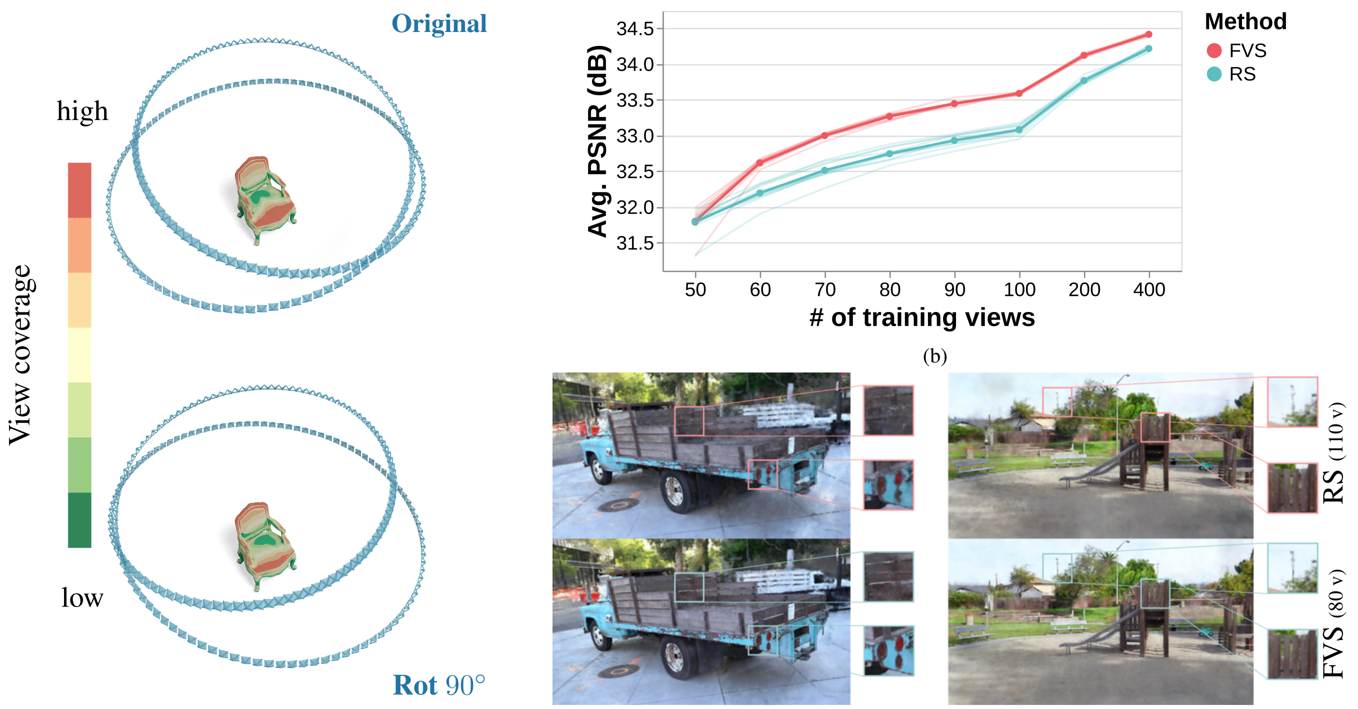
NeRF Director: Revisiting View Selection in Neural Volume Rendering

Wenhui Xiao[1,2], Rodrigo Santa Cruz[1,2], David Ahmedt-Aristizabal[1,2],
Olivier Salvado[1,2], Clinton Fookes[1], Leo Lebrat[1,2]
Queensland University of Technology[1], CSIRO Data61[2]
wenhui.xiao@hdr.qut.edu.au, {rodrigo.santacruz, leo.lebrat}@csiro.au
https://wenwhx.github.io/nerfdirector



(b)

(a)

(c)

Figure 1. Overview of NeRF Director. **(a)** Different test camera selections result in different error measurements on the target object, which ultimately can result in SOTA ranking inversion. **(b)** In a synthetic setting and for InstantNGP [25], a better view selection algorithm such as farthest view sampling (FVS) can reach 33 dB of PSNR using only 70 training views, whereas the traditional random sampling (RS) would require 150 views to achieve similar performances. **(c)** With fewer views, our view selection method outperforms traditional random view selection.

Abstract

Neural Rendering representations have significantly contributed to the field of 3D computer vision. Given their potential, considerable efforts have been invested to improve their performance. Nonetheless, the essential question of selecting training views is yet to be thoroughly investigated. This key aspect plays a vital role in achieving high-quality results and aligns with the well-known tenet of deep learning: "garbage in, garbage out". In this paper, we first illustrate the importance of view selection by demonstrating how a simple rotation of the test views within the most pervasive NeRF dataset can lead to consequential shifts in the performance rankings of state-of-the-art techniques. To address this challenge, we introduce a unified framework for view selection methods and devise a thorough benchmark to assess its impact. Significant improvements can be achieved without leveraging error or uncertainty estimation but focusing on uniform view coverage of the reconstructed object, resulting in a training-free approach. Using this technique, we show that high-quality renderings can be achieved faster by using fewer views. We conduct extensive experiments on both synthetic datasets and real-

istic data to demonstrate the effectiveness of our proposed method compared with random, conventional error-based, and uncertainty-guided view selection.

1. Introduction

Neural Radiance Fields (NeRFs) [24] can effectively learn the geometry and appearance of a 3D scene and have succeeded in novel view rendering. Ongoing research is devoted to enhancing NeRF's quality and efficiency by improving network architectures [3, 5, 10, 25, 33], and tackling the challenges from insufficient training views [12, 37, 38] or inaccurate camera poses [7, 20, 32, 35].

Another focus of recent efforts study how to sample 3D training primitives (e.g., views, rays, and points) to efficiently and effectively train a NeRF model. For instance, EfficientNeRF [11] manages to sample valid and pivotal points based on density and accumulated transmittance. Zhang et al. [39] propose to improve the learning of the radiance field without sacrificing quality significantly by selectively shooting rays in important regions. However, we argue that view selection should be the essential problem for data sampling. Views serve as *the root source of points and rays*, yet their selection remains an uncharted topic needing more comprehensive exploration.

View selection is crucial for both the evaluation and training of NeRF. On the one hand, a seamless rotation on the testing camera poses, leading to a different error measurement distribution on the reconstructed object as shown in Figure 1a, can yield an inversion in the rankings of the state-of-the-art (SOTA) NeRF methods. We underscore the importance of fair error assessment, which requires giving equal importance to every part of the scene. On the other hand, view selection also plays a pivotal role in training a NeRF model. It can be noticed from Figure 1b that in a noise-free synthetic setting, the performances of NeRF at convergence are influenced by the sampling method used to select its training views. Furthermore, Figure 1c indicates that fewer, well-chosen views can yield better novel-view rendering performance. With the ubiquity of smartphones with continually improving camera specifications, sourcing high-resolution data is no longer an issue. The prospect of a lightweight view selection algorithm is high as it contributes to improving the utility of deployable NeRF-based approaches. This motivates us to study and answer the question – *what is an effective way to select a given number of views from a large amount of training data to achieve better rendering performance?*

This paper introduces a comprehensive view selection assessment framework, NeRF Director, and explores the impact of different view selection schemes. First, we propose a robust method for generating a test split from a set of posed images without geometrical priors. It targets minimizing the sampling variance among testing views and empirically yields a more consistent evaluation. Then, we investigate typical sampling methods including random sampling (RS), and two types of heuristic sampling: farthest view sampling (FVS) and information gain-based sampling (IGS). The FVS is derived from farthest point sampling [8], and selects informative and spatially distributed training views. IGS methods allocate candidates with information gain in terms of error or uncertainty [27, 31], and each time picks the most profitable candidate. We also devise a variant of IGS methods leveraging Lloyd's algorithm [21] to mitigate the over-sampling effect of greedy view selection strategies. Finally, we conduct comprehensive experiments on both synthetic and real-world datasets.

The experimental results show that varying the choice of view selection schemes can result in a PSNR difference of up to 1.9 dB. For a fixed view budget, FVS reaches the converged quality of traditional RS significantly faster with up to 4× speedup. This indicates the critical impact of view selection on the performance of NeRF — a factor that should not be overlooked when the community continually improves the SOTA performance of NeRF models. Our detailed analysis yields interesting and important findings, listed below.

- The diversity within selected training views significantly contributes to the final reconstruction and should be given the highest initial consideration.
- IGS, being time-consuming, heavily depends on the accuracy of the adopted information. Placement based on noisy information may result in inferior results when compared to RS.
- IGS exhibits sensitivity to error and tend to cluster on complex regions of the scene, which may pose challenges for effective learning by NeRF. To enhance their performance, a relaxation step becomes necessary.

2. Related works

The importance of view sampling for traditional 3D reconstruction has been a thoroughly studied topic, and it is demonstrated that the reconstruction's quality heavily depends on viewpoint selection [1, 13, 23]. Nonetheless, this topic has not been extensively explored for NeRFs.

This paper focuses on exploring novel approaches to select training views from an extensive training pool so as to achieve optimal rendering quality. This section discusses existing work on training data sampling, which can be categorized into three types: uncertainty-guided, error-guided, and scene coverage-based methods.

Uncertainty-Guided Methods: Uncertainty estimation in neural networks plays a crucial role in various applications. It can be used for confidence assessment, quantifying information gain, and detecting outliers. Notably, in NeRF for example, NeRF-W [22] leverages uncertainty estimation to mitigate the influence of transient scene elements, while S-NeRF [30] incorporates uncertainty by sampling to encode the posterior distribution across potential radiance fields. Recent uncertainty-guided methods for view selection [6, 15, 19, 29, 31] assume that positioning the camera at the pose with the highest uncertainty will yield the highest reconstruction performance. Some methods [6, 15, 29] have integrated an uncertainty prediction module within the NeRF framework. Considering the cost of incorporating a new module into arbitrary NeRF model, the other trend of work [19, 31] avoids the modification of existing NeRF architecture. They estimate the uncertainty of reconstructed scenes based on predicted density and color. In contrast, Active-NeRF [27] considers information gain as the reduction of uncertainty in a candidate view, selecting the candidate view with the most reduction of uncertainty. However, these uncertainty-based methods directly rely on the quality of estimated uncertainty.

Error-guided methods: Rendering errors can be seen as a prior for guiding the ray sampling strategy. For instance, in the work of Zhang et al. [39] rays are directed at pixels with significant color changes and areas with higher rendered color loss, achieving faster NeRF training without compromising the competitive accuracy. A distortion-aware scheme [26] is adopted for effectively sampling rays in the 360° scene learning. In a different context, addressing the problem of NeRF model conversion via knowledge distillation, PVD-AL [9] actively selects views, rays, and points with the largest gap between the student and teacher models to enhance the student's understanding of critical knowledge. While these efforts may not directly address the question of view selection for the general setting of training a NeRF, they serve as inspiration for designing a method in an error-guided manner.

Scene coverage-based methods: Keyframe selection strategy by maximizing the coverage of the scene is crucial in Simultaneous Localization and Mapping (SLAM) [14, 36, 40] for tackling the forgetting issue. For example, NICE-SLAM [40] selects keyframes based on the overlap with existing frames, while H2-Mapping [14] focuses on maximizing the coverage of voxels in the scene. A progressive camera placement technique [17] is proposed for free-viewpoint navigation. This technique captures new views to achieve uniformity in ray coverage and angulation within a simulation system. Nevertheless, these approaches either rely on sensor-captured data like point clouds and depth images or require information within a radiance field. Differently, we focus on understanding the NeRF rendering performance when trained with different view selection algorithms.

3. Motivation – Providing a Robust Evaluation

This section analyzes the importance of view selection whilst evaluating different NeRF models and illustrates our motivation for investigating an effective view selection algorithm for NeRF. To carry out this experiment, we used the widely used NeRF Synthetic dataset [24], where cameras follow a trajectory resembling a lemniscate in the original test set. Using the Blender files provided by the authors, we generated 12 additional test sets by rotating the original cameras according to *z-axis*.

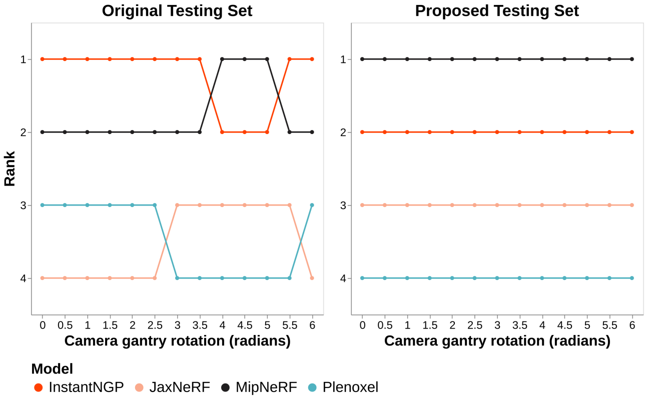


Figure 2. Ranking the rendering performance of four distinct NeRF models under various *z-axis* rotations of the test camera poses. **Left**: original test set. **Right**: proposed test set.

We compare the performance of four SOTA NeRF models—traditional NeRF [24], MipNeRF [3], Plenoxels [10] and InstantNGP [25]. Utilizing the checkpoints provided by the authors or adhering to the same training guidelines, we evaluate their performance across 13 sets of distinct camera poses. These models are ranked based on their average peak signal-to-noise ratio (PSNR) on each separate test set as visually presented in Figure 2. Notably, the ranking exhibited variations across various rotation scenarios; for instance, while InstantNGP excelled as the SOTA approach on the reference test pose, it was outperformed by MipNeRF in certain rotation scenarios.

To gain a deeper understanding of the impact of camera selection, we propose introducing the coverage density measure, supported on the mesh $\mathcal{M}$. Given a set of n-views $V = \{v_1, \ldots, v_n\}$ with associated rays $r^i_{j,k}$ for the j, k-th pixel of the i-th view, we compute the coverage measure $\mathfrak{C}$ defined by,

$$\mathfrak{C}(\mathcal{M}, V) = \frac{1}{\kappa} \sum_{l=1}^{M} \delta_{\mathbf{x}_l} \mathcal{W}(\mathcal{M}, V, \mathbf{x}_l), \tag{1}$$

$$\mathcal{W}(\mathcal{M}, V, \mathbf{x}_l) = \Big| \Big(\sum_{i=1}^{n} \sum_{\substack{1 \leq j \leq H \\ 1 \leq k \leq W}} (r^i_{j,k} \overset{1}{\cap} \mathcal{M}) \Big) \cap B^{\ell}_2(\mathbf{x}_l) \Big|,$$

with κ a normalization factor, $(\mathbf{x}_l)_{l \in 1 \cdots M}$ a uniform point-cloud discretizing $\mathcal{M}$ [18], where $B^{\ell}_2(\mathbf{x})$ is the L_2-ball of radius ℓ centered in $\mathbf{x}$ and $r \cap^1 \mathcal{M}$ denotes the first intersection between the ray r and the mesh $\mathcal{M}$. We display

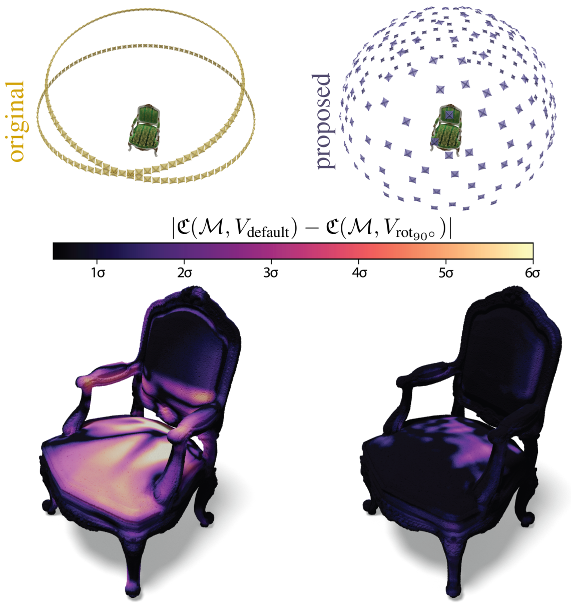

original
proposed
$|\mathfrak{C}(\mathcal{M}, V_{\text{default}}) - \mathfrak{C}(\mathcal{M}, V_{\text{rot}_{90^\circ}})|$
1σ
2σ
3σ
4σ
5σ
6σ

Figure 3. Visual comparison between original (left) and proposed (right) test set for NeRF Synthetic dataset. The top row visualizes the default (w/o. rotation) test cameras' distribution in the 3D space. The bottom displays the absolute difference of the coverage density measure between default and 90°. A lighter color indicates higher discrepancies in terms of the standard deviation σ.

$\mathfrak{C}(\mathcal{M}, V_{\text{default}})$ and $\mathfrak{C}(\mathcal{M}, V_{\text{rot}_{90^\circ}})$ in Figure 1a.

To circumvent the biased evaluation methodology, we devised a new uniform test set where all cameras are evenly distributed on a sphere centered on the reconstructed object and displayed in the Figure 3 (top). We visualize the absolute difference of $\mathfrak{C}(\mathcal{M}, V_{\text{default}})$ and $\mathfrak{C}(\mathcal{M}, V_{\text{rot}_{90^\circ}})$ for the original and the proposed test set in Figure 3 (bottom). We can observe that the proposed test set provides an even coverage of the scene, and thereby, effectively enables a robust evaluation shown in Figure 2 (right).

The effectiveness of view selection on test sets also motivates us to explore and develop an effective method for view selection on training sets that yields NeRF with improved rendering performance. We develop this idea in Section 4.

4. NeRF Director

This paper examines strategies to achieve optimal rendering quality by selecting training views from an extensive collection of images. Formally, given a set of training views $V = \{v_1, v_2, ..., v_N\}$ and a target view budget n, our objective is to select a subset of training views S that maximizes the rendering performance $\mathcal{Q}$ of NeRF. This problem can be formally defined as,

$$\underset{S \subseteq V}{\arg\max} \;\; \mathcal{Q}(S), \tag{2}$$

where $|S| = n$. To tackle this problem, this paper proposes two view selection methods — farthest view sampling (FVS) and information gain-based sampling (IGS).

Algorithm 1: Farthest View Sampling.

```
input  : $V \in v^n$ , $n \in \mathbb{N}$, $k \in \mathbb{N}$
         $\mathbf{d}(\bullet, \bullet) : v \times v \to \mathbb{R}^+$
output: $S$
$S \leftarrow \texttt{random\_sampler}(k, V)$
while $|S| < n$ do
    $v^* \leftarrow \underset{v \in V \setminus S}{\arg\max} \left[ \underset{s \in S}{\min}\ \mathbf{d}(v, s) \right]$
    $S \leftarrow S \cup \{v^*\}$
return $S$
```

4.1. Farthest View Sampling

Farthest point sampling (FPS) [8] is a classical algorithm typically applied to select a subset of points from extensive raw point cloud data. In this work, we adapt this algorithm for the view selection problem, aiming to efficiently capture features of the target scene by selecting representative views that are as different from each other as possible.

Algorithm 1 outlines our proposed FVS. Given a large set of training views V and an initially selected subset of views S, our FVS first identifies the nearest selected neighbor of each point in the remaining set $V \setminus S$. The algorithm then selects the candidate view with the maximum distance as the next addition to the subset S. The employed distance metric considers both the spatial expansion of cameras and the diversity of scene features captured by views.

Spatial expansion of cameras: We measure the spatial distance across cameras, denoted as $d_{spatial}$, by evaluating the distance between the camera centers of the candidate view c_v, and the selected view's camera center c_s. When all training views are captured from a common sphere, we adopt the great-circle distance d_{gc} as a metric. The distance is defined as,

$$\mathbf{d}_{gc}(c_v, c_s) = \arccos\left(c_v \cdot c_s\right). \tag{3}$$

In cases where training cameras are distributed throughout the scene, we employ the Euclidean distance to measure spatial separation,

$$\mathbf{d}_{euc}(c_v, c_s) = ||c_v - c_s||_2^2. \tag{4}$$

Scene diversity depicted in views: In an uncontrolled environment, the training images' perspective can be skewed away from the scene's central point, leading nearby cameras to capture entirely different visual content of the scene, e.g. two cameras at the same position orienting in two different directions. To address this challenge, we leverage the sparse 3D point cloud and 2D image correspondences computed by the underlying structure-from-motion (SfM) algorithm used to generate the poses of the training images.

Let $\mathcal{A} \in \mathbb{N}^{N \times N}$ be a symmetric matrix of pair-wise view similarities. We denote the similarity $\mathcal{A}_{ij}$ as the count of 3D points in the SfM's sparse point cloud, triangulated from 2D feature correspondences between views i and j. This measure takes into account visual content, field of view, and relative camera positioning between views. Then, using $\mathcal{A}$, the view photogrammetric distance $\mathbf{d}_{photo}$ is defined as,

$$\mathbf{d}_{photo}(v_i, v_j) = 1 - \frac{\mathcal{A}_{ij}}{\max(\mathcal{A})}. \tag{5}$$

Overall, our view distance $\mathbf{d}(\bullet, \bullet)$ is composed of two parts — $\mathbf{d}_{spatial}$ considering the spatial distance between camera centers of two views (either $\mathbf{d}_{gc}$ or $\mathbf{d}_{euc}$), and $\mathbf{d}_{photo}$ representing the difference in perception content about the scene. It can be expressed as,

$$\mathbf{d}(\bullet, \bullet) = \mathbf{d}_{spatial} + \alpha \, \mathbf{d}_{photo}, \tag{6}$$

where α is a positive hyper-parameter associated to photogrammetric distance.

4.2. Information Gain-based Sampling

While FVS selects n views from V based on a metric $\mathbf{d}$ without training a NeRF model, IGS is an incremental procedure deriving information gain from a checkpointed model to select novel views. As detailed in Algorithm 2, IGS begins by randomly sampling k views from V to establish the initial set of training views S. Subsequently, at each iteration i, it trains a NeRF model on S and evaluates the remaining views in $V \setminus S$. Employing different evaluation measures and sampling algorithms, it augments the current training set S by selecting l_i novel views from $V \setminus S$. Optionally, a density relaxation step can be performed on this augmented training set. This process continues until n new views are selected.

4.2.1 Sampling Views Maximizing Information Gain

At the core of this heuristic-based procedure lie two crucial components: the definition of information gain and the method for sampling views according to this quantity.

Target density construction: We propose to measure the error, in terms of the PSNR ranking, for every single remaining training view.

Algorithm 2: Information Gain-based Sampling.

```
input  : $V \in v^n$, $n \in \mathbb{N}$, $k \in \mathbb{N}$
         $(l_i)_{i=1 \ldots m}$    🗎 Number of added view list
output: $S$
$S \leftarrow$ random_sampler$(k, V)$
for $i \leftarrow 1$ to $m$ do
    $f_{\theta^\star} \leftarrow$ model_trainer$(S, f_\theta)$
    $\mu \leftarrow$ model_evaluator$(f_{\theta^\star}, V \setminus S)$
    $S^* \leftarrow$ $probabilitySampler(\mathbb{P}_\mu, V \setminus S, l_i)$
            🗎 Defined in Section 4.2.1.
    $S^* \leftarrow$ $relaxation(S^*, S)$
            🗎 (Optional) see Section 4.2.2.
    $S \leftarrow S \cup S^*$
return $S$
```

Sampling from the remaining view: We first introduce the Zipf view sampler. Given a set of q views $\{v_1, \ldots, v_q\}$ with associated error or uncertainty measurements $\mathbf{m} = (m_1, \ldots, m_q)$, we first define the greedy selection k as,

$$S^* \overset{k}{\sim} \mathcal{Z}\text{exp}(\mathbf{m}), \tag{7}$$

where $\overset{k}{\sim}$ describes the random selection of k elements without replacement, and $\mathcal{Z}$exp is a Pareto-Zipf law [28] with exponential weighting. The probability mass function is defined by,

$$f_i \propto \frac{e^{-\gamma \frac{\text{rank}(m_i)}{q-1}}}{K_q}, \tag{8}$$

with K_q a normalization factor, and γ a hyper-parameter controlling the sampling randomness. When $\gamma \to \infty$, this method is equivalent to the deterministic greedy sampler; conversely, when $\gamma \to 0$, this method is equivalent to RS.

As an alternative approach to exploring possible interactions between nearby error measurements within a probabilistic framework, we introduce the (M-vMF) view sampler. This sampler is built on a categorical mixture of the von Mises-Fisher (vMF) distribution [2]. This sampler assumes that each already sampled view induces a vMF distribution centered at its respective camera position on the unit sphere. The probability density function of this mixture model is given by,

$$f(x; \mathbf{v}, \mathbf{m}, \kappa) = \sum_{i=1}^{q} \alpha_i g(x; v_i, \kappa), \tag{9}$$

where $g(x; v_i, \kappa)$ is the induced vMF probability density function for view v_i, with a shared concentration hyper-parameter κ. The parameter κ regulates the dispersion of the distributions around their mode at the camera center of v_i. As $\kappa \to \infty$ the vMF density approaches a delta function located at v_i's camera center; conversely, when $\kappa \to 0$ this method is equivalent to a uniform distribution over a unit-sphere. Mathematically,

$$g(x; v_i, \kappa) = c_3(\kappa)\text{exp}(\kappa v_i^T x), \tag{10}$$

where $c_3(\kappa)$ is the standard vMF normalization term. The blending of these vMF components is controlled by the weights $\alpha = (\alpha_1, \ldots, \alpha_q)$, obtained through a softmax function applied to error measurements $\mathbf{m}$ with a temperature parameter σ. Specifically,

$$\alpha_i = \frac{\exp\left(\frac{\hat{m}_i}{\sigma}\right)}{\sum_{j=1}^{q} \exp\left(\frac{\hat{m}_j}{\sigma}\right)}, \tag{11}$$

where $\hat{m}_i = \frac{\max(\mathbf{m}) - m_i}{\max(\mathbf{m}) - \min(\mathbf{m})}$ is an inverse ranking function of the view's error, min-max normalized between zero and one.

To sample a new view location using this model, we begin by sampling from the categorical distribution controlled by α. Subsequently, we sample a 3D point from the corresponding vMF distribution. As this process does not ensure the existence of a view at the sampled location, we assign the closest view in $V \setminus S$ as the sampled view. Within this framework, regions with views of higher errors are more likely to be sampled.

4.2.2 Avoiding Oversampling Complex Object Parts

As further described in Section 4.2.1, we observe that purely greedy approaches tend to produce clusters of cameras in particular regions of the 3D space. Indeed, a scene may comprise more challenging parts to learn by neural-rendering primitive. Uncertainty or error will be more substantial for this specific scenario, resulting in Algorithm 2's proposal of novel training views clustering in this area. This over-exploitation behavior is detrimental; we empirically observe that it can lower the performance of the view proposal algorithm below that of the baseline RS. To tackle this problem, we introduce a relaxation step after the proposal of novel views via the view sampler, as described in Algorithm 2. We adapt the Lloyd-Max algorithm[21], commonly used for quantization, to uniformize the placement of the newly proposed camera. More specifically, we propose to build a uniform probability distribution whose support is defined by the convex hull of all available training cameras. After the Voronoi tessellation construction, this distribution is used to compute each cell's centroid. We apply the Lloyd iteration only to the new subset of camera S^*. More details on the implementation can be found in Supplementary.

5. Experiments

5.1. Experimental Setup

We experimented on two widely used datasets: NeRF Synthetic [24] and TanksAndTemples [16].

NeRF Synthetic: It contains 5 synthetic objects.[1] We reproduced the exact rendering settings and kept the original image resolution proposed in [24]. Each scene comprises 200 test images sampled as described in Section 3 and a pool of 300 views evenly distributed for training. We generated ten training sets to ensure reproducibility and statistical significance.

TanksAndTemples: It is a real-world dataset containing 4 scenes. Each scene comprises 251 to 313 training images and 25 to 43 test images. Due to significant bias in the testing view (see. Supplementary), we opted to combine original training and test images and resplit them while keeping the same number of test images for each scene. We followed the method described in Algorithm 1 to sample the new test set and kept the rest of the views for training.

Backbones and evaluation metrics: We conducted our experiment and analysis based on two SOTA NeRF models — InstantNGP [25] and Plenoxels [10]. As evaluation metrics, we consider peak signal-to-noise ratio (PSNR) (↑) and structural similarity index measure (SSIM) (↑) [34].

5.2. Implementation Details

We conduct a series of experiments with five repetitions using different random seeds and varying training/test sets for synthetic scenes. The process begins by randomly selecting an initial set of 5 views. Subsequently, we add 5 more views in each step, reaching a total of 30 views. The view selection process continues by adding 10 more views at each step until accumulating a total of 150 views. We train each model from scratch for each view selection choice and report novel-view rendering performance on our test set. We report results for RS and our proposed FVS and IGS, and for a comprehensive benchmark, we implement and compare two uncertainty-based IGS variants — ActiveNeRF [27] and Density-aware NeRF Ensembles [31]. These variants are built upon the InstantNGP backbone, and their implementation details are provided in Supplementary.

5.3. Evaluation on NeRF Synthetic Dataset

The results in terms of PSNR and SSIM for the InstantNGP backbone and an increasing number of training views across different view selection methods are depicted in Figure 4a and Figure 4b. It can be observed that FVS and IGS significantly outperform other view selection methods. Conversely, view selection methods from ActiveNeRF and Density-aware NeRF Ensembles exhibit inferior performance compared to RS. We attribute this gap in performance to two main factors: first, the uncertainty predicted does not consistently correlate with the expected reconstruction improvement for the candidate view; second, in certain scene areas, adding more views does not always result in decreased uncertainty, leading to oversampling,

[1] The dataset originally contains eight scenes, but with the data provided, we only managed to reproduce the original rendering quality for five of them.

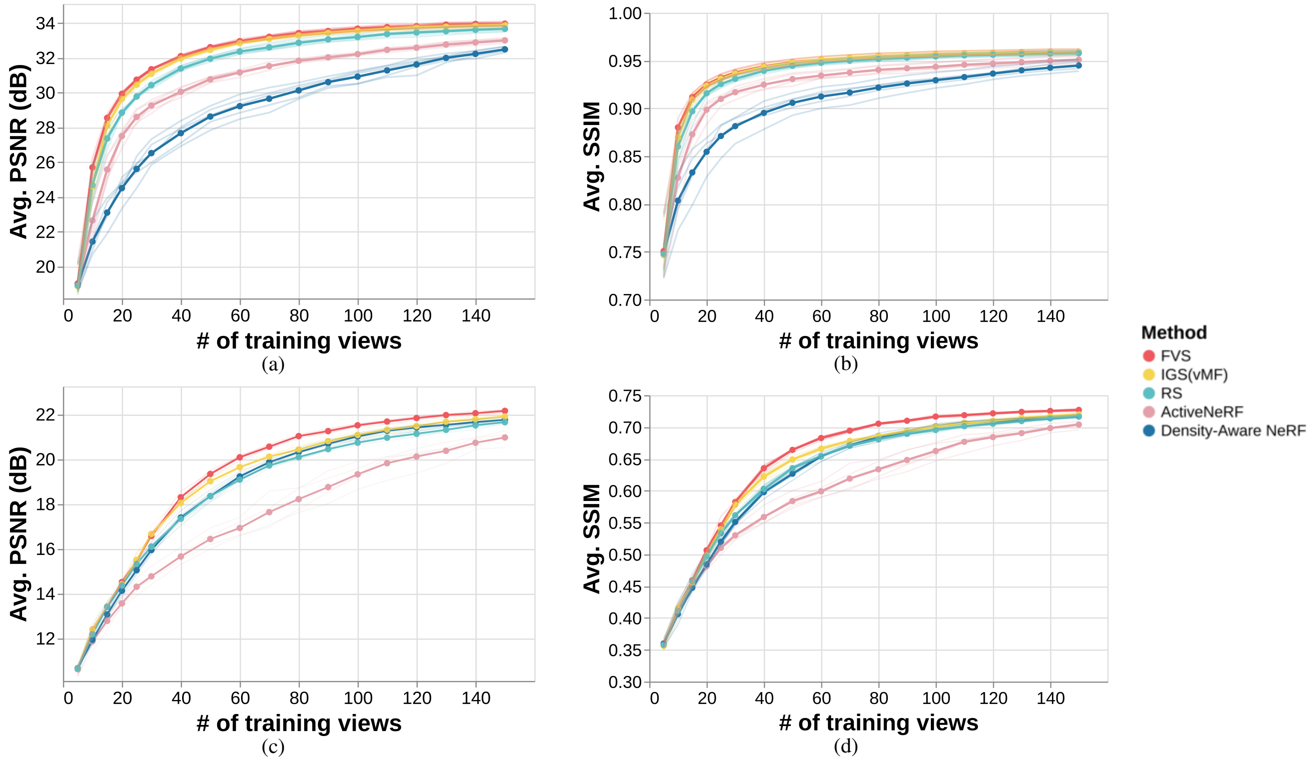


Figure 4. Quantitative comparisons of rendering quality along with the increase of used training views sampled by different view selection methods. **Top**: results on the NeRF Synthetic dataset in terms of PSNR (a) and SSIM (b). **Bottom**: results on the TanksAndTemples dataset in terms of PSNR (c) and SSIM(d). Low-opacity lines present the results for each repetition, while high-opacity lines present the average result across five repetitions.

which is not rectified by spatial regularization. Similar results are obtained for the Plenoxel backbone and are provided in Supplementary. We also provide the runtime cost analysis in Supplementary, showing that our proposed FVS can reach converged quality more efficiently than RS under the same view budget.

5.4. Evaluation on TankAndTemples Dataset

We extended our experiments to the TanksAndTemples dataset to assess the impact of different view selection methods on rendering performance for real-world data. Figure 4c and Figure 4d display the experimental results in terms of PSNR and SSIM. Notably, when the training view budget exceeds 30 views, FVS demonstrates superior performance, followed by IGS. In contrast to the results with the NeRF Synthetic dataset, view selection based on Density-aware NeRF Ensembles achieves better performance than RS in a higher view number regime (more than 60 views). This could be attributed to the improved uncertainty quantification on realistic data of the Density-aware NeRF Ensembles, where candidate views are not uniformly distributed. Intriguingly, ActiveNeRF provides the lowest performances for our test settings, and we attribute this to its exploration of a distinct training regime for NeRF (less than 30 views) and the limited pool of training views considered (100 views) as indicated in a recent study [17].

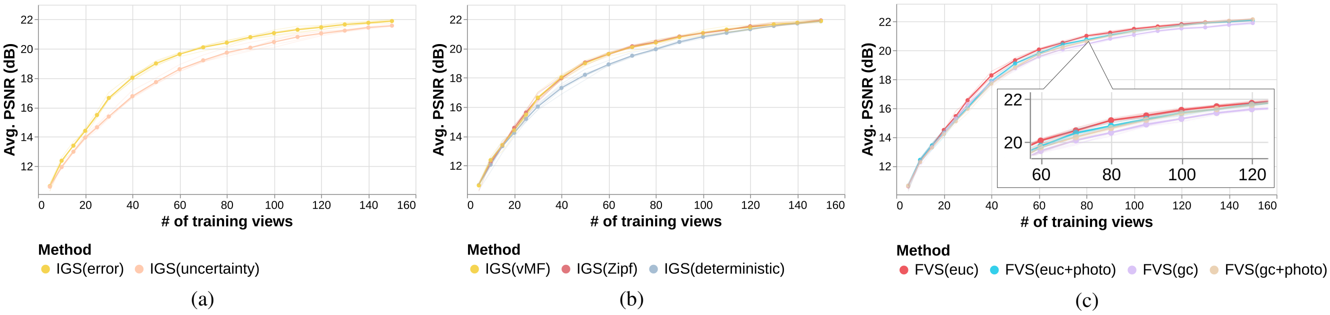
5.5. Ablation Study

This section introduces our ablation experiments on TanksAndTemples dataset to make a comprehensive analysis of the design of our proposed FVS and IGS. We use InstantNGP [25] as our backbone and PSNR as the evaluation metric. We conducted experiments following the description in Section 5.2.

5.5.1 Information Gain-based Sampling

Information type: There are two potential information types for IGS methods: error and uncertainty. We first explore the impact of different information gains on the performance of IGS. We implemented a variant of IGS based on uncertainty, which was quantified through Density-aware NeRF Ensembles [31]. Both error and uncertainty variants utilized the vMF view sampler with applied relaxation. Figure 5a provides a quantitative visualization of the comparison results. Error-based IGS consistently outperforms uncertainty-based IGS.

Sampling strategy: We further investigate the impact of different probabilistic mass functions on the IGS. We com-



(a) (b) (c)

Figure 5. Ablation studies of the information type (a) and the sampling strategy (b) in IGS, as well as different distance metrics in FVS (c) on the TanksAndTemple dataset.

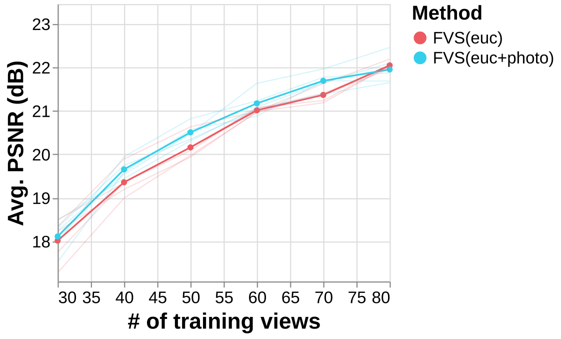


Figure 6. Comparison results between FVS ($\mathbf{d}_{euc}$) and FVS ($\mathbf{d}_{euc} + \mathbf{d}_{photo}$) in terms of PSNR on the scene *Playground*.

pared the M-vMF and the Zipf view samplers. Two Zipf view samplers were implemented with different γ settings: $\gamma = 10$ and $\gamma \to \infty$, representing the deterministic greedy. Quantitative results shown in Figure 5b indicate that view samplers, when combined with relaxation and consideration of nearby error measurements, can effectively select training views, thereby enhancing the performance of a NeRF.

5.5.2 Farthest View Sampling

We explore four combinations of spatial distance $\mathbf{d}_{spatial}$ and photogrammetric distance $\mathbf{d}_{photo}$. Specifically, we compare only $\mathbf{d}_{spatial}$ based on the great circle and the Euclidean distance, $\mathbf{d}_{gc}$ and $\mathbf{d}_{euc}$ respectively, as well as these two spatial distances separately combined with $\mathbf{d}_{photo}$. For methods using $\mathbf{d}_{gcd}$, we project all training views' camera centers onto a common sphere. The average quantitative results are presented in Figure 5c. It is evident that selecting views solely based on $\mathbf{d}_{gcd}$ could be insufficient for complex real-world datasets.

Discussion on the use of $\mathbf{d}_{photo}$: When looking at the results of each scene, we notice an interesting case in *Playground* shown in Figure 6, where cameras are distributed across the space, yet most share similar attention regions of the scene.

In such cases, the information provided by the camera center fails to indicate the area of the scene observed. Adopting $\mathbf{d}_{photo}$ is crucial for improving the scene diversity within selected training views. Despite generally offering better performances for our datasets, the use of FVS may be limited in complex indoor environments where relying solely on distances between camera centers may not adequately capture view similarity, especially in scenarios involving occlusions like indoor exploration.

6. Conclusion

We studied the role of view selection for NeRF in both training and testing. We first proposed a novel method to select test views reaching a more robust and reliable evaluation. We further proposed a novel view selection assessment framework, NeRF Director. We explored and introduced two view selection methods: farthest view sampling (FVS), considering the distance across cameras and the diversity of their content, and an improved information gain-based sampling (IGS) approach by incorporating relaxation to avoid clustering. Our experiments and analysis highlight the role of diversity in selected training views, caution against reliance on information gain-based methods with noisy information, and advocate for spatial relaxation to address sensitivity and cluster-related challenges in information gain-based methods for effective NeRF learning. We hope this will serve as a stepping stone in furthering research on this important topic.

Limitation and Future Works: Our proposed methods assume views captured by cameras with the same resolution and are designed under the object-centric setting. In our future work, we will consider unstructured configurations [4]. Also, the quality of available training views is another potential factor impacting NeRF's rendering performance. We will investigate its effect in our future work.

Acknowledgements: The work reported in this paper was partially supported by Australian Research Council (ARC) Discovery grant DP200101942.

References

[1] Sameer Agarwal, Noah Snavely, Ian Simon, Steven M. Seitz, and Richard Szeliski. Building Rome in a day. *Proceedings of the IEEE International Conference on Computer Vision*, pages 72–79, 2009. 2

[2] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, S Dhillon, and Suvrit Sra BANERJEE. Clustering on the Unit Hypersphere using von Mises-Fisher Distributions. *Journal of Machine Learning Research*, 6(46):1345–1382, 2005. 5

[3] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5835–5844. IEEE, 2021. 2, 3

[4] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2001*, pages 425–432, 2001. 8

[5] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. TensoRF: Tensorial Radiance Fields. In *Computer Vision – ECCV 2022*, pages 333–350, 2022. 2

[6] Le Chen, Weirong Chen, Rui Wang, and Marc Pollefeys. Leveraging Neural Radiance Fields for Uncertainty-Aware Visual Localization. *arXiv preprint arXiv:2310.06984*, 2023. 3

[7] 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. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8264–8273. IEEE, 2023. 2

[8] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Y. Zeevi. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, 6(9):1305–1315, 1997. 2, 4

[9] Shuangkang Fang, Yufeng Wang, Yi Yang, Weixin Xu, Heng Wang, Wenrui Ding, and Shuchang Zhou. PVD-AL: Progressive Volume Distillation with Active Learning for Efficient Conversion Between Different NeRF Architectures. *arXiv preprint arXiv:2304.04012*, 2023. 3

[10] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance Fields without Neural Networks. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5491–5500. IEEE, 2022. 2, 3, 6

[11] Tao Hu, Shu Liu, Yilun Chen, Tiancheng Shen, and Jiaya Jia. EfficientNeRF: Efficient Neural Radiance Fields. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2022-June:12892–12901, 2022. 2

[12] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis. *Proceedings of the IEEE International Conference on Computer Vision*, pages 5865–5874, 2021. 2

[13] Michal Jancosek, Alexander Shekhovtsov, and Tomas Pajdla. Scalable multi-view stereo. *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops 2009*, pages 1526–1533, 2009. 2

[14] Chenxing Jiang, Hanwen Zhang, Peize Liu, Zehuan Yu, Hui Cheng, Boyu Zhou, and Shaojie Shen. H2-Mapping: Real-time Dense Mapping Using Hierarchical Hybrid Representation. *IEEE Robotics and Automation Letters*, 8(10):6787–6794, 2023. 3

[15] Liren Jin, Xieyuanli Chen, Julius Rückin, and Marija Popović. NeU-NBV: Next Best View Planning Using Uncertainty Estimation in Image-Based Neural Rendering. *arXiv preprint arXiv:2303.01284*, 2023. 3

[16] Arno Knapitsch, Jaesik Park, Qian Yi Zhou, and Vladlen Koltun. Tanks and temples. *ACM Transactions on Graphics (TOG)*, 36(4), 2017. 6

[17] Georgios Kopanas and George Drettakis. Improving NeRF Quality by Progressive Camera Placement for Unrestricted Navigation in Complex Environments. In *International Symposium on Vision, Modeling, and Visualization (VMV)*, 2023. 3, 7

[18] Léo Lebrat, Rodrigo Santa Cruz, Clinton Fookes, and Olivier Salvado. MongeNet: Efficient Sampler for Geometric Deep Learning. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 16659–16668, 2021. 3

[19] Soomin Lee, Le Chen, Jiahao Wang, Alexander Liniger, Suryansh Kumar, and Fisher Yu. Uncertainty Guided Policy for Active Robotic 3D Reconstruction using Neural Radiance Fields. *IEEE Robotics and Automation Letters*, 7(4): 12070–12077, 2022. 3

[20] Chen Hsuan Lin, Wei Chiu Ma, Antonio Torralba, and Simon Lucey. BARF: Bundle-Adjusting Neural Radiance Fields. *Proceedings of the IEEE International Conference on Computer Vision*, pages 5721–5731, 2021. 2

[21] Stuart P. Lloyd. Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 28(2):129–137, 1982. 2, 6

[22] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7206–7215. IEEE, 2021. 2

[23] Oscar Mendez, Simon Hadfield, Nicolas Pugeault, and Richard Bowden. Taking the Scenic Route to 3D: Optimising Reconstruction from Moving Cameras. *Proceedings of the IEEE International Conference on Computer Vision*, 2017-October:4687–4695, 2017. 2

[24] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Computer Vision – ECCV 2020*, pages 405–421, Cham, 2020. Springer International Publishing. 2, 3, 6

[25] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a mul-

tiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4):102, 2022. 1, 2, 3, 6, 7

[26] Takashi Otonari, Satoshi Ikehata, and Kiyoharu Aizawa. Non-uniform Sampling Strategies for NeRF on 360{\textdegree} images. In *33rd British Machine Vision Conference 2022*, page 344, 2022. 3

[27] Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. ActiveNeRF: Learning where to See with Uncertainty Estimation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13693 LNCS:230–246, 2022. 2, 3, 6

[28] David M. W. Powers. Applications and Explanations of Zipf's Law. In *Proceedings of the Joint Conferences on New Methods in Language Processing and Computational Natural Language Learning*, pages 151–160. Association for Computational Linguistics, 1998. 5

[29] Yunlong Ran, Jing Zeng, Shibo He, Jiming Chen, Lincheng Li, Yingfeng Chen, Gimhee Lee, and Qi Ye. NeurAR: Neural Uncertainty for Autonomous 3D Reconstruction With Implicit Neural Representations. *IEEE Robotics and Automation Letters*, 8(2):1125–1132, 2023. 3

[30] Jianxiong Shen, Adria Ruiz, Antonio Agudo, and Francesc Moreno-Noguer. Stochastic Neural Radiance Fields: Quantifying Uncertainty in Implicit 3D Representations. *Proceedings - 2021 International Conference on 3D Vision, 3DV 2021*, pages 972–981, 2021. 3

[31] Niko Sünderhauf, Jad Abou-Chakra, and Dimity Miller. Density-aware NeRF Ensembles: Quantifying Predictive Uncertainty in Neural Radiance Fields. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9370–9376. IEEE, 2023. 2, 3, 6, 7

[32] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. SPARF: Neural Radiance Fields from Sparse and Noisy Poses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4190–4200, 2023. 2

[33] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022. 2

[34] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 6

[35] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. NeRF–: Neural Radiance Fields Without Known Camera Parameters. *arXiv preprint arXiv:2102.07064*, 2021. 2

[36] Xingrui Yang, Hai Li, Hongjia Zhai, Yuhang Ming, Yuqian Liu, and Guofeng Zhang. Vox-Fusion: Dense Tracking and Mapping with Voxel-based Neural Implicit Representation. *Proceedings - 2022 IEEE International Symposium on Mixed and Augmented Reality, ISMAR 2022*, pages 499–507, 2022. 3

[37] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural Radiance Fields from One or Few Images. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 4576–4585, 2020. 2

[38] Jian Zhang, Yuanqing Zhang, Huan Fu, Xiaowei Zhou, Bowen Cai, Jinchi Huang, Rongfei Jia, Binqiang Zhao, and Xing Tang. Ray Priors through Reprojection: Improving Neural Radiance Fields for Novel View Extrapolation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18355–18365. IEEE, 2022. 2

[39] Wenyuan Zhang, Ruofan Xing, Yunfan Zeng, Yu Shen Liu, Kanle Shi, and Zhizhong Han. Fast Learning Radiance Fields by Shooting Much Fewer Rays. *IEEE Transactions on Image Processing*, 32:2703–2718, 2023. 2, 3

[40] 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, 2022. 3