

Efficient Ray Sampling for Radiance Fields Reconstruction

Shilei Sun

Ming Liu

Beijing Institute of
Technology

China

slnsun@bit.edu.cn
bit411liu@bit.edu.cn

Zhongyi Fan

Yuxue Liu

Chengwei Lv

Beijing Institute of Technology, China

zyfanzy@foxmail.com
18800197258@163.com
1612484269@qq.com

Liquan Dong

Lingqin Kong

Beijing Institute of
Technology

China

kylind@bit.edu.cn
konglingqin3025@bit.edu.cn

Figure 1. We incorporate a novel ray sampling method guided by pixel regions and depth into the existing ENeRF framework [15]. Leveraging the DTU dataset, we pretrain our model to generalize across varied scenes, enabling reconstruction of neural radiance fields from merely 3 input views (a), while reserving other views for testing (b). Our model synthesizes photorealistic novel views (c) qualitatively exceeding ENeRF (d). Although artifacts persist, it can be greatly improved by fine-tuning our reconstruction on three images for 20 min (e), which has better quality than the ENeRF’s result from 1 hour fine-tuning (f) and NeRF’s result from 10.2h per-scen optimization (g).

Abstract

Accelerating neural radiance fields training is of substantial practical value, as the ray sampling strategy profoundly impacts network convergence. More efficient ray sampling can thus directly enhance existing NeRF models’ training efficiency. We therefore propose a novel ray sampling approach for neural radiance fields that improves training efficiency while retaining photorealistic rendering results. First, we analyze the relationship between the pixel loss distribution of sampled rays and rendering quality. This reveals redundancy in the original NeRF’s uniform ray sampling. Guided by this finding, we develop a sam-

pling method leveraging pixel regions and depth boundaries. Our main idea is to sample fewer rays in training views, yet with each ray more informative for scene fitting. Sampling probability increases in pixel areas exhibiting significant color and depth variation, greatly reducing wasteful rays from other regions without sacrificing precision. Through this method, not only can the convergence of the network be accelerated, but the spatial geometry of a scene can also be perceived more accurately. Rendering outputs are enhanced, especially for texture-complex regions. Experiments demonstrate that our method significantly outperforms state-of-the-art techniques on public benchmark

datasets.

1. Introduction

Neural radiance fields (NeRF) [18] represent 3D scenes as implicit radiance fields fitted by neural networks. Given an arbitrary 3D location and viewing direction, NeRF outputs an emitted color and volume density. The training process of NeRF first samples pixels from source view images. Numerous spatial points are then sampled along each corresponding ray path. MLPs estimate volume density and color for each directional point. Finally, volume rendering integrates these outputs to predict pixel colors, which are compared against ground truth pixels to supervise network training.

However, a significant limitation of NeRF is the extensive time required to fit scenes. Initial NeRF training on a novel scene with a single GPU spanned approximately two days. Subsequent works [28, 34, 41] revealed the bottleneck is not the learning radiance field itself, but inefficient oversampling of empty space during training. This aligns with the high sparsity of real 3D scenes. Yet NeRF’s initialization assumes a uniform density distribution, mismatched to the distributions of real-world surface. From this perspective, we expect that by learning the density function of NeRF to simulate the surface distribution in the real world, the fitting efficiency of NeRF to the scene geometry can be further improved.

Analysis of multiple NeRF models’ loss distributions during training reveals slower fitting efficiency in image regions with drastic color variations compared to other areas. This discrepancy manifests in 3D scene regions exhibiting significant detail texture and depth variations. Therefore, we propose the redundancy hypothesis for ray sampling in neural radiance fields. In addition to oversampling empty space, we find suboptimality in the ray sampling scheme itself during training. As 3D space is incrementally fitted, radiance fields first estimate the flat and low-texture areas more accurately, before less efficient fitting of regions with substantial depth and color changes. This follows the intrinsic fitting sequence of deep neural networks, which encode low-frequency information first, followed by the high-frequency information contained in intricately detailed textures.

Based on this hypothesis, we introduce a ray sampling strategy guided by pixel regions and depth boundaries. The key idea is to train with fewer sampled rays, yet more informative for scene fitting within source views. We increase sampling probability in areas with significant color and depth variations, greatly reducing redundant rays from other pixels to accelerate network fitting and improve efficiency. Our method enables faster per-scene optimization, achieving the same convergence effect in minutes through

fine-tuning reconstructions from given views. Moreover, our sampling strategy better perceives spatial geometry, improving rendering quality especially for rich-texture regions. Since ray sampling is ubiquitous in NeRF, our strategy can directly transfer to existing models. We demonstrate this by applying our approach to ENeRF [15], achieving state-of-the-art performance on both view synthesis and training time, as shown in Fig. 1.

2. Related work

Novel view synthesis. View synthesis methods typically rely on intermediate 3D scene representations. NeRF [18] pioneered photorealistic novel view synthesis with implicit scene functions. Unlike previous methods, NeRF attempts to recreate an implicit volume as the intermediate representation. Improvements based on the problems of NeRF include accelerating the training [3, 6, 19, 31, 39, 40] and rendering [2, 15, 16, 41, 43] process, targeting on dynamic scenes [8, 13, 22, 36], better generalization [4, 15, 35, 37, 42], training with fewer viewpoints [29, 33, 36]. There are also some researches on the application of NeRF. The process of estimating different model parameters (camera, geometry, materials, lighting parameters) from realistic data is called inverse rendering, which purpose is to estimate the pose of camera [5, 45], edit materials and lighting [30]. Controlled editing is a top priority in computer vision, and the development of NeRF in this area can be described as a separate branch, from EIDtNeRF [17] to GIRAFFE [21] (CVPR2021 Best Paper). These methods are mainly combined with GAN [9] and NeRF to achieve controlled editing, and the combination of diffusion model [38] and NeRF is becoming increasingly popular [14, 24, 26]. Another important area is digital human body, which includes the Facial Avatar [7, 27] and Human Bodies [15, 22, 36]. After a long time of development, NeRF has shown better and better performance for view synthesis.

Training and rendering acceleration. NeRF’s substantial computational requirements for training and rendering motivate research into efficiency improvements. For rendering acceleration, octrees-based [41] approaches decompose NeRF into numerous small MLPs to enable model distillation. Some methods [11, 15, 20] improve the sampling strategy to accelerate the rendering process. ENeRF [15] uses geometric feature bodies to guide the radiance fields sampling near surfaces, significantly improving the training and rendering efficiency. For training acceleration, DVGO [32] employs explicit discrete volume representations to rapidly simulate radiance fields, reducing training time to a few minutes. Plenoxels [40] uses a sparse voxel grid and spherical harmonics for fast viewpoint rendering. Instant-NGP [19] applies a multiresolution hash encoding and optimized GPU implementation to train high-quality neural graphics primitives in seconds. In summary, the methods of render-

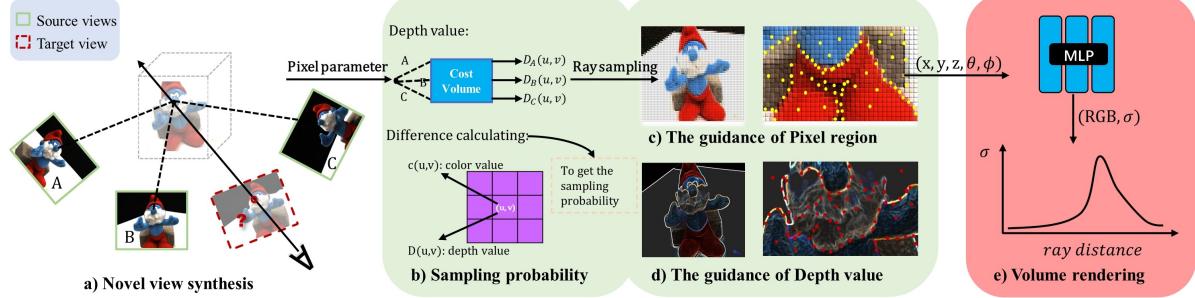


Figure 2. Overview of our ray sampling model. To render a novel target view, multi-view source images of static or dynamic scenes at one frame are utilized (a). The cost volume [15] is leveraged to obtain depth information, and we use it and color value to get the sampling probability of the pixel (u, v) (b). With the sampling probability of pixels from source view images I_i , during ray sampling, our model samples additional rays in pixel areas exhibiting substantial color variation (c). Similarly, guided by the depth-based sampling strategy, more rays are sampled in regions where geometric depth varies markedly (d). Finally, through volume rendering (e), our network outputs colors and densities along the sampled rays to render pixel color.

ing and training acceleration make NeRF more practical, from optimized model architectures and sampling strategies to discrete representations and GPU-optimized implementations.

However, current acceleration methods predominately focus on static scenes, with limited work on modeling dynamics [15, 34, 44]. Recently, Fourier PlenOctrees [34] extends PlenOctrees [41] to dynamic settings by representing time-varying density and color in the frequency domain. However, this approach necessitates intensive view sampling and prolonged training. In contrast, our proposed method enables efficient training for dynamic scene using only sparse camera viewpoints.

Ray sampling methods. The conventional ray sampling strategy of neural radiance fields utilizes uniform distribution [18], with evenly distributed ray locations across training images. Later work introduced adaptive sampling [15, 27, 29] guided by rendering loss to focus on inaccurately rendered regions. This approach is analogous to sample mining in deep learning training, where pixels exhibiting poor convergence are identified via loss map to receive additional sampling focus. There are also some ray sampling methods that sample fewer points in per-ray. [1, 4, 23] achieve acceleration by predicting surfaces and sampling near the surfaces, reducing the number of necessary sampling points for rendering at each ray. Uniquely, we sample varying ray densities based on depth boundaries and pixel regions. This simultaneously accelerates network convergence and improves geometry perception.

3. Method

Depth information of rough scenes is obtained from the depth estimation module in ENeRF [15]. An overview of our proposed method is illustrated in Fig. 2. First, the rendering process of neural radiance fields is introduced (Sec.

3.1). Then, through analysis of observations from several experiments in the training stage of neural radiance fields, we propose a redundant hypothesis regarding ray sampling for neural radiance fields (Sec. 3.2). By combining estimated scene depth information and image pixel distribution, we devise a ray sampling strategy guided by pixel regions (Sec. 3.3) and depth boundaries (Sec. 3.4). The entire network is trained in an end-to-end manner using RGB images as supervision (Sec. 3.5).

3.1. Preliminaries

Description of the entire framework of neural radiance fields (NeRF) : Given a 3D location $p = (x, y, z) \in \mathbb{R}^3$ and its associated 2D viewing direction $d = (\theta, \phi) \in \mathbb{R}^2$, the MLP network $F_\Theta : (p, d) \rightarrow (c, \sigma)$ maps the 5D coordinate point (p, d) to its corresponding volume density σ and directional emitted color $c = (r, g, b)$. The network parameters are then optimized by minimizing the color error at the pixel level from volume rendering.

During the training stage, a certain number of 5D coordinate points are sampled from each ray $r_i^k \in R_i$, where r_i^k represents the k th ray taken from the i th training image I_i , and R_i is the set of rays generated from images $I_i \in I, i = 1, 2, \dots, N$ (where N is the number of trained images). The volume rendering $V_\Theta : r_i^k \rightarrow \hat{c}_i^k$ then makes color predictions at the pixel level along the colors and densities of the sampled ray to obtain the color value of each pixel \hat{c}_i^k . The final optimization goal of the model is to minimize the volume rendering loss from all the sampled rays, as follows:

$$\min_\Theta \mathcal{L} = \frac{1}{NM} \sum_{r_i^k \in R_i} \|V_\Theta(r_i^k) - c_i^k\|_2^2, \quad (1)$$

where c_i^k is the true color of pixels, M is the number of R_i , which represents the number of sampled rays per image.

3.2. Efficiency of ray sampling

This section describes the problem of ray sampling in volume rendering from a holistic perspective. The key question is how to obtain all the sampled rays for the training images. Generally, the method based on classical radiance fields randomly selects M pixels in image I_i and emits M rays from the selected pixels along the observed direction. Since the ray parameters depend solely on the pixel position and viewing direction, $r_i(u, v)$ can represent a ray originating from pixel (u, v) . The ray direction aligns with the viewing direction of image I_i . In this manner, the positions of the sampled rays follow a uniform distribution across training images:

$$r_i(u, v) \sim U(I_i), u \in [0, H_i], v \in [0, W_i], \quad (2)$$

where H_i and W_i are the height and width of image I_i , respectively. However, this distribution fails to match the real-world surface, which is defined as $\alpha \sim E(w)$, where α is the scene density. $E(w)$ represents the density distribution that depends on scene geometry and appearance.

The uniform ray sampling strategy is effective in practice, but there remain some potential issues. First, this approach fails to fully capture contextual information in images, specifically the non-uniform distribution of color differences among pixels. In fact, there exist numerous minute regions in an image where successive pixels tend to exhibit similar colors. The color variation between certain pixel areas and their neighbors may be negligible, hence the sampled rays acquire less information in these regions.

Moreover, depth variation within a scene has a substantial impact on image information. Compared to regions with relatively flat depth, areas in a scene where depth undergoes abrupt changes contain more high-frequency signals. Rays in these regions necessitate sampling a greater number of 3D points to accurately determine their spatial depth. In trained radiance fields, this phenomenon commonly manifests as blurring of edges of three-dimensional objects. Consequently, in regions with flat color and depth variation, the rendering loss rapidly converges as the density of most sampled points along the rays approaches zero. In such regions, sampling fewer rays suffices to achieve a more precise fit of the radiance fields. In contrast, regions with stronger color and depth variation encompass more intricate scene information, necessitating sampling additional rays and spatial points to capture fine details.

Therefore, during the training phase, sampling a substantial number of rays is unwarranted for regions with relatively simple structure. Instead, focus should be directed toward areas with intricate structure. Loss in regions of intricate structure is always optimized more slowly. This demonstrates that existing methods of ray sampling during NeRF training contain redundancies, directly impacting the convergence efficiency of radiance fields.

Based on the preceding analysis, two sampling strategies are proposed to optimize the distribution of sampled rays for input images. The first strategy calculates the prior probability distribution based on the color of pixel region, while the second strategy dynamically adjusts the ray sampling distribution utilizing estimated depth. And we can make a learning density function of NeRF to simulate the surface distribution in the real world

$$r_i(u, v) \sim P_{\text{samp}}(I_i) = P_c(I_i) + P_d(I_i), \quad (3)$$

where P_{samp} is the probability density function, which can guide the radiance fields to sample rays. P_c and P_d represent the the probability density functions of pixel-guided and depth-guided, respectively.

3.3. Sampling ray with the guidance of pixel region

The information density across different regions of an input-view image largely depends on the intensity of color and depth variation. To identify high and low information density areas in the image, color variation in pixel regions is leveraged to capture image information and guide non-uniform ray sampling in radiance fields. Intuitively, a pixel exhibiting identical color to its neighborhood likely contains less information, whereas a pixel with substantial color differences from its surroundings encompasses more scene details. Therefore, we quantify the information richness level by computing variation intensity within pixel neighborhoods, subsequently replacing the uniform distribution of sampled rays during training.

In the ray sampling stage, we design a probability density function $P_c(u, v)$ that maps the position of pixel (u, v) to a prior probability, denoting the likelihood of sampled rays at that location. We define P_c as the standard deviation of a $n \times n$ neighborhood of pixel colors c centered on pixel (u, v) , as follows:

$$P_c(u, v) = \text{std}(c(u, v)) = \sqrt{\frac{1}{n^2} \sum_{x', y'} [c(x', y') - \bar{c}]^2}, \\ x' \in [u - \frac{n-1}{2}, u + \frac{n-1}{2}], \quad (4) \\ y' \in [v - \frac{n-1}{2}, v + \frac{n-1}{2}],$$

where \bar{c} denotes the mean color of the $n \times n$ pixels centered on pixel (u, v) . In the experiments, we set $n = 3$.

To facilitate computation, the standard deviation equation is utilized for code implementation:

$$\text{std}(c(u, v)) = \sqrt{E(c(x', y')^2) - E(c(x', y'))^2}. \quad (5)$$

If pixel (u, v) exhibits identical color to its neighbors, then $P_c(u, v) = 0$. Conversely, if the pixel color differs substantially from the surrounding color, $P_c(u, v)$ will be

higher, denoting this position contains more scene details. In the 2D pixel space, pixels with a higher value of P_c typically corresponds to this area with sharper color variation. Correspondingly, in 3D space, some of these pixels match planes with richer texture details, while others match locations with sharper density variation, representing the boundary surfaces of 3D objects. To balance the discrepancy between maximum and minimum values for P_c , normalization is necessitated, as depicted:

$$P'_c(u, v) = \frac{\text{clamp}(s, \max(P_c(u, v)))}{\max(P_c(u, v))}, \quad (6)$$

where the threshold s is defined as $0.01 \times \text{mean}(P_c(u, v))$. Values below s will be scaled to s to avoid sampling too few rays at certain locations, which can prevent model underfitting. The normalization function P'_c generates a specific probability distribution for each input source view. After normalization, distribution $P'_c(u, v)$ is within the interval $[0, 1]$, then the distribution P'_c can substitute the uniform distribution to sample rays utilized for NeRF training:

$$r_i(u, v) \sim P'_c(u, v), u \in [0, H_i], v \in [0, W_i], \quad (7)$$

where $r_i(u, v)$ represents the sampled ray from pixel (u, v) of image I_i . More details about *clamp* are in the supplementary material.

3.4. Depth-guided ray sampling

Analogous to the guided sampling of pixel regions delineated in Sec. 3.3, we devise a depth-guided ray sampling strategy based on the scene depth probability distribution. The depth prediction follows the coarse-to-fine approach in ENeRF [15]. In ENeRF, for pixel (u, v) situated in the target view's feature map, linear interpolation of the depth probability volume $P_i(u, v)$ is requisite to ascertain its probability at a specific depth plane L_i . The depth value $D(u, v)$ at pixel (u, v) constitutes the expectation of the depth probability distribution, where L_i denotes the different depth planes sampled from $[L_{\min}, L_{\max}]$:

$$D(u, v) = \sum_{i=1}^{N_d} P_i(u, v) \cdot L_i(u, v), \quad (8)$$

where N_d is the number of sampled depth planes.

Applying the scheme analogous to Eqn. 4 on the desired depth map of the rendering view (from Eqn. 8) obtains the sampling probability distribution guided by the depth boundary

$$P_d(u, v) = \text{std}(D(u, v)) = \sqrt{\frac{1}{n^2} \sum_{x', y'} [D(x', y') - \bar{D}]^2}, \quad (9)$$

where \bar{D} denotes the mean depth value. The x' and y' is also same as Eqn. 4. The same normalization operation

is utilized to avoid underfitting for certain rays, getting the normalization $P'_d(u, v)$.

In the training stage, two sampling strategies guided by pixel region and depth boundary impact the ray sampling process through weighted fusion, as delineated in the following equation:

$$P_{\text{samp}}(u, v) = \beta P'_c(u, v) + (1 - \beta) P'_d(u, v), \quad (10)$$

where β represents a weight parameter that modulates the proportional contributions of the two probability distributions for fusion. Given that the depth information is progressively fitted during training, the value of β incrementally rises from 0 to 0.5 across training iterations. Please refer to the supplementary material for further details on the analysis of sampling efficiency and methodology.

3.5. Training

We integrate our ray sampling method, guided by pixel region and depth boundary, into ENeRF [15]. During training, our ray sampling approach samples fewer rays per iteration than ENeRF, while achieving improved performance. Consequently, the ENeRF model employing our approach occupies less GPU memory. Thus, \mathcal{L}_{mse} and \mathcal{L}_{perc} can be utilized for supervision:

$$\begin{aligned} \mathcal{L}_{mse} &= \frac{1}{N_r} \sum_{i=1}^{N_r} \| \hat{C}_i - C_i \|^2, \\ \mathcal{L}_{perc} &= \frac{1}{N_i} \sum_{i=1}^{N_i} \| \Phi(\hat{I}_i) - \Phi(I_i) \|, \end{aligned} \quad (11)$$

where \mathcal{L}_{mse} represents the mean square error between the color of rendered pixels and ground-truth pixels. \mathcal{L}_{perc} denotes perceptual losses obtained by sampling image blocks. N_r signifies the number of sampled rays per iteration and N_i denotes the number of image blocks. Φ represents the definition of perceptual function [15]. \hat{C}_i and C_i signify the rendered and ground-truth colors, respectively.

The final loss function is as follows:

$$\mathcal{L} = \mathcal{L}_{mse} + \lambda' \mathcal{L}_{perc}. \quad (12)$$

In the experiments, we set $\lambda' = 0.01$.

4. Experiments

4.1. Implementation Details

Experiments. For the ENeRF model [15] utilizing our ray sampling strategy, the same experimental setup is followed. The model is exclusively trained on the DTU dataset [10] to accomplish rapid generalization for novel scenes. Subsequently, the model undergoes fine-tuning for a few epochs on specified novel scenes. Rays are sampled under the guidance of pixel region and depth boundary, where 2 spatial

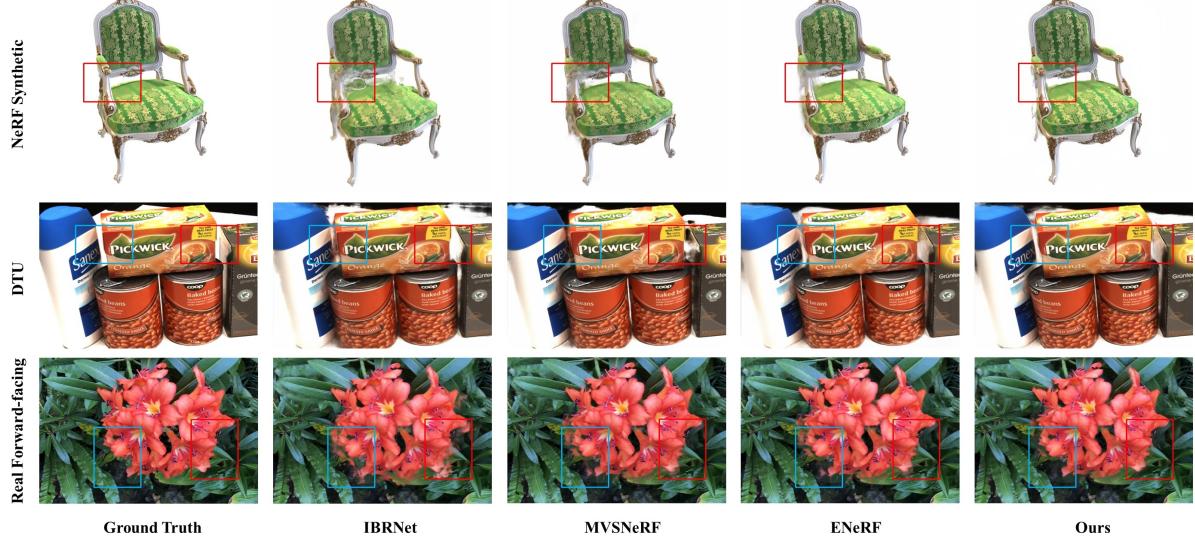


Figure 3. Qualitative comparison of view synthesis results under the generalization setting.

Table 1. Quantitative results of static scenes. “Generalization” means that the model is not additionally fine-tuned. At the bottom, results are presented of different baselines fine-tuning their models on specific scenes. Our approach attains the optimal performance on PSNR and LPIPS, alongside outstanding results on SSIM. The baseline results are obtained from ENeRF [15].

Methods	Training settings	DTU			NeRF Synthetic			Real Forward-facing		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
IBRNet	Generalization (only pre-trained)	26.04	0.917	0.191	22.44	0.874	0.195	21.79	0.786	0.279
MVSNeRF		26.63	0.931	0.168	23.62	0.897	0.176	21.93	0.795	0.252
ENeRF		27.61	0.956	0.091	26.65	0.947	0.072	22.78	0.808	0.209
Ours		27.85	0.941	0.059	26.82	0.949	0.062	23.01	0.821	0.194
IBRNet(1h)	Fine-tuning on given scene	31.35	0.956	0.131	25.62	0.939	0.111	24.88	0.861	0.189
MVSNeRF(15min)		28.51	0.933	0.179	27.07	0.931	0.168	25.45	0.877	0.192
ENeRF(15min)		28.73	0.956	0.093	27.20	0.951	0.066	24.59	0.857	0.173
ENeRF(1h)		28.87	0.957	0.090	27.57	0.954	0.063	24.89	0.865	0.159
Ours(15min)		29.30	0.953	0.092	27.32	0.943	0.066	25.04	0.861	0.166
Ours(30min)		30.61	0.956	0.088	27.62	0.952	0.055	26.19	0.882	0.0145

points are sampled along each ray. In experiments, ablation studies are conducted to assess the performance of various ray sampling strategies on rendering outcomes.

Datasets. We use the same train-test split and evaluation setting from ENERF [15]. For each test scene, 16 views are chosen as seen views and the remaining 4 views as novel views for evaluation. The framework is pre-trained on the DTU dataset to learn generalizable neural radiance fields. Real Forward-facing and NeRF Synthetic datasets [18] are also tested. For dynamic scenes, the ZJU-MoCap [22] and our produced digital human dataset are evaluated. Our dataset employs identical setup to ZJU-MoCap and additionally provides synchronized and calibrated multi-view videos with simple backgrounds and high-quality masks. The input image resolution is 512×512 for dynamic scenes.

Baselines. For static scenes, comparisons are made with state-of-the-art generalizable radiance field methods [4, 15, 35, 42] under both direct generalization and fine-tuning on

particular scenes, including IBRNet [35], MVSNeRF [4], and ENeRF [15]. The same settings as ENeRF are followed on the aforementioned benchmarks, with baseline results obtained from ENeRF. For dynamic scenes, comparisons are drawn against DNeRF [25], IBRNet, and ENeRF. For all baselines [4, 15, 25, 35], their released code is utilized with retraining under our experimental setting.

4.2. Comparisons on view synthesis

Comparisons on static scenes. All baselines [4, 15, 35] are compared on these datasets utilizing identical input views, with each scene tested on 4 novel views. Quantitative results for our model and other baselines on the NeRF Synthetic, Real Forward-facing, and DTU datasets are presented in Tab. 1, demonstrating that our method achieves state-of-the-art performance. Qualitative results are shown in Fig. 3 and 4. From the quantitative and qualitative results, it can be seen that our method achieves obvious per-

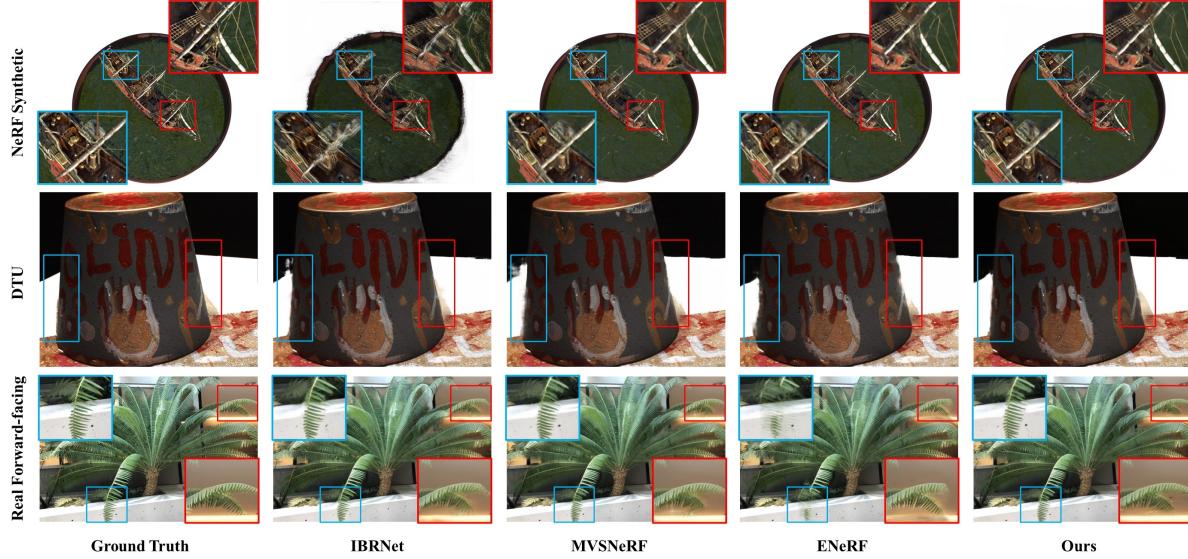


Figure 4. Qualitative comparison of view synthesis results under fine-tuning on the given scene setting.

formance optimization in better rendering effect than existing models. Specifically, improvements are attained at sampled boundaries by directly applying the pre-trained model and briefly fine-tuning. This signifies that rendering quality is enhanced by our ray sampling approach, especially in regions with enriched texture. Implementation of our ray sampling technique on other baselines [4, 35] can also boost network convergence efficiency and rendering quality, as elaborated in the supplementary material.

Comparisons on dynamic scenes. The ZJU-MoCap and our dataset provide multi-view human-body images with background removed. For DNeRF [25], identical settings to ENeRF are utilized, with the video sequence divided into several sub-sequences. We validate the performance of our method in modeling dynamic human bodies using ZJU-MoCap and our dataset. Quantitative results are summarized in Tab. 2, while qualitative results for our proposed approach are depicted in Fig. 5. The quantitative analysis demonstrates our method realizes state-of-the-art performance after abbreviated fine-tuning time. This signifies our approach enhances training efficiency (less fine-tuning time) while sustaining significant rendering capability.

4.3. Analysis of convergence efficiency

Owing to the various ray sampling strategies proposed herein, the training efficiency of our model is markedly superior to existing generalizable neural radiance fields [4, 15, 35, 42]. The Adam optimizer [12] is utilized for training on the RTX 3090 GPU with an initial learning rate of 5e-4, halving the learning rate after 30k iterations (1.7x faster than ENeRF). The model converges after 100k iterations (2x faster than ENeRF). Fig. 6 presents a comparison

Table 2. Quantitative results of dynamic scenes. “0” indicates that the pre-trained model is directly used in the corresponding scene. All methods are evaluated on the same machine with corresponding fine-tuning time.

Methods	Training time	ZJU-MoCap			Our dataset		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
DNeRF	10h	26.53	0.889	0.154	24.38	0.877	0.136
IBRNet	0	29.46	0.947	0.094	26.89	0.920	0.065
IBRNet	2h	32.38	0.968	0.065	28.03	0.926	0.059
ENeRF	0	31.21	0.970	0.041	29.41	0.949	0.082
ENeRF	2h	32.52	0.978	0.030	30.29	0.952	0.080
Ours	0h	32.49	0.972	0.031	31.15	0.964	0.066
Ours	1h	33.61	0.977	0.025	31.81	0.969	0.062

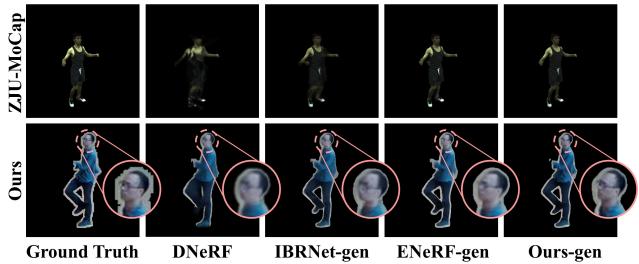


Figure 5. Qualitative comparison of view synthesis results on dynamic scenes. With the exception of DNeRF, all methods directly apply the pre-trained model to the scene without fine-tuning.

of training time and PSNR for our proposed model, ENeRF, and MVSNeRF using the RTX 3090 GPU on the DTU dataset under random parameter initialization.

In this figure, merely the PSNR and time curves of the first 10 epochs are delineated. It can be discerned that the convergence rapidity of our proposed methodology is conspicuously superior to that of ENeRF and MVSNeRF dur-

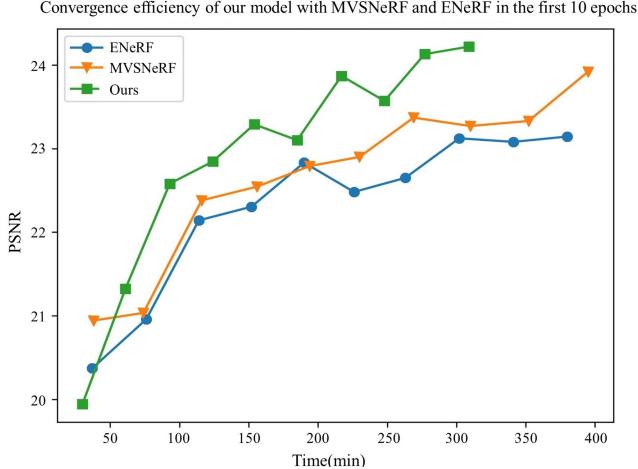


Figure 6. **Comparison of the efficiency of convergence of our method with other baselines.** Our model attains superior PSNR values (higher is preferable) given identical training durations, and accomplishes equivalent training results in markedly shorter convergence time versus MVSNeRF and ENeRF. This demonstrates that our approach achieves optimal convergence efficiency.

ing the training process. The predominant rationale is that we elect a more efficient ray sampling strategy, which allocates more training resources to regions with intricate information. This accelerates the convergence of the neural radiance fields on one hand. On the other hand, it also facilitates the radiance fields to perceive the scene depth more accurately, thereby ameliorating the final rendering quality.

4.4. Ablation studies

To compare the convergence acceleration afforded by the aforementioned ray sampling strategies, on the DTU dataset, we compare the convergence efficiency and results of ray sampling approaches guided by pixel regions and depth boundaries, respectively, against an adaptive ray sampling approach. The earliest convergence timepoints and associated rendering qualities for the different ray sampling strategies are appraised. Subsequently, the models are trained further without guided strategies until convergence, thereby evaluating the final performance enhancement conferred by the sampling strategies.

The first row of Tab. 3 delineates the convergence performance and time utilizing the default random ray sampling approach for neural radiance fields. The divergent sampling methodologies are ultimately trained for identical timespans to compare the final rendering qualities. As depicted in Tab. 3, the proposed ray sampling strategies accelerate the convergence rate of neural radiance fields to varying extents. With adequate prolonged training, they conspicuously refine the rendering quality.

Among the approaches, the ray sampling methodology

Table 3. **Quantitative ablation of the design choices on DTU dataset.** “Pixel region” and “Depth boundary” denote guiding the ray sampling from the neural radiance fields based on pixel regions and depth boundaries, respectively. The rendering resolution is configured to 512×640 .

Pixel region	Depth boundary	Adaptive ray sampling	PSNR (convergence)	Convergence time(h)	PSNR
✓	✓	✓	27.53	22.5	27.53
			26.63	14.2	27.58
	✓	✓	26.97	13.8	27.61
		✓	26.92	14.6	27.11
✓	✓	✓	27.65	8.4	27.85
✓	✓	✓	27.43	8.2	27.55

guided by pixel regions and depth boundaries demonstrates the most expeditious convergence, attaining comparable quality to the final rendering of the default random sampling upon convergence. These experimental findings corroborate our hypothesis regarding ray sampling redundancy, and validate the efficacy of our sampling strategy. The adaptive sampling approach analogously enhances training efficiency, akin to the online example mining technique in deep learning. However, the data indicates the convergence acceleration effect of adaptive sampling strategy is worse than our ray sampling strategies, with the combination of all three failing to confer additional improvements. This implies the sampling strategy guided by pixel regions and depth boundary subsumes the advantages of adaptive sampling, while proffering prior information more congruous with real-world scene observation, as reflected in the loss distribution. More details regarding the qualitative results of ablation experiments and adaptive ray sampling can be found in the supplementary material.

5. Conclusion

We propose a ray sampling approach that enhances both training efficiency and neural rendering quality. Our model samples rays under the guidance of pixel regions and depth boundaries, thereby densifying sampling in areas with greater color and depth variations. By concentrating training and fine-tuning on regions with intricate textures, our model effectively mitigates the redundancy of the sampled rays. The ENeRF framework [15] integrating our ray sampling method demonstrates superior convergence efficiency and rendering quality compared to the vanilla ENeRF [15], while also generalizing better across diverse testing datasets with abbreviated training durations, surpassing concurrent works [4, 15, 25, 35]. With merely 2-3 input views, our model requires dramatically fewer iterations for fine-tuning on novel scenes to achieve competitive performance versus these baselines [4, 15, 25, 35]. Applying our sampling approach to existing models [4, 35] also confers improvements, further validating the broad utility of our method.

References

- [1] Relja Arandjelović and Andrew Zisserman. Nerf in detail: Learning to sample for view synthesis. *ArXiv*, abs/2106.05264, 2021. 3
- [2] Junli Cao, Huan Wang, Pavlo Chemerys, Vladislav Shakhrai, Ju Hu, Yun Fu, Denys Makoviichuk, S. Tulyakov, and Jian Ren. Real-time neural light field on mobile devices. *ArXiv*, abs/2212.08057, 2022. 2
- [3] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. *ArXiv*, abs/2203.09517, 2022. 2
- [4] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14104–14113, 2021. 2, 3, 6, 7, 8, 12
- [5] Yue Chen, Xingyu Chen, Xuan Wang, Qi Zhang, Yu Guo, Ying Shan, and Fei Wang. Local-to-global registration for bundle-adjusting neural radiance fields. *ArXiv*, abs/2211.11505, 2022. 2
- [6] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12872–12881, 2021. 2
- [7] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8645–8654, 2020. 2
- [8] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5692–5701, 2021. 2
- [9] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014. 2
- [10] Rasmus Ramsbøl Jensen, A. Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 406–413, 2014. 5
- [11] Petr Kellnhofer, Lars Jebe, Andrew Jones, Ryan P. Spicer, Kari Pulli, and Gordon Wetzstein. Neural lumigraph rendering. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4285–4295, 2021. 2
- [12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *Corr*, abs/1412.6980, 2014. 7
- [13] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6494–6504, 2020. 2
- [14] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. *ArXiv*, abs/2211.10440, 2022. 2
- [15] Haotong Lin, Sida Peng, Zhen Xu, Yunzhi Yan, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Efficient neural radiance fields for interactive free-viewpoint video. *SIGGRAPH Asia 2022 Conference Papers*, 2021. 1, 2, 3, 5, 6, 7, 8, 11, 12, 13, 14
- [16] David B. Lindell, Julien N. P. Martel, and Gordon Wetzstein. Autoint: Automatic integration for fast neural volume rendering. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14551–14560, 2020. 2
- [17] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan C. Russell. Editing conditional radiance fields. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5753–5763, 2021. 2
- [18] 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. *ArXiv*, abs/2003.08934, 2020. 2, 3, 6
- [19] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41:1 – 15, 2022. 2
- [20] Thomas Neff, Pascal Stadlbauer, Mathias Parger, Andreas Kurz, Joerg H. Mueller, Chakravarty Reddy Alla Chaitanya, Anton Kaplanyan, and Markus Steinberger. Donerf: Towards real-time rendering of compact neural radiance fields using depth oracle networks. *Computer Graphics Forum*, 40, 2021. 2
- [21] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11448–11459, 2020. 2
- [22] Sida Peng, Yuanqing Zhang, Yinghao Xu, Qianqian Wang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Neural body: Implicit neural representations with structured latent codes for novel view synthesis of dynamic humans. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9050–9059, 2020. 2, 6
- [23] Martin Piala and Ronald Clark. Terminerf: Ray termination prediction for efficient neural rendering. *2021 International Conference on 3D Vision (3DV)*, pages 1106–1114, 2021. 3
- [24] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *ArXiv*, abs/2209.14988, 2022. 2
- [25] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10313–10322, 2020. 6, 7, 8
- [26] Amit Raj, Srinivas Kaza, Ben Poole, Michael Niemeyer, Nataniel Ruiz, Ben Mildenhall, Shiran Zada, Kfir Aberman, Michael Rubinstein, Jonathan T. Barron, Yuanzhen Li, and Varun Jampani. Dreambooth3d: Subject-driven text-to-3d generation. *ArXiv*, abs/2303.13508, 2023. 2
- [27] Amit Raj, Michael Zollhoefer, Tomas Simon, Jason M. Saragih, Shunsuke Saito, James Hays, and Stephen Lombardi. Pva: Pixel-aligned volumetric avatars. *ArXiv*, abs/2101.02697, 2021. 2, 3

- [28] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14315–14325, 2021. 2
- [29] Konstantinos Rematas, An Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Thomas A. Funkhouser, and Vittorio Ferrari. Urban radiance fields. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12922–12932, 2021. 2, 3
- [30] Pratul P. Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T. Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7491–7500, 2020. 2
- [31] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5449–5459, 2021. 2
- [32] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5449–5459, 2021. 2
- [33] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8238–8248, 2022. 2
- [34] Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yanshun Zhang, Yingliang Zhang, Minye Wu, Lan Xu, and Jingyi Yu. Fourier plenocubes for dynamic radiance field rendering in real-time. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13514–13524, 2022. 2, 3
- [35] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas A. Funkhouser. Ibrnet: Learning multi-view image-based rendering. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4688–4697, 2021. 2, 6, 7, 8, 12
- [36] Chung-Yi Weng, Brian Curless, Pratul P. Srinivasan, Jonathan T. Barron, and Ira Kemelmacher-Shlizerman. Humannerf: Free-viewpoint rendering of moving people from monocular video. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16189–16199, 2022. 2
- [37] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5428–5438, 2022. 2
- [38] Ling Yang, Zhilong Zhang, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Ming-Hsuan Yang, and Bin Cui. Diffusion models: A comprehensive survey of methods and applications. *ArXiv*, abs/2209.00796, 2022. 2
- [39] Lior Yariv, Peter Hedman, Christian Reiser, Dor Verbin, Pratul P. Srinivasan, Richard Szeliski, Jonathan T. Barron, and Ben Mildenhall. Bakedsdf: Meshing neural sdf’s for real-time view synthesis. *ArXiv*, abs/2302.14859, 2023. 2
- [40] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5491–5500, 2021. 2
- [41] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenotrees for real-time rendering of neural radiance fields. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5732–5741, 2021. 2, 3
- [42] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4576–4585, 2020. 2, 6, 7, 12
- [43] Heng Yu, Joel Julin, Zoltán Ádám Milacskai, Koichiro Niinuma, and László A. Jeni. Dylin: Making light field networks dynamic. *ArXiv*, abs/2303.14243, 2023. 2
- [44] Heng Yu, Joel Julin, Zoltán Ádám Milacskai, Koichiro Niinuma, and László A. Jeni. Dylin: Making light field networks dynamic. *ArXiv*, abs/2303.14243, 2023. 3
- [45] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12776–12786, 2021. 2

Supplementary Material

A. Method Details

Adaptive ray sampling. The adaptive ray sampling approach enables the radiance fields to adaptively sample rays in regions with inaccurate rendering. This process is analogous to sample mining during training in conventional deep learning, striving to identify pixels with inadequate convergence via the loss function and confer greater priority.

The optimization of the adaptive ray sampling comprises two stages. Initially, a uniform distribution of pixels s_i is randomly sampled from the input source views to update the network parameters and compute the loss. This can be achieved by partitioning each image into an $[8 \times 8]$ grid and calculating the mean loss within each square region R_j ($j = 1, 2, \dots, 64$):

$$\mathcal{H}_i[j] = \frac{1}{|r_j|} \sum_{(u,v) \in r_j} e_i^g[u, v] + e_i^p[u, v], \quad (13)$$

where $r_j = s_i \cap R_j$ represents the pixels uniformly sampled from R_j . These statistics are normalized into a probability distribution:

$$f_i[j] = \frac{\mathcal{H}_i[j]}{\sum_{m=1}^{64} \mathcal{H}_i[m]}. \quad (14)$$

This distribution is subsequently utilized to uniformly resample a number of pixels quantified by $n_i \cdot f_i[j]$ in each region (where n_i represents the total sample count per image), thus allocating more samples to areas exhibiting elevated loss. The adaptive ray sampling approach is delineated in Fig. 7.

Analysis of training convergence. The loss distribution of NeRF model during the training process is analyzed, and it is found that the fitting efficiency is slower in the image region with drastic color changes than in other regions. This difference shows significant detail texture and depth changes in the 3D scene area. As shown in Fig. 8, this phenomenon aligns with rendering characteristics of NeRF.

Fig. 9 demonstrates the learning rate of the radiance fields converges more quickly in regions exhibiting negligible color and depth variations. However, areas with pronounced color and depth variations, typifying complex scenes, necessitate extended durations for detailed rendering.

This analysis is further substantiated by examining the loss distribution during ENeRF [15] training. As depicted in the third row of Fig. 9, regions with sparse textures predominantly converge early in the training phase, while greater time is requisite for gradual convergence in areas with dense depth boundaries and color textures. This indicates redundancy in the existing ray sampling strategy, directly hampering the convergence efficiency of radiance fields.

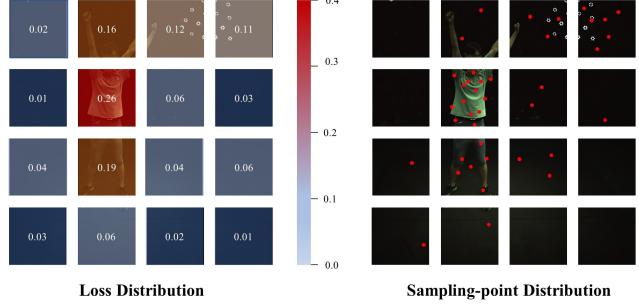


Figure 7. Sampling probability distribution of the adaptive ray sampling method.

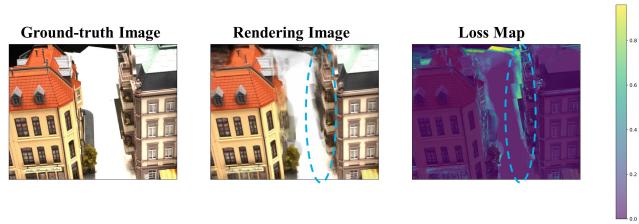


Figure 8. The training process of NeRF. Regions with sharp texture variation exhibit larger loss values in loss map. Correspondingly, these areas demonstrate poorer rendering performance in rendering image. Experiments reveal areas of more detailed texture require prolonged training time, with slower convergence rate.

Our ray sampling methods. Employing depth-guided ray sampling, we sample more rays in regions exhibiting pronounced depth variations. As portrayed in Fig. 10, greater sampling probability is conferred at depth boundaries, where heightened depth fluctuations with richer details necessitate densified ray sampling during the training stage. We additionally implement pixel-guided sampling, assigning higher probabilities when the color difference between the pixel (u, v) and its neighbors is substantial, as demonstrated in Fig. 11. This vividly exemplifies the sampling methodology under our proposed pixel-guided approach.

Clamp function. The *clamp* function will return the parameter value if the parameter is within the numerical range between the minimum and maximum values, the maximum value if the parameter is larger than the range, and the minimum value if the parameter is smaller than the range. Such as *clamp(minnumber, maxnumber, parameter)*. *clamp(4, 6, 22)* returns 6 because 22 is greater than 6 and 6 is the largest value in the range, *clamp(4, 6, 2)* returns 4 because 2 is less than 4 and 4 is the smallest value in the range, *clamp(4, 6, 5)* returns 5 because the value is in the range.

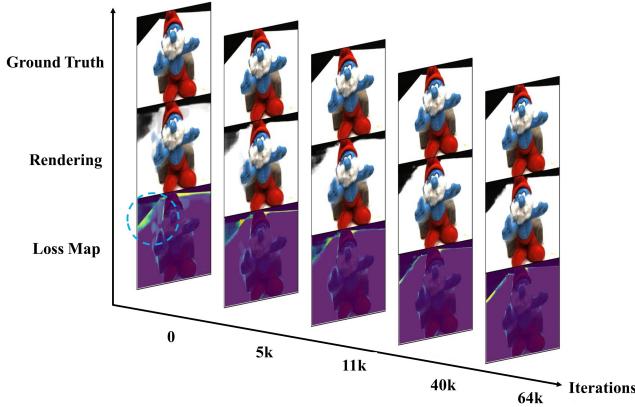


Figure 9. Training process and loss distribution. It is evident that regions containing simple details exhibit rapid convergence within 64 epochs of training. However, complex areas maintain substantial loss values necessitating further convergence, yielding relatively inferior renderings.

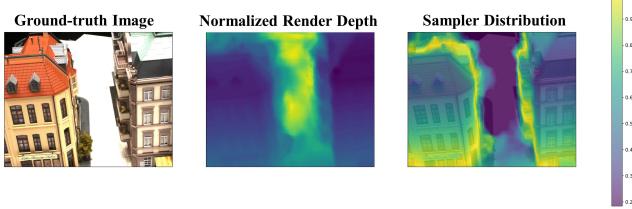


Figure 10. The distribution of depth-guided ray sampling method.

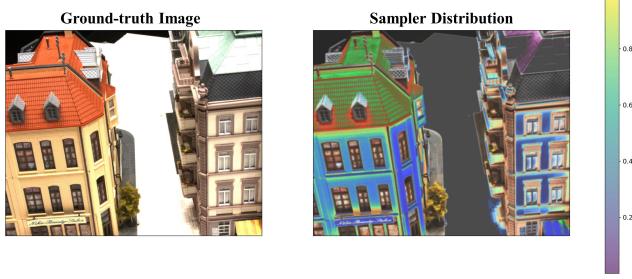


Figure 11. The distribution of pixel-guided ray sampling method.

B. More Experimental Results

Analysis of rendering quality. Our ray sampling approach samples more rays in regions with intricate details, as reflected in dramatic variations of color and depth. In this manner, the scene geometry is perceived with enhanced accuracy, and greater attention is conferred to areas with rich texture, consequently engendering marked improvements compared to the vanilla ENeRF model devoid of our method, as exemplified in Fig. 12. By comparing the generalization and fine-tuning results between our model and EN-

Table 4. Quantitative results with these models on the NeRF Synthesis and DTU datasets. The means of “ori” and “ft” are the same as Fig. 14

Methods	Training settings	DTU			NeRF Synthetic		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
IBRNet-ori	Generalization (only pre-trained)	26.04	0.917	0.191	22.44	0.874	0.195
MVSNeRF-ori		26.63	0.931	0.168	23.62	0.897	0.176
IBRNet-ft		27.38	0.937	0.166	23.77	0.915	0.181
MVSNeRF-ft		27.53	0.940	0.143	24.83	0.924	0.172
IBRNet-ori(1h)	Fine-tuning on given scene	31.35	0.956	0.131	25.62	0.939	0.111
MVSNeRF-ori(15min)		28.51	0.933	0.179	27.07	0.931	0.168
IBRNet-ft(1h)		31.90	0.958	0.122	26.82	0.944	0.102
MVSNeRF-ft(15min)		29.28	0.941	0.167	28.10	0.936	0.155

eRF [15], with all models fine-tuned for 2k iterations, our model demonstrates superior details compared to ENeRF. Therefore, our model not only expedites the network convergence, but also augments rendering quality under analogous convergence circumstances.

Visual ablation results. Visual ablation results are presented in Fig. 13. Relative to uniform sampling strategy, superior rendering is achievable solely utilizing either pixel-guided or depth-guided approaches, at accelerated training velocities. The combination of both ray sampling methodologies, with ray sampling under the joint guidance of pixel region and depth boundary, confers additional refinements in rendering quality and convergence efficiency compared to uncombined ray sampling methodology, as portrayed in Fig. 13.

Other baselines with our method. Ray sampling constitutes a ubiquitous process within NeRF modeling paradigms; therefore, our proposed ray sampling method can be directly integrated into various extant NeRF frameworks [4, 35, 42] to enhance training efficiency and rendering fidelity. To demonstrate the generalization capacity, we incorporate our approach into MVSNeRF [4] and IBRNet [35] separately and evaluate the modified baselines on NeRF Synthesis and DTU datasets relative to the original methods. Experiments adopt identical configurations as described previously.

Integrating our ray sampling methodology engenders 1.7× (IBRNet) and 2× (MVSNeRF) accelerations in convergence speed versus the original frameworks. Guidance from pixel regions and depth boundaries in sampling confers additional gains in training speed alongside superior rendering fidelity. By concentrating sampling probabilities in areas exhibiting pronounced depth and color variations given equal fine-tuning durations, qualitative outcomes after generalization and fine-tuning showcase marked improvements in these regions, as portrayed in Fig. 14. Quantitative results are tabulated in Tab. 4.

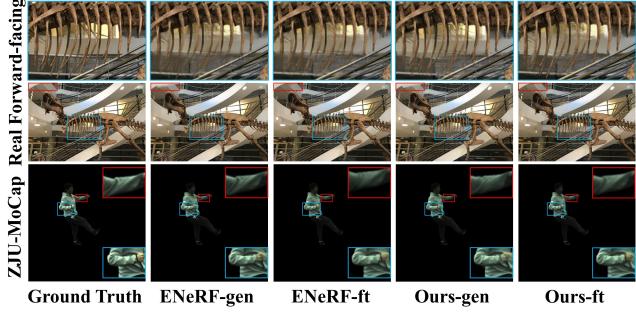


Figure 12. **Qualitative comparison highlighting improvements in regions with abundant textures.** “-gen” denotes that these pre-trained models are directly deployed on the input images without fine-tuning. “-ft” signifies that these pre-trained models undergo fine-tuning on the provided scenes for 2k iterations.

C. Per-scene breakdown

Tab. 5, 6, 7 delineate the per-scene breakdown of the quantitative results presented in the manuscript for the NeRF Synthetic, DTU and Real Forward-facing datasets, respectively. These results align with the averaged metrics depicted in the paper. The quantitative data demonstrates our methodology attaining comparable performance to other baselines. The baseline results are obtained from ENeRF [15].

Table 5. Quantitative comparison on the DTU dataset.

Scan	#1	#8	#21	#103	#114
Metric	PSNR↑				
IBRNet	25.97	27.45	20.94	27.91	27.91
MVSNeRF	26.96	27.43	21.55	29.25	27.99
ENeRF	28.85	29.05	22.53	30.51	28.86
Ours	29.43	29.62	24.35	30.79	28.92
IBRNet-ft(1h)	31.00	32.46	27.88	34.40	31.00
MVSNeRF-ft(15min)	28.05	28.88	24.87	32.23	28.47
ENeRF-ft(15min)	29.81	30.06	22.50	31.57	29.72
ENeRF-ft(1h)	30.10	30.50	22.46	31.42	29.87
Ours-ft(15min)	30.87	30.32	24.89	32.01	30.28
Ours-ft(30min)	31.11	30.88	24.95	32.25	30.97
Metric	SSIM↑				
IBRNet	0.918	0.903	0.873	0.950	0.943
MVSNeRF	0.937	0.922	0.890	0.962	0.949
ENeRF	0.958	0.955	0.916	0.968	0.961
Ours	0.963	0.956	0.931	0.969	0.960
IBRNet-ft(1h)	0.955	0.945	0.947	0.968	0.964
MVSNeRF-ft(15min)	0.934	0.900	0.922	0.964	0.945
ENeRF-ft(15min)	0.964	0.958	0.922	0.971	0.965
ENeRF-ft(1h)	0.966	0.959	0.924	0.971	0.965
Ours-ft(15min)	0.967	0.964	0.935	0.977	0.962
Ours-ft(30min)	0.968	0.967	0.935	0.980	0.964
Metric	LPIPS↓				
IBRNet	0.190	0.252	0.179	0.195	0.136
MVSNeRF	0.155	0.220	0.166	0.165	0.135
ENeRF	0.086	0.119	0.107	0.107	0.076
Ours	0.078	0.114	0.089	0.104	0.075
IBRNet-ft(1h)	0.129	0.170	0.104	0.156	0.099
MVSNeRF-ft(15min)	0.171	0.261	0.142	0.170	0.153
ENeRF-ft(15min)	0.074	0.109	0.100	0.103	0.075
ENeRF-ft(1h)	0.071	0.106	0.097	0.102	0.074
Ours-ft(15min)	0.059	0.107	0.074	0.100	0.072
Ours-ft(30min)	0.57	0.105	0.71	0.098	0.072



Figure 13. Visual ablation results on DTU dataset. “w/o our method” is similar to ENeRF [15].

Table 6. Quantitative comparison on the NeRF Synthetic dataset.

Metric	Chair	Drums	Ficus	Hotdog	Lego	Materials	Mic	Ship
Metric	PSNR↑							
IBRNet	24.20	18.63	21.59	27.70	22.01	20.91	22.10	22.36
MVSNeRF	23.35	20.71	21.98	28.44	23.18	20.05	22.62	23.35
ENeRF	28.44	24.55	23.86	34.64	24.98	24.04	26.60	26.09
Ours	29.43	25.77	25.12	34.82	25.25	24.41	27.00	26.32
Metric	SSIM↑							
IBRNet	0.888	0.836	0.881	0.923	0.874	0.872	0.927	0.794
MVSNeRF	0.876	0.886	0.898	0.962	0.902	0.893	0.923	0.886
ENeRF	0.966	0.953	0.931	0.982	0.949	0.937	0.971	0.893
Ours	0.968	0.956	0.934	0.983	0.952	0.938	0.969	0.896
Metric	LPIPS↓							
IBRNet	0.144	0.241	0.159	0.175	0.202	0.164	0.103	0.369
MVSNeRF	0.282	0.187	0.211	0.173	0.204	0.216	0.177	0.244
ENeRF	0.043	0.056	0.072	0.039	0.075	0.073	0.040	0.181
Ours	0.032	0.044	0.065	0.039	0.073	0.065	0.039	0.167
Metric	IBRNet-ft(1h)							
IBRNet	0.079	0.133	0.082	0.093	0.105	0.093	0.040	0.257
MVSNeRF	0.129	0.197	0.171	0.094	0.176	0.167	0.117	0.294
ENeRF	0.033	0.047	0.069	0.031	0.073	0.063	0.021	0.190
Ours	0.030	0.045	0.071	0.028	0.070	0.059	0.017	0.183
Metric	Ours-ft(15min)							
Ours	0.032	0.046	0.068	0.032	0.070	0.065	0.020	0.188
Ours	0.028	0.041	0.070	0.027	0.067	0.056	0.015	0.162

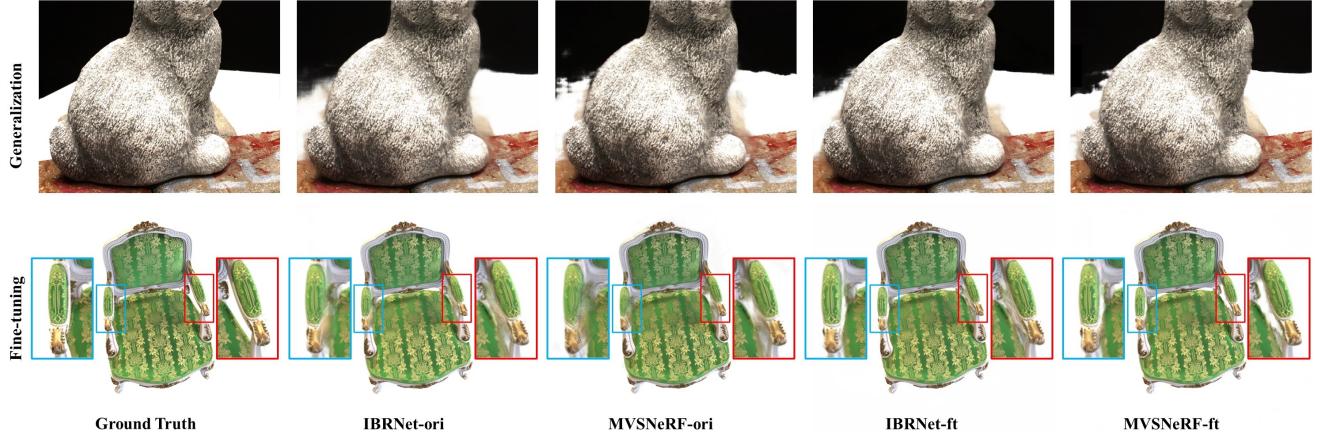


Figure 14. Qualitative results on NeRF Synthesis and DTU datasets. “IBRNet-ori” and “MVSNeRF-ori” mean that these models without our ray sampling method. “IBRNet-ft” and “MVSNeRF-ft” mean that these models are modified with our method.

Table 7. Quantitative comparison on the Real Forward-facing dataset.

	Fern	Flower	Fortress	Horns	Leaves	Orchids	Room	Trex
Metric	PSNR↑							
IBRNet	20.83	22.38	27.67	22.06	18.75	15.29	27.26	20.06
MVSNeRF	21.15	24.74	26.03	23.57	17.51	17.85	26.95	23.20
ENeRF	20.88	24.78	28.63	23.51	17.78	17.34	28.94	20.37
Ours	21.14	25.85	28.70	23.53	18.02	17.41	28.94	20.62
IBRNet-ft(1h)	22.64	26.55	30.34	25.01	22.07	19.01	31.05	22.34
MVSNeRF-ft(15min)	23.10	27.23	30.43	26.35	21.54	20.51	30.12	24.32
ENeRF-ft(15min)	21.84	27.46	29.58	24.97	20.95	19.17	29.73	23.01
ENeRF-ft(1h)	22.08	27.74	29.58	25.50	21.26	19.50	30.07	23.39
Ours-ft(15min)	22.53	27.90	29.54	25.48	21.67	19.76	30.04	23.55
Ours-ft(30min)	23.42	28.37	30.88	26.17	22.84	21.02	31.01	24.94
Metric	SSIM↑							
IBRNet	0.710	0.854	0.894	0.840	0.705	0.571	0.950	0.768
MVSNeRF	0.638	0.888	0.872	0.868	0.667	0.657	0.951	0.868
ENeRF	0.727	0.890	0.920	0.866	0.685	0.637	0.958	0.778
Ours	0.742	0.904	0.931	0.864	0.692	0.635	0.962	0.782
IBRNet-ft(1h)	0.774	0.909	0.937	0.904	0.843	0.705	0.972	0.842
MVSNeRF-ft(15min)	0.795	0.912	0.943	0.917	0.826	0.732	0.966	0.895
ENeRF-ft(15min)	0.758	0.919	0.940	0.893	0.816	0.710	0.963	0.855
ENeRF-ft(1h)	0.770	0.923	0.940	0.904	0.827	0.725	0.965	0.869
Ours-ft(15min)	0.764	0.920	0.938	0.898	0.821	0.717	0.961	0.863
Ours-ft(30min)	0.791	0.944	0.938	0.925	0.843	0.750	0.982	0.874
Metric	LPIPS↓							
IBRNet	0.349	0.224	0.196	0.285	0.292	0.413	0.161	0.314
MVSNeRF	0.238	0.196	0.208	0.237	0.313	0.274	0.172	0.184
ENeRF	0.235	0.168	0.118	0.200	0.245	0.308	0.141	0.259
Ours	0.222	0.153	0.107	0.185	0.232	0.293	0.140	0.248
IBRNet-ft(1h)	0.266	0.146	0.133	0.190	0.180	0.286	0.089	0.222
MVSNeRF-ft(15min)	0.253	0.143	0.134	0.188	0.222	0.258	0.149	0.187
ENeRF-ft(15min)	0.220	0.130	0.103	0.177	0.181	0.266	0.123	0.183
ENeRF-ft(1h)	0.197	0.121	0.101	0.155	0.168	0.247	0.113	0.169
Ours-ft(15min)	0.206	0.129	0.114	0.158	0.180	0.254	0.121	0.175
Ours-ft(30min)	0.185	0.106	0.081	0.142	0.157	0.238	0.113	0.158