Spatial Annealing for Efficient Few-shot Neural Rendering

Yuru Xiao, Deming Zhai*, Wenbo Zhao, Kui Jiang, Junjun Jiang, Xianming Liu

Harbin Institute of Technology
hitxyr@stu.hit.edu.cn, (zhaideming, wbzhao, jiangkui, jiangjunjun, csxm)@hit.edu.cn

Abstract

Neural Radiance Fields (NeRF) with hybrid representations have shown impressive capabilities for novel view synthesis, delivering high efficiency. Nonetheless, their performance significantly drops with sparse input views. Various regularization strategies have been devised to address these challenges. However, these strategies either require additional rendering costs or involve complex pipeline designs, leading to a loss of training efficiency. Although FreeNeRF has introduced an efficient frequency annealing strategy, its operation on frequency positional encoding is incompatible with the efficient hybrid representations. In this paper, we introduce an accurate and efficient few-shot neural rendering method named **S**patial **A**nnealing regularized **NeRF** (**SANeRF**), which adopts the pre-filtering design of a hybrid representation. We initially establish the analytical formulation of the frequency band limit for a hybrid architecture by deducing its filtering process. Based on this analysis, we propose a universal form of frequency annealing in the spatial domain, which can be implemented by modulating the sampling kernel to exponentially shrink from an initial one with a narrow grid tangent kernel spectrum. This methodology is crucial for stabilizing the early stages of the training phase and significantly contributes to enhancing the subsequent process of detail refinement. Our extensive experiments reveal that, by adding merely one line of code, SANeRF delivers superior rendering quality and much faster reconstruction speed compared to current few-shot neural rendering methods. Notably, SANeRF outperforms FreeNeRF on the Blender dataset, achieving 700× faster reconstruction speed.

Introduction

Neural Radiance Field (NeRF) (Mildenhall et al. 2021), represented by a multi-layer perception (MLP), is a powerful learning-based tool for detailed 3D reconstruction and high-fidelity novel view synthesis. However, NeRF requires dense input views or days of training time, limiting its application in real-world scenarios such as autonomous driving, robotics, and AR/VR. Although grid-based hybrid presentations (Müller et al. 2022; Hu et al. 2023) have been introduced to improve training efficiency by accelerating convergence and reducing training time from hours to minutes, they still suffer from abnormal convergence and weak geometry awareness when dealing with sparse input views, a problem known as few-shot neural rendering.

Current few-shot neural rendering methods prioritize mitigating overfitting and enhancing geometry reconstruction, yet often neglect reconstruction efficiency. For instance, RegNeRF (Niemeyer et al. 2022), DietNeRF (Jain, Tancik, and Abbeel 2021), and VGOS (Sun et al. 2023) utilize patch-based geometry regularization or semantic consistency regularization, which involves computationally intensive forward processing on randomly sampled rendered patches. Alternatively, some methods require substantial modifications to the base architecture (Zhu et al. 2024) or depend on 3D prior from pre-trained networks (Wang et al. 2023; Xiao et al. 2024), reducing pipeline efficiency. There is an urgent need for methods that integrate into existing efficient architectures to improve robustness against sparse views while maintaining computational efficiency.

FreeNeRF (Yang, Pavone, and Wang 2023) adopts a coarse-to-fine training approach to address the few-shot neural rendering problem, akin to implicit geometry regularization. While FreeNeRF significantly enhances rendering performance in few-shot scenarios with minimal code adjustments, its training process remains time-consuming and labor-intensive due to its MLP-based architecture. Moreover, its implementation using frequency positional encoding is incompatible with hybrid acceleration architectures. Exploring an efficient way to embed the frequency annealing concept into hybrid-represented neural fields holds the promise of enhancing both practicality and performance.

Introducing frequency annealing in hybrid representations requires explicitly modeling the frequency band-limit of hybrid represented fields. While current methods employ pre-filtering strategies (Hu et al. 2023; Liu et al. 2024) to mitigate high-frequency information for anti-aliasing, they often lack explicit analysis and formulation of the frequency bandwidth. Therefore, conducting frequency band limit analysis and modeling for grid-based neural fields is crucial for advancing the efficiency of few-shot neural rendering.

In this paper, we derive the explicit formation of frequency bandwidth for a pre-filtering-driven hybrid representation through an analysis of the filtering process. Building on this analysis, we introduce a more generalized form of frequency regularization in the spatial domain, termed spa-

*Corresponding authors

tial annealing, by examining the relationship between spatial area sampling and frequency bandwidth. Our method is applicable to both implicit and hybrid representations and can be implemented with just one line of code on two well-known pre-filtering-driven neural fields: MipNeRF and TriMipRF, characterized by implicit and hybrid representations, respectively. Extensive experiments on synthetic datasets and Blender (Mildenhall et al. 2021) in a few-shot context demonstrate the effectiveness and efficiency of our approach. It not only outperforms the original TriMipRF by approximately 3dB in PSNR but also surpasses the state-of-the-art few-shot neural rendering technique, FreeNeRF (Yang, Pavone, and Wang 2023), by 0.3dB in PSNR. Additionally, our method achieves a training speed that is 700× faster than FreeNeRF, marking a significant advancement in both performance and efficiency. The main contributions of our work are summarized as follows:

- We establish the analytical formulation of the frequency band limit for a pre-filtering designed hybrid representation architecture by analyzing its filtering process.
- We explore a general form of frequency annealing in the spatial domain based on frequency band limit analysis, which can be integrated into both implicit and hybrid representations with just one line of code.
- We present extensive experimental results demonstrating that our method achieves superior efficiency and enhanced performance. Specifically, it outperforms FreeNeRF by 0.3 dB in PSNR on the Blender dataset while achieving training speeds 700 times faster.

Related Work

Neural Radiance Fields. Neural Radiance Field (NeRF) (Mildenhall et al. 2021) is a distinguished 3D representation method renowned for its ability to render realistic novel views. It utilizes a Multilayer Perceptron (MLP) network to implicitly associate 3D coordinates with radiance attributes, including density and color. Over the past few years, the research community has developed numerous NeRF extensions, enhancing its application across a variety of domains such as novel view synthesis (Martin-Brualla et al. 2021; Barron et al. 2021, 2022; Mildenhall et al. 2022), surface reconstruction (Oechsle, Peng, and Geiger 2021; Wang et al. 2021), dynamic scene modeling (Fang et al. 2022; Liu et al. 2023), and 3D content generation (Gu et al. 2023; Chen et al. 2023; Hong et al. 2023). Although NeRF has pioneered several advancements across different research areas, it encounters significant challenges. The method particularly struggles with slow reconstruction speed and a heavy dependence on the density of input views, limiting its efficiency and practicality in broader applications. These challenges highlight the need for ongoing research and development to optimize NeRF's performance and expand its usability in the field of 3D representation and beyond.

Few-shot Neural Rendering. NeRF grapples with accurately fitting the low-frequency geometry when processing sparse input views. Many approaches incorporate 3D information like sparse point clouds (Deng et al. 2022), estimated depth (Wang et al. 2023; Cao, Rockwell, and Johnson 2022; Roessle et al. 2022), or dense correspondences (Truong et al. 2023; Lao et al. 2024) for enhanced supervision. However, integrating additional algorithms or models to acquire extra 3D data adds complexity and implementation challenges to the overall process. Beyond leveraging 3D information, some methods aim to directly regularize geometry using strategies such as random patch-based semantic consistency (Jain, Tancik, and Abbeel 2021), or geometric regularization (Kwak, Song, and Kim 2023; Niemeyer et al. 2022; Sun et al. 2023; Seo et al. 2023). These techniques, however, tend to prolong training times because of the added rendering burden on extra sampled patches. Learning-based methods (Lin et al. 2023; Yu et al. 2021; Chen et al. 2021; Liu et al. 2022; Chibane et al. 2021), on the other hand, seek to train a network on extensive multi-view datasets to internalize a geometric prior, yet these approaches require costly pre-training and additional fine-tuning steps.

Recently, FreeNeRF (Yang, Pavone, and Wang 2023) addresses the issue of under-constrained geometry in sparse view settings by progressively increasing the frequency of positional encoding. This technique eliminates the need for additional 3D information and does not increase the computational cost of the base architecture, positioning it as a robust baseline for few-shot neural rendering applications. While the coarse-to-fine training approach of FreeNeRF has demonstrated effectiveness in few-shot scenarios, it still necessitates lengthy training periods. Furthermore, its method of applying a mask to frequency positional encoding is not readily adaptable to various NeRF acceleration architectures. Motivated by FreeNeRF's achievements, our goal is to develop a straightforward solution for a NeRF acceleration architecture, aiming to enhance performance in few-shot settings without compromising training efficiency.

Preliminaries and Motivation

Frequency Regularization. FreeNeRF (Yang, Pavone, and Wang 2023) addresses the few-shot challenge by progressively increasing the frequency of positional encoding throughout the training process, a technique known as frequency regularization. This approach is straightforwardly implemented using a modulated mask as

$$\gamma'(t, T, \mathbf{x}) = \gamma(\mathbf{x}) \odot \mathbf{M}(t, T, L) \tag{1}$$

with

$$M_i(t, T, L) = \begin{cases} 1 & \text{if } i \leq \frac{t \cdot L}{T} + 3 \\ \frac{t \cdot L}{T} - \lfloor \frac{t \cdot L}{T} \rfloor & \text{if } \frac{t \cdot L}{T} + 3 < i \leq \frac{t \cdot L}{T} + 6 \\ 0 & \text{if } i > \frac{t \cdot L}{T} + 6 \end{cases}, \tag{2}$$

where t and T represent the current and total iteration steps, respectively, γ and γ' correspond to the initial and masked positional encodings, L controls the maximum frequency of positional encoding, and M_i is the i-th element of mask $\mathbf{M}$, which linearly expands throughout the training process.

TriMipRF. TriMipRF (Hu et al. 2023) introduces a pre-filtering strategy for tri-plane representation. Due to the incompatibility of integrated positional encoding (IPE) introduced by MipNeRF (Barron et al. 2021) with this hybrid

model, TriMipRF performs area-sampling by querying features on the three planes with plane levels correlating with the projected radius of the sampling sphere within the cone. The relevant levels are denoted by $\lfloor l \rfloor$ and $\lceil l \rceil$, where l signifies the query level corresponding to the sphere's radius:

$$l = \log_2 (\frac{\tau}{\ddot{r}}), \tag{3}$$

where $\ddot{r} = T_0/\sqrt{\pi}$ represents the base radius associated with base planes with maximum resolution r_0, τ denotes the radius of the sampling sphere, and T_0 is the period of the base planes. The plane level $[l] \in \mathbb{N}$ has a relationship with the plane resolution as $r_{[l]} = r_0/2^{[l]}$. For detailed formation, please refer to TriMipRF (Hu et al. 2023).

Method

Spatial Annealing in Implicit Representation

We start by examining the relationship between frequency regularization (see Eq. 1) and the pre-filtering strategy based on integrated positional encoding (IPE). We consider a multivariate Gaussian distribution in 3D space, depicted as

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

where $\boldsymbol{\mu}$ represents the position, and $\boldsymbol{\Sigma}$ denotes the covariance matrix. For straightforward comparison, we model the Gaussian as isotropic, characterized by a diagonal $\boldsymbol{\Sigma}$ with uniform elements σ_f^2, reflecting the sample space size. Following MipNeRF (Barron et al. 2021), we calculate the Gaussian's integrated positional encoding γ_G as

$$\gamma_G(\boldsymbol{\mu}) = \gamma'_G(\boldsymbol{\mu}) \odot \mathbf{M}_L \tag{5}$$

with

$$\begin{aligned} \mathbf{M}_L &= [e^{-\frac{1}{2}\sigma_f^2}, e^{-\frac{1}{2}\sigma_f^2}, ..., e^{-2^{2L-3}\sigma_f^2}, e^{-2^{2L-3}\sigma_f^2}] \\ \gamma'_G(\boldsymbol{\mu}) &= [\sin(\boldsymbol{\mu}), \cos(\boldsymbol{\mu}), ..., \sin(2^{L-1}\boldsymbol{\mu}), \cos(2^{L-1}\boldsymbol{\mu})], \end{aligned} \tag{6}$$

where $\gamma'_G(\boldsymbol{\mu})$ represents the frequency positional encoding of $\boldsymbol{\mu}$, and $\mathbf{M}_L$ denotes a low-pass mask applied to this encoding, with L controlling the maximum encoded frequency. The mask's structure aligns with a Gaussian distribution, featuring a covariance $\sigma_i^2 = 1/\sigma_f^2$, where σ_i^2 regulates the frequency bandwidth.

Upon comparing Eq. 6 with Eq. 1, we observe a notable similarity between the IPE and frequency regularization: both apply a low-pass mask to the original positional encoding, with the mask's form being the primary difference. Drawing inspiration from FreeNeRF (Yang, Pavone, and Wang 2023), which linearly expands the frequency mask (effectively exponentially increasing the frequency bandwidth), we adopt an exponential growth model for σ_i^2, defined as $\sigma_i^2 = f_s 2^x$, where x is the step increment corresponding to the iteration count and f_s controls the initial frequency bandwidth. Concurrently, the spatial Gaussian's covariance σ_f^2 decreases exponentially, represented as

$$\sigma_f^2 = \frac{1}{f_s 2^x}. \tag{7}$$

Consequently, we deduce that the frequency regularization introduced by FreeNeRF (Yang, Pavone, and Wang 2023) can be executed by inversely adjusting the spatial sample space within the pre-filtering strategy, as detailed in Eq. 7. This insight guides the exploration of the form of spatial annealing strategy in hybrid representation.

Spatial Annealing in Hybrid Representation

Filtering Process. Deducing the relationship between the spatial sampling strategy and frequency bandwidth in hybrid representation requires a detailed examination of the filtering process. Building on recent research into grid-based models (Zhao et al. 2024; Shabanov et al. 2024), we construct the mathematical form of the tri-plane represented field and deduce the filtering process of the base architecture, which subsequently motivates the establishment of the spatial annealing strategy in hybrid representation.

Based on recent research, the band limit of an optimized field can be determined by the sampling kernel under specific conditions (Shabanov et al. 2024). We analyze the sampling process of this optimized neural field directly. General grid-based models (Zhao et al. 2024; Shabanov et al. 2024) reconstruct the optimized continues field $\mathbf{f}(\mathbf{x})$ from discrete feature points by applying a convolution process with a kernel function κ:

$$\hat{\mathbf{f}}(\mathbf{x}) = \hat{\kappa}_\omega(\mathbf{x}) * \sum_n \mathbf{f}(\mathbf{x}) \cdot \delta(\mathbf{x} - \mathbf{x}_n), \tag{8}$$

where δ is the impulse function, ω is the frequency bandwidth of the kernel, and $x_n \in \mathbb{R}^3$ are discrete grid points. In current hybrid representation fields, the kernel function can be implemented using simple linear interpolation, equivalent to a conical kernel $\Lambda_\omega(\mathbf{x})$. For regular grids with period T, the frequency bandwidth of the kernel is approximated by $\omega = \frac{2\pi}{T}$, as the conical kernel is not an ideal low-pass filter.

In TriMipRF, the reconstructed signal $\hat{\mathbf{f}}^m(\mathbf{x}, [l])$ for each plane m at a specific level $[l]$ can be represented as

$$\hat{\mathbf{f}}^m(\mathbf{x}, [l]) = \Lambda^m_{\omega_{[l]}}(\mathbf{x}) * \sum_n \mathbf{f}^m(\mathbf{x}, [l]) \cdot \delta(\mathbf{x} - \mathbf{x}_n^{[l]}), \tag{9}$$

where $m \in \{xy, xz, yz\}$ represents the label of each plane, $\Lambda^m_{\omega_{[l]}}(\mathbf{x})$ represents a two-dimensional conical kernel, and $\{\mathbf{x}_n^{[l]}\}$ is the uniform lattice with level $[l]$. The bandwidth of each plane at a specific level $[l]$ can be represented as $\omega_{[l]} = \frac{2\pi}{2^{[l]} T_0}$, where $T_0 = \frac{B_{max} - B_{min}}{r_0}$ represents the period of the grid with maximum resolution r_0 and bounding box size $B_{max} - B_{min}$. We label the bandwidth ω corresponding to the level l as ω_l. TriMipRF utilizes multi-resolution triplanes to model varying levels of detail (LOD), it queries features from adjacent integral levels of l corresponding to the size of the sphere (refer to Eq. 3) as

$$\tilde{\mathbf{f}}^m(\mathbf{x}, l) = \Lambda_{\omega_c}(l) * \sum_{[l]} \hat{\mathbf{f}}^m(\mathbf{x}, l) \cdot \delta(l - [l]), \tag{10}$$

where $\Lambda_{\omega_c}(l)$ is a conical kernel with constant bandwidth ω_c, and $[\cdot]$ represents discrete integral values. TriMipRF utilizes a pre-filtering strategy across the level dimension to filter out fields with higher-frequency bandwidths that cannot

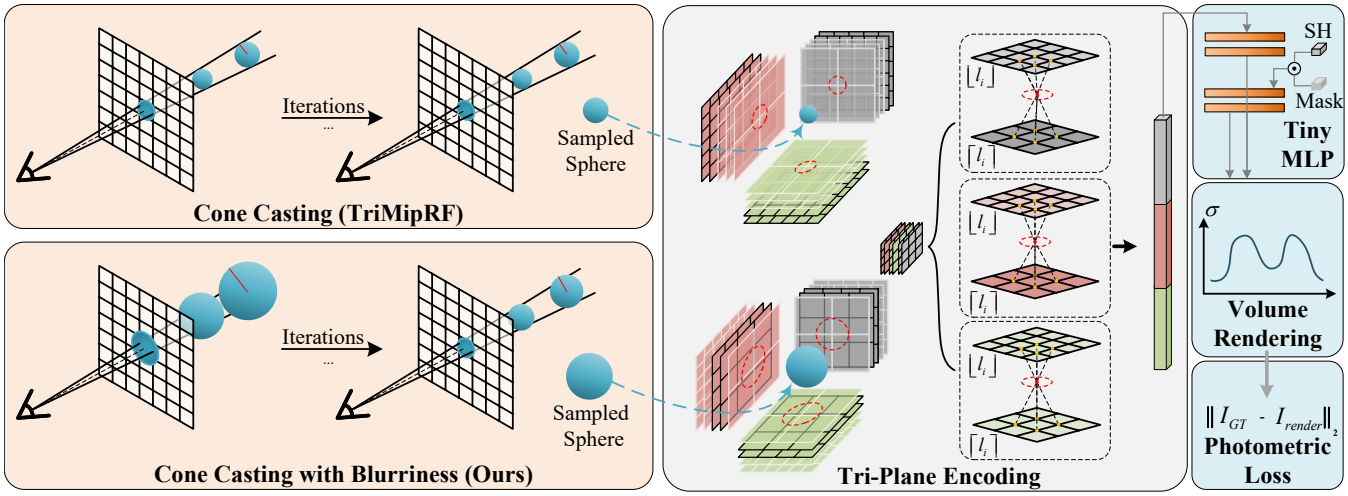


Figure 1: An overview of the complete framework. We introduce an efficient few-shot neural rendering method utilizing TriMipRF. Initially, we set the sample sphere's radius larger than that of the base sphere to optimize low-frequency geometry, as depicted in the bottom left corner. During training, we progressively reduce the sphere's radius through exponential decay, thereby refining local details within the reconstructed global structure.

be well-sampled by discrete rays, thereby addressing aliasing artifacts. This approach constructs a continuous field along the level dimension within a continuous range of sampling spaces. Equation 10 can also be rewritten in the form of linear interpolation across the level dimension:

$$\tilde{\mathbf{f}}^m(\mathbf{x}, l) = \frac{\lceil l \rceil - l}{\lceil l \rceil - \lfloor l \rfloor} \hat{\mathbf{f}}^m(\mathbf{x}, \lfloor l \rfloor) + \frac{l - \lfloor l \rfloor}{\lceil l \rceil - \lfloor l \rfloor} \hat{\mathbf{f}}^m(\mathbf{x}, \lceil l \rceil). \tag{11}$$

The continuous field represented by each feature plane can be considered a linear combination of the two fields at specific levels $\lceil l \rceil$ and $\lfloor l \rfloor$. Therefore, the bandwidth ω_m of the combined field lies within the range $(\omega_{\lceil l \rceil}, \omega_{\lfloor l \rfloor})$ as

$$\omega_{\lceil l \rceil} \leq \omega_m \leq \omega_{\lfloor l \rfloor}. \tag{12}$$

The final feature field can be represented as the concatenation of components corresponding to each plane. Since the sampling process operates independently based on the same kernel and identical querying level l, the bandwidth ω of the final field is determined by ω_m.

Spatial Annealing Strategy. We aim to explore the relationship between the frequency bandwidth ω_m and the size of spatial sampling space. Since the bandwidth of the field cannot be analytically represented, we approximate ω_m by

$$\omega_m \approx \frac{2\pi}{2^l T_0}. \tag{13}$$

The approximation consistently satisfies the inequality given by Eq. 12, since Eq. 13 is monotonic along level l. Such an approximation has an error bound of $\omega_{\lfloor l \rfloor} - \omega_{\lceil l \rceil}$. Inspired by FreeNeRF, we exponentially increase the frequency bandwidth from an initial value f_s as $\omega_m = f_s \cdot 2^x$, where x is the step increment corresponding to the iteration count. As TriMipRF aligns level l with the size of the sampling sphere according to Eq. 3, the relationship between the size of the

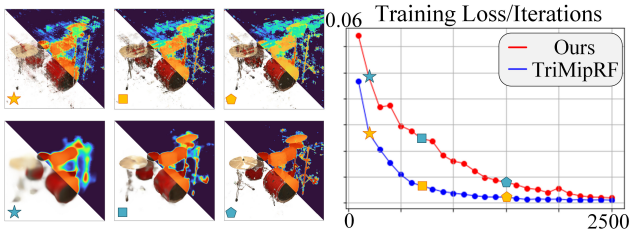


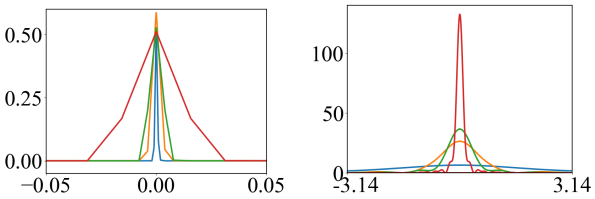
Figure 2: Comparison results during the training procedure. The training loss curve reveals that TriMipRF exhibits premature convergence early, resulting in the underfitting of the geometry depicted on the left. Conversely, our spatial annealing strategy effectively addresses this challenge.

sphere τ and the increment step x can be computed as

$$\tau = \frac{2\sqrt{\pi}}{f_s 2^x}. \tag{14}$$

We observe that Eq. 14 and Eq. 7 exhibit similar forms, both exponentially decreasing the size of the sampling space from an initial value. Consequently, we establish a universal form of frequency regularization in the spatial domain that is suitable for both implicit and hybrid representations.

Fig. 1 provides an overview of our method. We have developed a coarse-to-fine training strategy based on TriMipRF (Hu et al. 2023). Initially, we implement blurring in the rendering process by increasing the radius r of the sample sphere, as depicted in the bottom-left corner of Fig. 1. At the beginning of training, this approach aids in optimizing the reconstruction of global geometric structures, crucial for mitigating overfitting issues common in few-shot scenarios, as illustrated in Fig. 2. The radius of the sample area is systematically reduced following an exponential decay de-



(a) GTK spatial (b) GTK spectrum

Figure 3: Visualization of GTK and its spectrum. We vary the size r of the sampling sphere and measure the 1-D grid tangent kernel along with its frequency spectrum. The red, green, yellow, and blue curves correspond to sphere radii of $r = 0.07$, $r = 0.01$, $r = 0.007$, and $r = 0.001$, respectively.

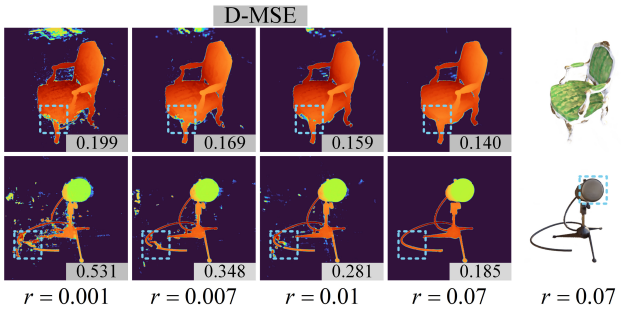


Figure 4: Qualitative rendered depth results with varying size of sampling sphere.

scribed in Eq. 14. This gradual reduction aims to progressively shift the training focus towards enhancing local geometric and textural details. The specific steps of this annealing process are outlined as

$$r_i = \begin{cases} \tau + \dfrac{f_s}{2^{\vartheta x}}, i < T \\ \tau, i \geq T \end{cases}, x = \left\lfloor \frac{iN_{split}}{T} \right\rfloor, \tag{15}$$

where N_{split} denotes the total number of decrement steps, i represents the current iteration count, and T indicates the stopping point of annealing. f_s determines the initial size of the sphere, while ϑ governs the rate of decrement.

Assessing the Geometry Awareness of Spatial Blurriness. In this paper, we strategically modulate the size of the sample sphere starting from an initially large value. This approach introduces blurriness corresponding to low-frequency band limit (refer to Eq. 13), which enhances geometry awareness during the initial training stage. To verify this insight, we conduct a spectrum analysis of the grid-based model utilizing grid tangent kernel (GTK) theory (Zhao et al. 2024). We initially simplified Eq. 11 as follows:

$$\begin{aligned} \tilde{\mathbf{f}}^m(\mathbf{x}, l) = & \frac{\lceil l \rceil - l}{\lceil l \rceil - \lfloor l \rfloor} \sum_n \varphi^{\lfloor l \rfloor}(\mathbf{x}_n^{\lfloor l \rfloor}) \boldsymbol{w}_n^{\lfloor l \rfloor} + \\ & \frac{l - \lfloor l \rfloor}{\lceil l \rceil - \lfloor l \rfloor} \sum_n \varphi^{\lceil l \rceil}(\mathbf{x}_n^{\lceil l \rceil}) \boldsymbol{w}_n^{\lceil l \rceil}, \end{aligned} \tag{16}$$

where $\boldsymbol{w}_n^{\lfloor l \rfloor}$ and $\boldsymbol{w}_n^{\lceil l \rceil}$ are features stored in grid points at levels $\lfloor l \rfloor$ and $\lceil l \rceil$, respectively, and φ represents the interpolation weights. As the GTK of a grid-based model remains stationary during training (Zhao et al. 2024), the GTK during training can be computed by the optimized neural field:

$$[\boldsymbol{G}_g(t)]_{i,j} = \left\langle \frac{\partial \tilde{\mathbf{f}}^m(\mathbf{x}_i, l)}{\partial \boldsymbol{w}}, \frac{\partial \tilde{\mathbf{f}}^m(\mathbf{x}_j, l)}{\partial \boldsymbol{w}} \right\rangle, \tag{17}$$

where $\boldsymbol{w}$ denotes features stored in grid points. We numerically analyze the 1D spatial GTK of the TriMip-represented field and its spectrum, as shown in Fig. 3. The results indicate that with a larger sampling sphere, the spectrum becomes narrower, which biases training more towards low-frequency information, especially low-frequency geometry (Zhao et al. 2024). Such spectral bias performs exceptionally well when handling extreme scenarios, such as sparse views that specifically lack multi-view geometry consistency. As the training iteration progresses, the size of the sample sphere exponentially decreases, corresponding to a wider GTK spectrum, which shifts the training focus towards high-frequency details. Although current research has reached this insight for implicit representation (Yang, Pavone, and Wang 2023), we expand this insight to grid-based hybrid representation utilizing recently introduced GTK analysis.

We conduct experiments to validate this insight. We vary the sampling sphere size r from a small value (0.001) to a relatively larger value (0.07), aligning with the setting of the GTK analysis, and evaluate the depth mean square error (D-MSE) between the rendered depth results and the ground truth (GT) depth maps, as shown in Fig. 4. We observe that the D-MSE decreases with increasing sampling sphere size. Additionally, qualitative results indicate that a larger sampling area with blurred rendered results corresponds to more complete geometry.

Experiments

Datasets and Metrics. We evaluate our method using the Blender dataset (Mildenhall et al. 2021), which comprises 8 synthetic scenes observed from a surround view perspective. Consistent with the FreeNeRF (Yang, Pavone, and Wang 2023) and DietNeRF (Jain, Tancik, and Abbeel 2021), we train our model on 8 views identified by the IDs 26, 86, 2, 55, 75, 93, 16, and 73. The evaluation is conducted on 25 images, evenly selected from the test set. All images are downsampled by a factor of 2 using an average filter. For quantitative analysis, we report the average scores across all test scenes for PSNR, SSIM, and LPIPS.

Degree Truncation of Spherical Harmonic Encoding. When the input views are sparse, fitting the view-dependent colors becomes challenging. To address this and cover a broader range of unseen view directions, we directly mask the higher levels of the Spherical Harmonic Encoding (Verbin et al. 2022), which can be represented as

$$M_i(n) = \begin{cases} 1, i \leq \sum_{j=0}^{n-1} 2j+1 \\ 0, i > \sum_{j=0}^{n-1} 2j+1 \end{cases}, \tag{18}$$

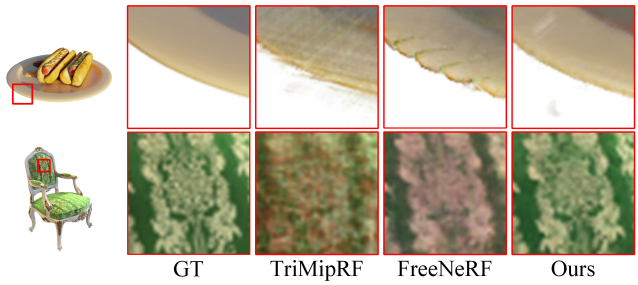


Figure 5: Qualitative results on Blender. We present qualitative comparisons of our method with the base architecture TriMipRF and the FreeNeRF few-shot baseline, utilizing 8 input views consistent with the FreeNeRF configuration.

where n represents the truncated level and M_i is the i-th element of the Spherical Harmonic mask $\mathbf{M}$. The mask zeroes out all Spherical Harmonic components above level n while retaining all lower-level components.

Implementation Details. We implement our methodology based on the TriMipRF codebase (Hu et al. 2023), utilizing the PyTorch framework (Paszke et al. 2019). Building on TriMipRF's original configurations, we introduce additional hyperparameters tailored to our spatial annealing strategy. Specifically, we initialize the sphere size f_s at 0.15, set the decrease rate ϑ to 0.2, define the total number of decrement steps N_{split} as 30, and establish the stop point T at $2K$. In the few-shot setting, there is a significant reduction in the number of input rays. Consequently, we limit the maximum training steps for both our method and the baseline TriMipRF to one-tenth of those employed in TriMipRF's full-view setting. We train our model using the AdamW optimizer (Loshchilov and Hutter 2017) for $2.5K$ iterations, with a learning rate of 2×10^{-3} and a weight decay of 1×10^{-5}. Regarding the degree truncation in Spherical Harmonic Encoding, we consistently set the truncated level n to 2 across all experiments.

Blender Dataset

Fig. 5 and Tab. 1 show the qualitative and quantitative comparisons on the Blender dataset (Mildenhall et al. 2021), respectively. The qualitative analysis reveals that the TriMipRF (Hu et al. 2023) produces noticeable distorted rendered results in the "chair" and "hotdog" scenes. In contrast, our approach, which requires only a few additional lines of code, effectively addresses these issues. When compared with the latest few-shot baseline, FreeNeRF (Yang, Pavone, and Wang 2023), our method shows superior performance, particularly in detail-rich areas highlighted in red boxes. Specific examples include the texture in the "chair" scene and the edge of the plate in the "hotdog" scene.

Our method demonstrates significant advancements in quantitative results, outperforming TriMipRF (Hu et al. 2023) by nearly 3 dB in PSNR with similar 35 seconds reconstruction time. Furthermore, the evaluated training durations demonstrate that our method achieves faster reconstruction speed. Notably, while the FreeNeRF (Yang, Pavone, and Wang 2023) necessitates 7.5 hours for training, and DietNeRF (Jain, Tancik, and Abbeel 2021), with its patch-based semantic regularizer, requires an extensive 26 hours, our approach achieves a remarkable 700× reconstruction speed compared to FreeNeRF. In addition to its efficiency, our method also delivers superior quantitative outcomes, exceeding FreeNeRF by 0.27 dB in PSNR. While these results are impressive, extending the training duration has the potential to further enhance performance. To this end, we increase the training iterations to $5K$, while keeping all other settings unchanged. As shown in Tab. 1, our method not only surpasses FreeNeRF with a 0.55 dB gain in PSNR but also delivers a substantial 350× acceleration.

Method	PSNR↑	SSIM↑	LPIPS↓	Train↓
NeRF	14.93	0.687	0.318	6 h
MipNeRF	18.74	0.830	0.240	9 h
Simplified NeRF	20.09	0.822	0.179	2 h
TriMipRF	21.63	0.818	0.183	**35 s**
DietNeRF	23.15	0.866	0.109	26 h
DietNeRF, $\mathcal{L}_{\text{MSE}}$ ft	23.59	0.874	0.097	32 h
InfoNeRF	22.01	0.852	0.133	7 h
MixNeRF	23.84	0.878	0.103	9 h
FreeNeRF	24.26	0.883	0.098	7.5 h
VGOS	21.32	0.861	0.168	240 s
FSGS	24.64	**0.895**	0.095	180 s
DNGaussian	24.31	0.886	**0.088**	110 s
Ours ($2.5K$ iterations)	24.53	0.881	0.102	**35 s**
Ours ($5K$ iterations)	**24.81**	0.882	0.101	70 s

Table 1: Quantitative results on Blender with 8 input views.

Method Analysis

Frequency Regularization vs Spatial Annealing

In previous section, we provide a theoretical analysis that clarifies the relationship between the pre-filtering strategy and frequency regularization based on integrated positional encoding. To demonstrate the strategy's validity and to evaluate our method's performance across various base architectures, we apply it using the JAX MipNeRF framework (Barron et al. 2021). In this setup, we adjust the initial sphere size to 0.2, establish the decrease rate ϑ at 1, and set the total number of decrease steps N_{split} to 33.

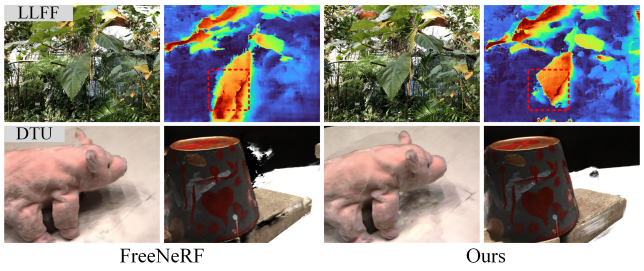


Figure 6: Qualitative comparisons on LLFF and DTU with 3 input views.

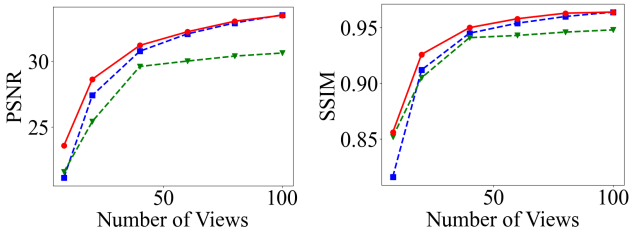
Method	LLFF			DTU		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
MipNeRF	14.62	0.351	0.495	7.64	0.227	0.655
RegNeRF	19.08	0.587	0.336	15.33	0.621	0.341
FreeNeRF	19.63	0.612	0.308	18.02	**0.680**	**0.318**
Ours	**19.67**	**0.616**	**0.288**	**18.15**	0.675	0.320

Table 2: Quantitative results on LLFF and DTU with 3 input views.

Fig. 6 and Tab. 2 present qualitative and quantitative comparisons of our method with FreeNeRF (Yang, Pavone, and Wang 2023). Quantitatively, our method matches or slightly outperforms FreeNeRF, showing up to a 0.05 dB and 0.1 dB increase in PSNR on the LLFF dataset (Mildenhall et al. 2019) and the DTU dataset (Jensen et al. 2014), respectively. Qualitatively, our method provides improved novel views and depth maps. For instance, while FreeNeRF struggles to accurately reconstruct the leaf edges in the first row and the ear of the toy in the second row, our method preserves geometric consistency and produces more detailed rendered results. This advantage likely stems from our method's greater adaptability. Unlike frequency regularization, constrained by a fixed frequency range, our method dynamically adjusts the sample space to a wider range, thereby enhancing global geometry reconstruction.

Robustness to View Sparsity

We validate our method alongside the TriMipRF (Hu et al. 2023) using varying numbers of input views to assess our method's resilience to view sparsity, as shown in Fig. 7. We use the first n images from the Blender's training set as input, where n is varied among 8, 20, 40, 60, 80, and 100, with the dataset containing a total of 100 images. We fix the total iteration steps at 10,000 and the spatial annealing strategy's endpoint at 2,000. The curve of Fig. 7 demonstrates that our method outperforms TriMipRF under different input view conditions. Notably, it shows marked improvements in scenarios with sparse input views, achieving a 2.50 dB and 1.22 dB higher PSNR with 8 and 20 input views, respectively.



(a) PSNR (b) SSIM

Figure 7: Validation of SANeRF's robustness to view sparsity compared with TriMipRF and FreeNeRF.

Ablation Studies

To thoroughly assess the efficacy of our proposed method, we conduct qualitative and quantitative ablation studies using the Blender dataset (Mildenhall et al. 2021) with 8 input views, as detailed in Fig. 8 and Tab. 3. We evaluate the contributions of the newly introduced spatial annealing strategy (SA) and the spherical harmonic mask (SHM) separately. The findings reveal that the spatial annealing strategy significantly enhances performance in a few-shot setting, achieving a 1.2 dB increase in PSNR on the Blender dataset with SA alone. Regarding qualitative results, the rendered RGB images and depth maps demonstrate that our spatial annealing strategy primarily enhances the reconstruction of low-frequency geometry but does not effectively address color distortions. Conversely, the SHM effectively mitigates color distortions but fails to resolve geometric underfitting. The combination of both strategies results in well-established geometry and photorealistic rendering. These results underscore the substantial improvements achieved by our method.

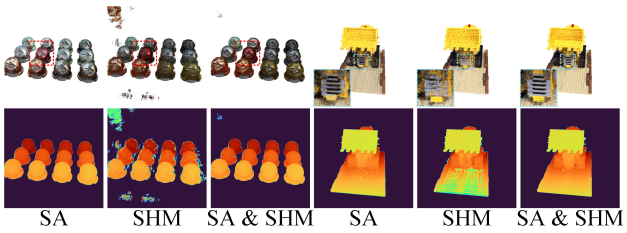


Figure 8: Qualitative ablation study on the Blender dataset.

	SA	SHM	PSNR↑	SSIM↑	LPIPS↓	Train
TriMipRF	×	×	21.63	0.818	0.183	35 s
Ours	✓	×	22.83	0.856	0.120	35 s
	×	✓	22.27	0.825	0.173	
	✓	✓	24.53	0.881	0.102	

Table 3: Quantitative ablation study on the Blender dataset.

Conclusion

In this study, we present a novel spatial annealing strategy tailored for a hybrid architecture equipped with a pre-filtering strategy. This approach adaptively modifies the spatial sampling size using a meticulously designed annealing process, enhancing geometric reconstruction and detail refinement. Remarkably, our approach requires minimal modifications to the base architecture's code to achieve state-of-the-art performance in few-shot scenarios while preserving efficiency. We acknowledge that our spatial annealing strategy possesses the potential to enhance the training stability of diverse architectures crafted with pre-filtering.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grants 92270116 and 61932022. Additionally, this research was financially supported by the Open Research Fund from Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), under Grant No.GML-KF-24-09.

References

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5855–5864.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5470–5479.

Cao, A.; Rockwell, C.; and Johnson, J. 2022. Fwd: Real-time novel view synthesis with forward warping and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15713–15724.

Chen, A.; Xu, Z.; Zhao, F.; Zhang, X.; Xiang, F.; Yu, J.; and Su, H. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF international conference on computer vision*, 14124–14133.

Chen, H.; Gu, J.; Chen, A.; Tian, W.; Tu, Z.; Liu, L.; and Su, H. 2023. Single-stage diffusion nerf: A unified approach to 3d generation and reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2416–2425.

Chibane, J.; Bansal, A.; Lazova, V.; and Pons-Moll, G. 2021. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7911–7920.

Deng, K.; Liu, A.; Zhu, J.-Y.; and Ramanan, D. 2022. Depth-supervised nerf: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12882–12891.

Fang, J.; Yi, T.; Wang, X.; Xie, L.; Zhang, X.; Liu, W.; Nießner, M.; and Tian, Q. 2022. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, 1–9.

Gu, J.; Trevithick, A.; Lin, K.-E.; Susskind, J. M.; Theobalt, C.; Liu, L.; and Ramamoorthi, R. 2023. Nerfdiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*, 11808–11826. PMLR.

Hong, Y.; Zhang, K.; Gu, J.; Bi, S.; Zhou, Y.; Liu, D.; Liu, F.; Sunkavalli, K.; Bui, T.; and Tan, H. 2023. Lrm: Large reconstruction model for single image to 3d. *arXiv preprint arXiv:2311.04400*.

Hu, W.; Wang, Y.; Ma, L.; Yang, B.; Gao, L.; Liu, X.; and Ma, Y. 2023. Tri-miprf: Tri-mip representation for efficient anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 19774–19783.

Jain, A.; Tancik, M.; and Abbeel, P. 2021. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5885–5894.

Jensen, R.; Dahl, A.; Vogiatzis, G.; Tola, E.; and Aanæs, H. 2014. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 406–413.

Kwak, M.-S.; Song, J.; and Kim, S. 2023. GeCoNeRF: few-shot neural radiance fields via geometric consistency. In *Proceedings of the 40th International Conference on Machine Learning*, 18023–18036.

Lao, Y.; Xu, X.; Liu, X.; Zhao, H.; et al. 2024. CorresNeRF: Image Correspondence Priors for Neural Radiance Fields. *Advances in Neural Information Processing Systems*, 36.

Lin, K.-E.; Lin, Y.-C.; Lai, W.-S.; Lin, T.-Y.; Shih, Y.-C.; and Ramamoorthi, R. 2023. Vision transformer for nerf-based view synthesis from a single input image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 806–815.

Liu, J.; Hu, W.; Yang, Z.; Chen, J.; Wang, G.; Chen, X.; Cai, Y.; Gao, H.-a.; and Zhao, H. 2024. Rip-NeRF: Anti-aliasing Radiance Fields with Ripmap-Encoded Platonic Solids. *arXiv preprint arXiv:2405.02386*.

Liu, Y.; Peng, S.; Liu, L.; Wang, Q.; Wang, P.; Theobalt, C.; Zhou, X.; and Wang, W. 2022. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7824–7833.

Liu, Y.-L.; Gao, C.; Meuleman, A.; Tseng, H.-Y.; Saraf, A.; Kim, C.; Chuang, Y.-Y.; Kopf, J.; and Huang, J.-B. 2023. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13–23.

Loshchilov, I.; and Hutter, F. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Martin-Brualla, R.; Radwan, N.; Sajjadi, M. S.; Barron, J. T.; Dosovitskiy, A.; and Duckworth, D. 2021. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7210–7219.

Mildenhall, B.; Hedman, P.; Martin-Brualla, R.; Srinivasan, P. P.; and Barron, J. T. 2022. 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*, 16190–16199.

Mildenhall, B.; Srinivasan, P. P.; Ortiz-Cayon, R.; Kalantari, N. K.; Ramamoorthi, R.; Ng, R.; and Kar, A. 2019. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4): 1–14.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM transactions on graphics (TOG)*, 41(4): 1–15.

Niemeyer, M.; Barron, J. T.; Mildenhall, B.; Sajjadi, M. S.; Geiger, A.; and Radwan, N. 2022. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5480–5490.

Oechsle, M.; Peng, S.; and Geiger, A. 2021. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5589–5599.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Roessle, B.; Barron, J. T.; Mildenhall, B.; Srinivasan, P. P.; and Nießner, M. 2022. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12892–12901.

Seo, S.; Han, D.; Chang, Y.; and Kwak, N. 2023. Mixnerf: Modeling a ray with mixture density for novel view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20659–20668.

Shabanov, A.; Govindarajan, S.; Reading, C.; Goli, L.; Rebain, D.; Yi, K. M.; and Tagliasacchi, A. 2024. BANF: Band-limited Neural Fields for Levels of Detail Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20571–20580.

Sun, J.; Zhang, Z.; Chen, J.; Li, G.; Ji, B.; Zhao, L.; and Xing, W. 2023. VGOS: voxel grid optimization for view synthesis from sparse inputs. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, 1414–1422.

Truong, P.; Rakotosaona, M.-J.; Manhardt, F.; and Tombari, F. 2023. Sparf: Neural radiance fields from sparse and noisy poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4190–4200.

Verbin, D.; Hedman, P.; Mildenhall, B.; Zickler, T.; Barron, J. T.; and Srinivasan, P. P. 2022. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5481–5490. IEEE.

Wang, G.; Chen, Z.; Loy, C. C.; and Liu, Z. 2023. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9065–9076.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *Advances in Neural Information Processing Systems*, 34: 27171–27183.

Xiao, Y.; Liu, X.; Zhai, D.; Jiang, K.; Jiang, J.; and Ji, X. 2024. SGCNeRF: Few-Shot Neural Rendering via Sparse Geometric Consistency Guidance. *arXiv preprint arXiv:2404.00992*.

Yang, J.; Pavone, M.; and Wang, Y. 2023. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8254–8263.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4578–4587.

Zhao, Z.; Fan, F.; Liao, W.; and Yan, J. 2024. Grounding and Enhancing Grid-based Models for Neural Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19425–19435.

Zhu, H.; He, T.; Li, X.; Li, B.; and Chen, Z. 2024. Is vanilla MLP in neural radiance field enough for few-shot view synthesis? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20288–20298.