\pi\mathbf{p}), \cos(2^{L-1}\pi\mathbf{p})). \end{aligned} \quad (6)$$

We set L as 10 and 2 for spatial coordinates and view directions, respectively. Our AFNet has five layers, the input feature dimension is 120, and hidden dimensions are 256, 256, 256 and 128. The output dimension is 11 (including 3-channel raw image and 8-channel feature map). Gated convolution block used in U-Net is shown in Fig. 11. For upsampling and downsampling in U-Net, we use nearest neighbor interpolation and average pooling respectively.

Training. During training stage, we conduct data augmentation by random scaling and cropping, and we found that random scaling is critical for the training of U-Net. The loss function is shown as follows:

$$\mathcal{L} = \mathcal{L}_{\text{RGB}} + \lambda \mathcal{L}_{\text{perceptual}}, \quad (7)$$

and we set λ as 5e-3 in our experiments. Thanks to the geometric prior, our model requires only ten minutes of training to achieve realistic rendering, but takes about 10 hours to almost converge, as shown in Fig. 12. In order to achieve

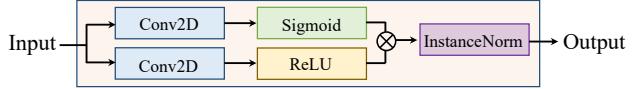


Figure 11. Gated convolution block.

complete convergence, more than ten hours of fine-tuning is required.

A.2. Point Cloud Geometry Optimization

Denoising We set $k = 16$, *i.e.*, keep 16 points in each pixel buffer. Each point \mathbf{p}_i in point cloud is assigned an opacity parameter α_i , and the color \mathbf{c}_i is predicted by MLP. The ray color is obtained as the following discrete volume rendering equation:

$$\hat{C}(\mathbf{r}) = \sum_{i=0}^{k-1} T_i \alpha_i \mathbf{c}_i, \quad \text{where } T_i = \prod_{j=0}^{i-1} (1 - \alpha_j). \quad (8)$$

We optimize the L2 distance of predicted image and ground truth with sparse regularization as follows:

$$\begin{aligned} \mathcal{L} &= \mathcal{L}_{\text{RGB}} + \lambda_{\text{sparse}} \mathcal{L}_{\text{sparse}}, \\ \mathcal{L}_{\text{RGB}} &= \|\hat{C} - C_{gt}\|_2^2, \\ \mathcal{L}_{\text{sparse}} &= \frac{1}{N} \sum_{i=1}^N [\log(\alpha_i) + \log(1 - \alpha_i)]. \end{aligned} \quad (9)$$

We set $\lambda_{\text{sparse}} = 5e-4$ in our experiments, and the learning rate of MLP and opacity parameters α are set as 5e-4 and 0.01, respectively. And for scenes without background, we also add transparency loss, *i.e.* L2 distance of predicted transparency and ground truth transparency. When the training converges, we remove those low-opacity points, as shown in Fig. 13.

In experiments, we perform 4-20 loops (depends on the raw point cloud quality) of point cloud denoising and completion step alternately. For every denoising step, we train for 300 Epochs. Fig. 14 shows the depth map computed by volume rendering equation (replace colors as depth values of sample points) during optimization. Fig. 15 shows the point cloud after completion. Although the completion step introduces additional noise, it would be removed in the denoising stage of the next iteration.

A.3. Editing Details

In practice, the part of point cloud that needs to be edited (*i.e.*, the object) is stored in the form of a mask. We use PCL library [37] to implement a simple interactive selection function, as shown in Fig. 17. In fact, users can obtain this mask by selecting points using any interactive software.

For scene composition, since the points on the edge of the object will expand, the mask of the scene will also expand outward. We design a simple strategy to avoid artifacts at edges by shrinking the masks determined by the index buffers, please refer to the code for specific implementation.

Training Time	0s	26s	117s	5min	15min	30min	1h	2h	10h
View #0									
View #25									
PSNR	6.77	18.95	24.02	26.59	29.34	30.57	32.68	33.74	35.05
View #0									
View #24									
PSNR	8.69	17.18	24.16	26.56	28.94	30.24	31.16	32.23	33.21

Figure 12. Rendering results during training. One Epoch consumes about 12.8s on NeRF-Synthetic. Our model requires only ten minutes of training to achieve realistic rendering, but takes about 10 hours to almost converge. In order to achieve complete convergence, more than ten hours of fine-tuning is required.

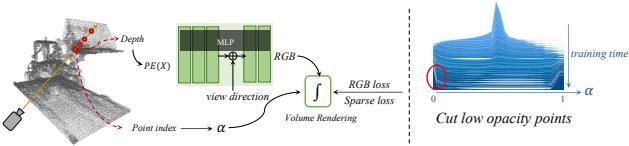


Figure 13. Denoising step of preprocessing.

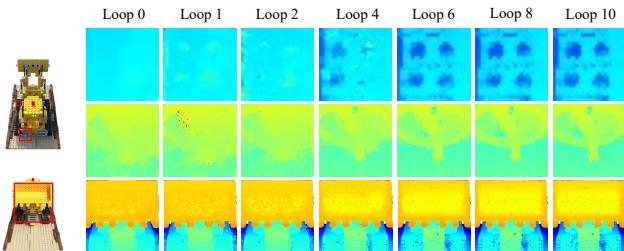


Figure 14. Depth map during geometry optimization. We remove outliers near the track and shovel.

B. Experimental Details

B.1. Datasets

- NeRF-Synthetic [21] is a high quality synthetic dataset containing pathtraced images of 8 objects. For each object, there are 100 frames for training and 200 frames for testing. The initial point clouds we used are generated by MVSNet [57]. We set the training size



Figure 15. Original and completed point cloud. When an empty buffer is detected, we add a set of points with opacity parameters along the pixel ray in the point clouds, shown as red points. Although the completion step introduces additional noise, it would be removed in the denoising stage of the next iteration.

as 800×800 and the scaling factor is $[0.5, 1.5]$ of the side length. At test time, we render with the original resolution 800×800 .

- We use a subset of Tanks and Temples [16] dataset, which is from NSVF [19], containing five scenes of real objects. We also use the foreground masks provided by NSVF. Each scene contains 152-384 images of size 1920×1080 . Due to the size limitation of U-Net, we resize the test resolution to 1920×1056 , and training size is 640×640 . The initial point clouds we used are provided by MVSNet [57].
- ScanNet [5] is a RGBD dataset of indoor scenes. We evaluate on *scene0000_00*, *scene0043_00*, and *scene0045_00*, as done in NPBG++ [33] and follow



Figure 16. Some rendering results on test views of Tanks and Temples dataset.

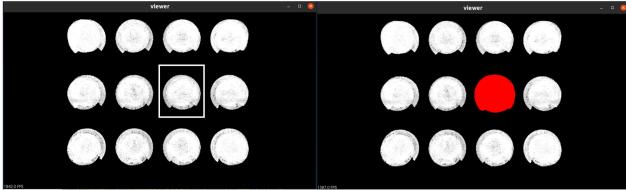


Figure 17. Point cloud selection interface.

their training-test split. Specifically, if there are more than 2000 frames in the scene, we select every 20-th image such that training views would not be too sparse. Otherwise, we take 100 frames with an equal interval in the image stream. Then, we select 10 frames at a fixed interval for testing and the rest for training. We set the training size as 720×720 and the test resolution is 960×1200 . The initial point clouds are obtained by the provided depth maps.

- DTU [14] is a multi-view stereo dataset with a resolution of 1200×1600 . We evaluate on *scan110*,

scan114 and *scan118*, and use the same point clouds and training-test split as NPBG++ [33]. We mask out the background in training using the binary segmentation masks provided by IDR [58]. The training patch size is set as 800×800 and the test resolution 1200×1600 .

- For ToyDesk [56] dataset, we only perform training and editing without evaluation.

B.2. Compared Methods

- NPBG [1]: A famous point-based rendering method, which uses point-wise features to encode the appearance of each surface point and an U-Net for decoding.
- NPBG++ [33]: The improved version of NPBG, which predicts the descriptors with a feature extractor and makes the neural descriptors view-dependent.
- Huang *et al.* [13]: The state-of-the-art point-based neural rendering, which combines explicit point clouds and implicit radiance mapping. However, its performance is still lower than that of NeRF, due to the weak

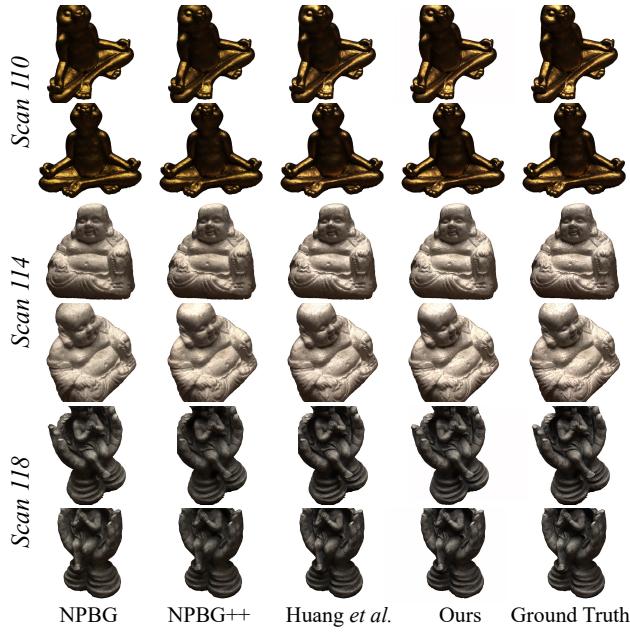


Figure 18. Qualitative comparison on DTU dataset.

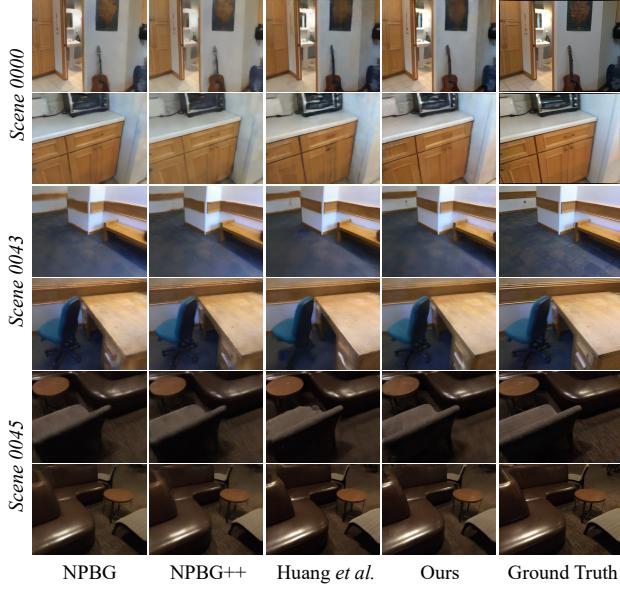


Figure 19. Qualitative comparison on ScanNet dataset.

frequency expressiveness and lack of geometric optimization.

- CCNeRF [48]: The latest editable variant of NeRF [21], which represents a 3D scene as a compressible 3D Tensor. Due to the massive sampling and calculation of tensor decomposition, the rendering speed is only 1.05 FPS.

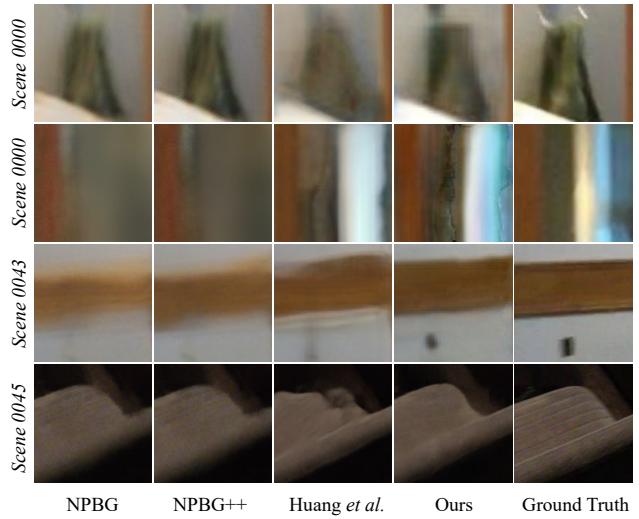


Figure 20. Qualitative comparison in detail on ScanNet dataset. Our rendering results are closer to ground truth.

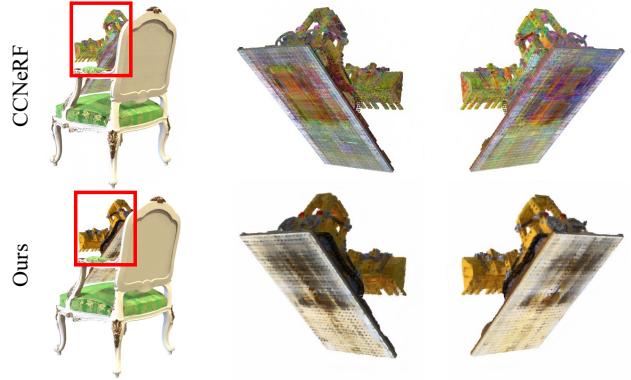


Figure 21. Extreme views of Lego scene that may be encountered in editing. We render reasonable results while CCNeRF renders distorted colors.

B.3. Detailed Results

We present some rendering results on Tanks and Temples dataset in Fig. 16, and qualitative comparisons on DTU and ScanNet datasets in Figs. 18 and 19, respectively. A comparison of details on ScanNet dataset is shown in Fig. 20. Although the rendering results look similar overall as seen from Figs. 18 and 19, we are better at some details, as shown in Fig. 20. Due to the blurring of some training images in ScanNet dataset, some rendering results of novel views are also blurred. For the compared methods, we all use the same experimental configuration, which is fair. We present quantitative evaluation for each scene of each dataset in Tabs. 5, 6 and 7. More editing results are shown in Figs. 21, 22 and 23.



Figure 22. Editing results, including object-level editing and scene composition.

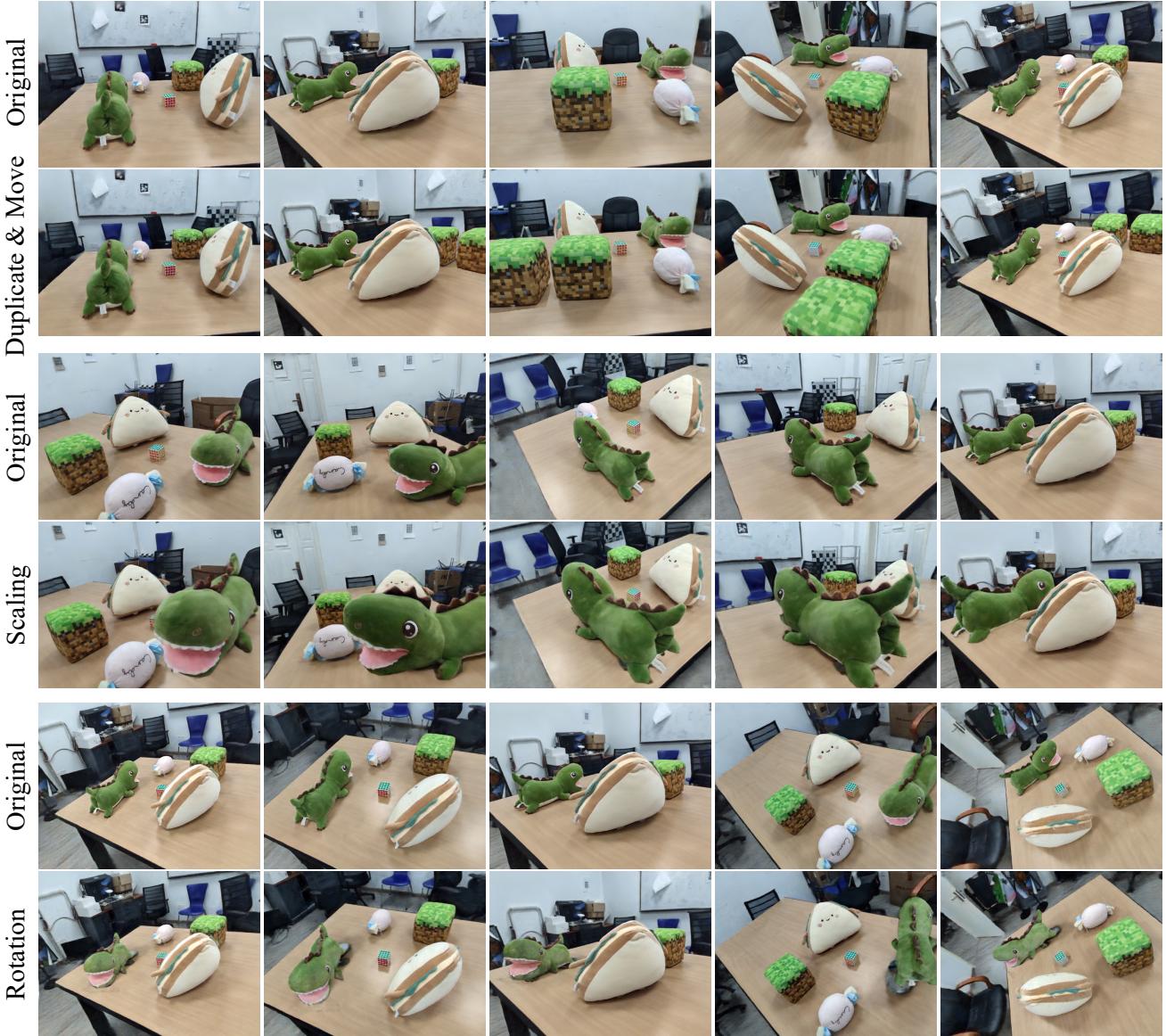


Figure 23. Object translation, rotation and scaling on ToyDesk [56] dataset. As can be seen from the last line of results, the rotation of the green object will expose the untrained local space, resulting in artifacts, which are also reflected in the results of Object-NeRF [56].

Method	NeRF-Synthetic									
	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship	Mean	
PSNR↑										
NeRF [21]	33.00	25.01	30.13	36.18	32.54	29.62	32.91	28.65	31.01	
CCNeRF-CP [48]	33.63	24.23	29.40	35.27	32.94	28.34	32.81	27.77	30.55	
CCNeRF-HY-S [48]	34.37	24.76	30.04	36.04	33.66	28.96	33.53	28.38	31.22	
NPBG [1]	28.81	23.57	28.23	32.03	27.72	27.24	31.16	26.04	28.10	
NPBG++ [33]	28.72	23.60	28.11	32.22	27.84	27.12	31.23	26.11	28.12	
Huang <i>et al.</i>	31.13	24.51	29.09	33.20	26.62	28.03	32.94	26.14	28.96	
Ours	33.06	25.95	32.19	35.82	31.56	29.69	33.64	27.97	31.24	
SSIM↑										
NeRF [21]	0.967	0.925	0.964	0.974	0.961	0.949	0.980	0.856	0.947	
CCNeRF-CP [48]	0.964	0.906	0.950	0.966	0.957	0.923	0.971	0.842	0.935	
CCNeRF-HY-S [48]	0.976	0.918	0.962	0.978	0.969	0.935	0.983	0.853	0.947	
NPBG [1]	0.954	0.902	0.942	0.960	0.919	0.922	0.970	0.812	0.923	
NPBG++ [33]	0.961	0.910	0.947	0.961	0.923	0.925	0.972	0.822	0.928	
Huang <i>et al.</i>	0.953	0.924	0.958	0.964	0.902	0.945	0.983	0.824	0.932	
Ours	0.974	0.938	0.971	0.974	0.956	0.955	0.986	0.845	0.950	
LPIPS↓										
NeRF [21]	0.046	0.091	0.044	0.121	0.050	0.063	0.028	0.206	0.081	
CCNeRF-CP [48]	0.037	0.111	0.055	0.057	0.037	0.082	0.031	0.196	0.076	
CCNeRF-HY-S [48]	0.036	0.109	0.054	0.056	0.036	0.080	0.030	0.192	0.074	
NPBG [1]	0.047	0.093	0.046	0.055	0.089	0.076	0.037	0.171	0.077	
NPBG++ [33]	0.048	0.092	0.043	0.053	0.089	0.074	0.030	0.178	0.076	
Huang <i>et al.</i>	0.040	0.068	0.035	0.038	0.085	0.050	0.014	0.159	0.061	
Ours	0.025	0.065	0.026	0.028	0.045	0.046	0.015	0.142	0.049	

Table 5. PSNR↑, SSIM↑ and LPIPS↓ on each scene of NeRF-Synthetic dataset.

Method	Tanks and Temples					Mean
	Barn	Caterpillar	Family	Ignatius	Truck	
PSNR↑						
NeRF [21]	24.05	23.75	30.29	25.43	25.36	25.78
CCNeRF-CP [48]	25.84	24.02	32.13	27.24	25.84	27.01
CCNeRF-HY-S [48]	26.34	24.48	32.75	27.76	26.34	27.53
NPBG [1]	24.86	22.05	30.84	26.50	25.59	25.97
NPBG++ [33]	24.90	22.22	30.67	26.98	25.45	26.04
Huang <i>et al.</i> [13]	25.34	23.09	30.65	27.01	25.68	26.35
Ours	27.01	24.67	32.36	28.83	26.56	27.79
SSIM↑						
NeRF [21]	0.750	0.860	0.932	0.920	0.860	0.864
CCNeRF-CP [48]	0.807	0.864	0.934	0.916	0.872	0.879
CCNeRF-HY-S [48]	0.827	0.886	0.957	0.939	0.894	0.901
NPBG [1]	0.841	0.848	0.940	0.928	0.887	0.889
NPBG++ [33]	0.842	0.863	0.943	0.933	0.878	0.892
Huang <i>et al.</i> [13]	0.841	0.855	0.946	0.936	0.886	0.893
Ours	0.847	0.876	0.953	0.941	0.895	0.902
LPIPS↓						
NeRF [21]	0.395	0.196	0.098	0.111	0.192	0.198
CCNeRF-CP [48]	0.310	0.223	0.078	0.099	0.192	0.180
CCNeRF-HY-S [48]	0.304	0.219	0.076	0.097	0.188	0.177
NPBG [1]	0.219	0.182	0.072	0.076	0.138	0.137
NPBG++ [33]	0.197	0.181	0.075	0.068	0.131	0.130
Huang <i>et al.</i> [13]	0.195	0.179	0.069	0.077	0.131	0.130
Ours	0.179	0.175	0.066	0.073	0.131	0.125

Table 6. PSNR↑, SSIM↑ and LPIPS↓ on each scene of Tanks and Temples dataset.

Method	ScanNet				DTU			
	Scene0000	Scene0043	Scene0045	Mean	Scan110	Scan114	Scan118	Mean
PSNR↑								
NeRF [21]	22.08	25.98	29.15	25.74	25.55	27.42	27.78	26.92
CCNeRF-CP [48]	21.14	25.89	26.91	24.65	25.88	27.89	26.61	26.79
CCNeRF-HY-S [48]	21.38	26.22	27.91	25.17	26.23	28.12	27.34	27.23
NPBG [1]	22.24	25.27	27.75	25.09	24.65	26.74	26.62	26.00
NPBG++ [33]	22.05	25.51	28.26	25.27	24.84	26.72	26.67	26.08
Huang <i>et al.</i> [13]	23.79	25.26	28.59	25.88	25.05	26.87	26.75	26.22
Ours	24.35	26.15	29.48	26.66	26.45	28.31	27.06	27.27
SSIM↑								
NeRF [21]	0.729	0.869	0.743	0.780	0.909	0.894	0.924	0.909
CCNeRF-CP [48]	0.695	0.849	0.779	0.774	0.904	0.894	0.922	0.907
CCNeRF-HY-S [48]	0.701	0.854	0.788	0.781	0.909	0.896	0.925	0.910
NPBG [1]	0.695	0.830	0.686	0.737	0.907	0.873	0.904	0.895
NPBG++ [33]	0.742	0.859	0.716	0.772	0.910	0.865	0.909	0.895
Huang <i>et al.</i> [13]	0.748	0.844	0.789	0.794	0.917	0.863	0.919	0.900
Ours	0.754	0.863	0.792	0.803	0.915	0.892	0.916	0.908
LPIPS↓								
NeRF [21]	0.588	0.466	0.558	0.537	0.194	0.217	0.182	0.198
CCNeRF-CP [48]	0.599	0.457	0.569	0.542	0.181	0.191	0.161	0.178
CCNeRF-HY-S [48]	0.594	0.456	0.566	0.539	0.171	0.188	0.154	0.171
NPBG [1]	0.474	0.421	0.482	0.459	0.124	0.143	0.123	0.130
NPBG++ [33]	0.457	0.410	0.477	0.448	0.125	0.148	0.121	0.131
Huang <i>et al.</i> [13]	0.440	0.389	0.415	0.415	0.124	0.154	0.117	0.132
Ours	0.418	0.369	0.412	0.400	0.122	0.143	0.122	0.129

Table 7. PSNR↑, SSIM↑ and LPIPS↓ on each scene of ScanNet and DTU datasets.