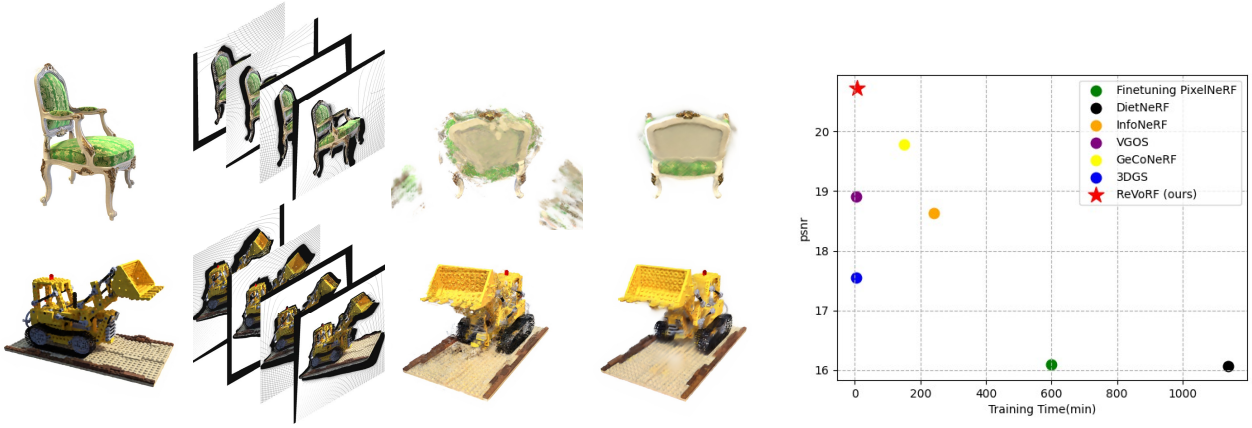
Learning with Unreliability: Fast Few-shot Voxel Radiance Fields with Relative Geometric Consistency

Yingjie Xu[1,2*] Bangzhen Liu[2*] Hao Tang[1,3] Bailin Deng[4] Shengfeng He[1†]
[1]Singapore Management University [2]South China University of Technology
[3]Nanjing University of Science and Technology [4]Cardiff University



(a) Input Views (b) Warped Images (c) w/o Unreliability (d) Ours (e) Comparisons on the Realistic Synthetic 360° dataset [25]

Figure 1. We present ReVoRF, a voxel-based framework designed to capitalize on the unreliability inherent in warped novel views. (b) demonstrates the warping outcomes, where black holes signify unmatched pixels from the original view. (c) illustrates the results of training when these holes are masked out, which unfortunately results in ambiguous geometric structures. In contrast, (d) showcases our approach's ability to maintain correct geometric consistency. ReVoRF achieves this by leveraging relational depth prior knowledge within these unreliable hole regions. Our approach demonstrates the best reconstruction quality while being one of the fastest few-shot approaches in (e).

Abstract

We propose a voxel-based optimization framework, ReVoRF, for few-shot radiance fields that strategically address the unreliability in pseudo novel view synthesis. Our method pivots on the insight that relative depth relationships within neighboring regions are more reliable than the absolute color values in disoccluded areas. Consequently, we devise a bilateral geometric consistency loss that carefully navigates the trade-off between color fidelity and geometric accuracy in the context of depth consistency for uncertain regions. Moreover, we present a reliability-guided learning strategy to discern and utilize the variable quality across synthesized views, complemented by a reliability-aware voxel smoothing algorithm that smoothens the transition between reliable and unreliable data patches. Our approach allows for a more nuanced use of all available data, promoting enhanced learning from regions previously considered unsuitable for high-quality reconstruction. Extensive experiments across diverse datasets reveal that our approach attains significant gains in efficiency and accuracy, delivering rendering speeds of 3 FPS, 7 mins to train a 360° scene, and a 5% improvement in PSNR over existing few-shot methods. Code is available at https://github.com/HKCLynn/ReVoRF.

1. Introduction

Neural Radiance Fields (NeRF) have revolutionized the fields of novel view synthesis and 3D reconstruction by leveraging an implicit function optimized from a collection of 2D images [2, 13, 15, 25, 28]. Despite their remarkable rendering capabilities, NeRFs are hampered by the substantial cost and time required to gather dense image sets for a given scene [6, 42, 48]. This challenge has spurred the development of Few-shot NeRF, an emerging domain focused on reconstructing 3D scenes with minimal image data [6, 8, 12, 19].

The performance of NeRF in accurately reconstructing geometry and texture diminishes when faced with sparse observations, as it tends to overfit the limited views available [8, 50]. To address this issue, there has been a push in recent studies to fortify NeRF with additional priors [47, 48], including semantic relations [14], depth cues [50], and entropy constraints [18]. These enhancements strive to extract maximum information from limited data. However, the reconstructions are inherently limited by the insufficient

*The first two authors contributed equally.
†Corresponding author (*shengfenghe@smu.edu.sg*).

coherence of the sparse views provided.

Recent research has explored overcoming the challenges posed by very limited observations through pseudo-view synthesis [3, 19, 54]. By using known camera poses and coarse depth estimates, these methods generate warped images from sparse viewpoints to enhance cross-view consistency. However, as shown in Fig. 1b, these generated images often include noisy areas with artifacts, which, if used for learning, can lead to inconsistent training signals and compromise scene integrity. To address this, Kwak et al. [19] implement self-occlusion aware masking to exclude unreliable regions. While this selective masking successfully filters out areas of uncertainty, it also introduces voids, presenting a conundrum: refining these images can bring in inconsistent noise and floaters, as illustrated in Fig. 1c, yet the limited number of usable samples necessitates using all available pseudo supervision for quality reconstruction.

In light of the issues identified with unreliable warped areas, our paper proposes a novel method for fully exploiting these uncertain regions to achieve multi-view consistency learning. The rationale behind our method is that, although absolute supervision is not reliable in those disoccluded regions, we observe that they maintain consistent relative depth relationships. The unreliability of certain regions in warped images can still bear geometric resemblances to their original view counterparts. We find that local depth information within these images can indicate geometric disparities, offering a self-supervised signal that aids in discerning the geometry of regions lacking precise textural information. While reliable regions offer more accurate supervision, our approach seeks to fully utilize all the information present, exploiting depth cues in coarsely warped images to inform the learning process across both reliable and traditionally discarded unreliable areas.

Drawing from the insights above, we propose a novel voxel-based optimization framework, ReVoRF, tailored for fast and multi-view consistent reconstruction of few-shot radiance fields, which incorporates the relative depth priors from several aspects. The objective of this work is to concurrently explore information from both dependable and less reliable regions within the warped novel view images. In the first step, we randomly warp the sparse images onto a series of novel views, subsequently delineating reliable and unreliable regions based on the pixel-wise correlation between the input and novel view. We then introduce a bilateral geometric consistency loss to enable self-training on novel synthesized images. This loss encompasses a reconstruction term in a bilateral manner, including a color and density regularization term for reliable regions and a relative depth consistency term for unreliable regions, respectively. While the former term aims at explicitly learning the geometric context of the reliable regions, the relative depth regularization is applied for implicitly exploring the geometric consistency guided by relative depth. Moreover, we integrate unreliability into our voxelization of scene features: 1) a reliability-guided learning strategy that dynamically adjusts learning priorities towards more reliable regions; 2) a reliability-aware voxel smoothing procedure that preserves structural integrity in reliable zones and mitigates inconsistencies in less reliable ones, ensuring a balanced and coherent scene reconstruction. As illustrated in Fig. 1e, assisted by both bilateral geometric consistency loss and reliability-aware regularization, our method is the second fastest while achieving the best reconstruction fidelity, with a large margin over the others.

Our contributions can be summarized as follows:

- We present the first attempt to explore pseudo-views unreliability within few-shot radiance fields, presenting the first framework to incorporate these areas for enhanced multi-view consistency learning with a bilateral geometric consistency loss.
- We introduce a reliability-guided learning strategy and voxelization smoothing procedure that tailors the learning process to the reliability of data, thus optimizing the training emphasis for improved reconstruction quality in few-shot radiance fields.
- We demonstrate superior performance of our approach against existing state-of-the-art few-shot methods in efficiency and accuracy, through extensive experiments on both synthetic and real-world datasets.

2. Related Work

Neural Radiance Fields (NeRF). NeRF [2, 5, 13, 23, 25] have emerged as a significant advancement in 3D reconstruction and novel view synthesis. These methods employ an implicit function to represent a 3D scene, enabling the extraction of detailed geometric and textural information from a set of multi-view images. Subsequent researchers have broadened the scope of NeRF applications, including generative modeling [11, 41, 53], video synthesis [9, 21, 35], and scene editing [22, 49]. Despite the impressive rendering quality, the training of vanilla NeRF often spans several days for a single scene reconstruction. Recent advancements [10, 28, 38, 39] have endeavored to mitigate this computational burden. Approaches such as DVGO [4, 38, 39] employ dense voxel grids in conjunction with shallow multilayer perceptrons to expedite the reconstruction process. Similarly, Plenoxels [10] utilizes sparse voxel grids, and Instant-NGP [28] employs a multi-resolution hash table to delineate the radiance field more efficiently. Diverse from these methods, which typically require dense inputs, we aim to address the challenge of achieving fast and high-fidelity scene reconstruction in the case where only a few observed views are available.

Few-Shot NeRFs. Recent advances [5, 14, 18, 48, 50, 52] have sought to reduce the dependency on densely collected

data for scene reconstruction, leveraging sparse inputs and scene priors. Notably, PixelNeRF [48] and StereoRF [7] utilize local semantic relationships across multiple scenes, while MVSNeRF [5] incorporates cost volume to enhance performance. These methodologies, however, require pretraining on numerous scenes to acquire necessary scene priors. Further developments [14,18,32,40,47] have introduced various regularization techniques to maximize the utility of sparse input views. InfoNeRF [18] enhances ray adjacency consistency through entropy regularization. DietNeRF [14] facilitates cross-view semantic consistency by harnessing the semantic space of the pretrained CLIP [32], while DiffusioNeRF [45] explores the diffusion prior of pretrained diffusion models. Additionally, FreeNeRF [47] applies frequency regularization, and VGOS [40] introduces voxel regularization to optimize both feature representation and density. Another group of research focuses on augmenting sparse inputs with synthetically generated views. RapNeRF [50] utilizes geometric re-projection for novel view extrapolation, while VmNeRF [3] employs depth maps for view-morphing. GeCoNeRF [19] aims to refine geometric consistency by separating reliable regions from warped images and discarding unreliable areas prone to self-occlusion. Our work diverges from these approaches by considering the inherent information of both reliable and unreliable regions of the novel view images, facilitating cross-view geometric consistency.

Unreliability/Uncertainty Modeling. In the rapid development of NeRF, the incorporation of uncertainty modeling has become crucial for achieving robustness in 3D reconstruction from sparse views. Previous efforts have employed diverse strategies, including Bayesian approaches [16, 29] and evidential neural networks [1,34], to quantify uncertainty in neural networks. In the context of NeRF, uncertainty has been harnessed to enhance rendering and guide input capture. Some methods assume Gaussian noise in RGB space for pixel-wise uncertainty [23, 31], employ volumetric entropy for scene geometry [20, 46], or adopt variational inference or Latent Variable Modeling for radiance field uncertainty as seen in S-NeRF [37] and CF-NeRF [36]. However, these approaches have not comprehensively addressed uncertainty quantification in unseen regions. Our approach differs from these methods by not only capturing uncertainty in the geometry and appearance of visible areas but also explicitly accounting for unseen spaces, including occluded points, which previous methods have not considered. This distinction allows for a more nuanced and accurate reconstruction, promising to elevate the fidelity of few-shot NeRF models.

3. Methodology

3.1. Preliminaries

NeRF [25] represents a scene as a continuously differentiable function f via a Multi-Layer Perceptron (MLP). Given a 3D position $\mathbf{x} \in \mathbb{R}^3$ and the associated 2D viewing directions $\mathbf{d} \in \mathbb{R}^2$, NeRF maps them into a volume density $\sigma \in \mathbb{R}$ and an RGB value $\mathbf{c} \in \mathbb{R}^3$, such that: $(\mathbf{c}, \sigma) = f(\gamma(\mathbf{x}), \gamma(\mathbf{d}))$, where the γ is a positional encoding that projects $\mathbf{x}$ and $\mathbf{d}$ into a higher dimensional feature space [43]. With a ray parameterized as $\mathbf{r}_p(t) = \mathbf{o} + t\mathbf{d}_p$ cast from the camera's optical center $\mathbf{o}$ along direction $\mathbf{d}_p$, the expected color $\hat{C}(\mathbf{r}_p)$ of pixel p is rendered as follows:

$$\hat{C}(\mathbf{r}_p) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}_p(t))\mathbf{c}(\mathbf{r}_p(t), \mathbf{d}_p)\, dt, \tag{1}$$

where t_n and t_f are the near and far bounds of the ray for sampling, and $T(t) = \exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))\, ds\right)$ denotes the cumulative transparency along the ray from t_n to t. Therefore, the NeRF can be optimized by a reconstruction loss between the rendered color $\hat{C}(\mathbf{r})$ and the real color $C(\mathbf{r})$:

$$L_{rgb} = \sum_{\mathbf{r}\in\mathcal{R}} \|\hat{C}(\mathbf{r}) - C(\mathbf{r})\|^2, \tag{2}$$

where $\mathcal{R}$ denotes the set of training rays.

3.2. Overview

We propose ReVoRF, a novel voxel-based optimization framework tailored for fast and multi-view consistent scene reconstruction from sparse input views. Our key idea is to incorporate the unreliability information for fully exploring the warped novel view images. By treating depth priors as the unreliability metrics, ReVoRF facilitates the reconstruction of the few-shot radiance field from several aspects, including the multi-view consistency learning (Sec. 3.3) and the regularization of voxel features (Sec. 3.4). The overall pipeline of ReVoRF is displayed in Fig. 2.

3.3. Unreliability for Multi-view Consistency

In this section, we explore the potential of unreliability in facilitating multi-view geometric consistency via the proposed bilateral geometric consistency loss.

Novel View Warping. Starting from a set of sparse input images I_r^i for $i \in 1, ..., N_r$, where i represents the view number and N_r is a small number, *e.g.*, $N_r = 3$ or $N_r = 4$, we propose to synthesize novel view images $I_{s\leftarrow r}^{i,j}$ through a fast and flexible warping process on several novel views $j \in 1, ..., N_s$. To preserve the cross-view consistency, the warping is guided by a coarse depth map D_r^i, where the depth value on each pixel p is accumulated by the density along each ray r_p omitted from the camera: $D_r^i(p) = \int_{t_f}^{t_n} T(t)\sigma(r_p(t))t\, dt$. Subsequently, we obtain the warped image $I_{s\leftarrow r}^{i,j}$ through a cross-view transformation $H_{s\leftarrow r}$, which deforms each pixel p_r of I_r^i to its correspond-

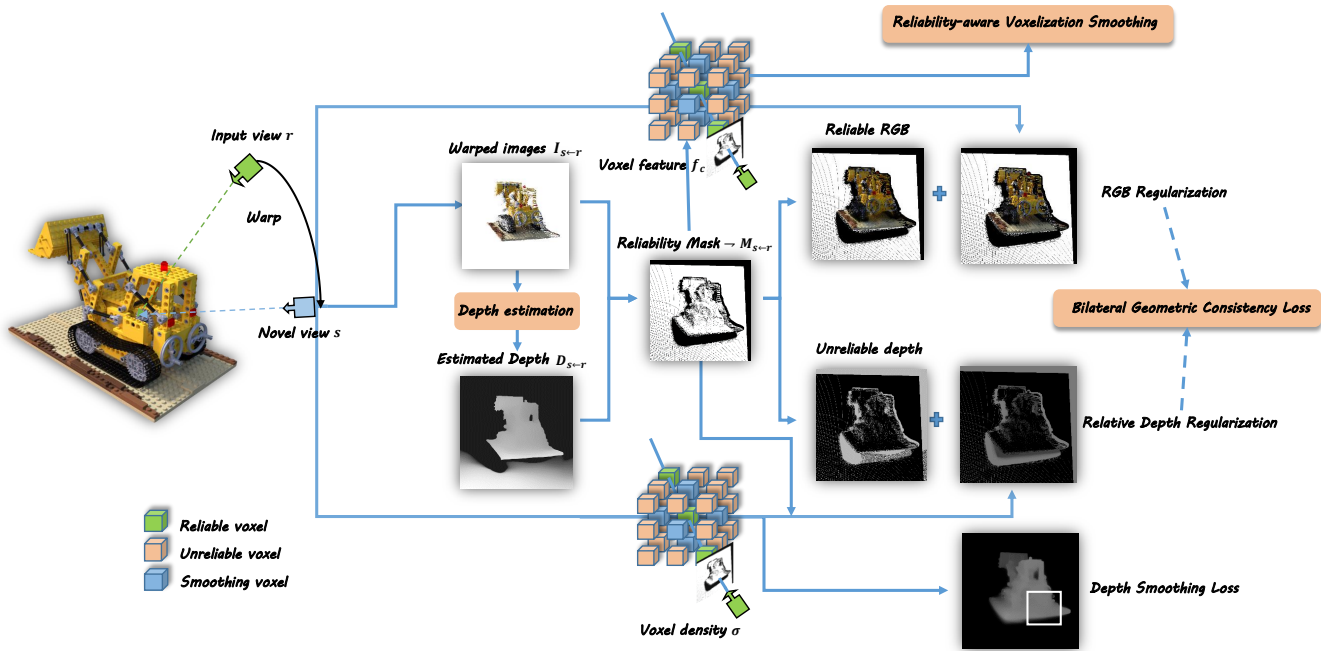


Figure 2. Overview of our proposed ReVoRF. Specifically, we first warp the sparse images onto several novel views and determine both the dependable and unreliable regions. Based on the dependability of each image region, we introduce a bilateral geometric consistency loss for multi-view consistent learning, which is composed of a color and density regularization term for reliable regions and a relative depth consistency term for unreliable regions. These two terms are responsible for explicitly learning the reliable geometric contents and implicitly exploring the geometric consistency via the guidance of relative depth, respectively. For voxel feature regularization, we integrate the unreliability through a reliability-guided learning strategy and a reliability-aware voxel smoothing procedure. By prioritizing the learning of more reliable regions and mitigating the inconsistencies in less reliable ones, ReVoRF ensures a more balanced and coherent reconstruction.

ing position p_s on the target view:

$$\begin{aligned} p_{s\leftarrow r} &= H_{s\leftarrow r}(p_r) \\ &= f_{s\leftarrow w}(f_{w\leftarrow r}(p_r)), \end{aligned} \tag{3}$$

$f_{w\leftarrow r}$ is a mapping matrix from the pixel coordinate of I_r^i to the world coordinate and $f_{s\leftarrow w}$ is the inverse operation to the coordinate of $I_{s\leftarrow r}^{i,j}$, such that:

$$f_{w\leftarrow r}(p_r) = D_r^i(p_r)T_r^{-1}K_r^{-1}p_r, \tag{4}$$

$$f_{s\leftarrow w}(p_w) = K_s T_s(p_w), \tag{5}$$

where K and T represent the camera's intrinsic and extrinsic parameter matrices in their corresponding views, respectively. Since the warping function is not surjective, voids may occur in $I_{s\leftarrow r}^{i,j}$. We empirically obtain a binary mask M_{warp}, where the pixels of void areas are set as 1 to identify the initial unreliable regions. To further refine the mask, we employ the cross-view pixel correspondences within the world coordinate of I_r^i and $I_{s\leftarrow r}^{i,j}$. To achieve this, we obtain the pseudo depth D_s^j of view j by rendering from the radiance fields. Following Eq. 4, we map the pixel of I_r^i and $I_{s\leftarrow r}^{i,j}$ into the same coordinate and obtain the correlation map M_{cor} by comparing the distance between each pixel pair $(p_r, p_{s\leftarrow r})$:

$$M_{cor}(p_{s\leftarrow r}) = \begin{cases} 1, & \|f_{w\leftarrow r}(p_r) - f_{w\leftarrow s}(p_{s\leftarrow r})\|_2 > \epsilon. \\ 0, & \text{otherwise.} \end{cases} \tag{6}$$

In this way, the final unreliability mask can be calculated as follows:

$$M_{s\leftarrow r} = M_{cor} \cup M_{warp}. \tag{7}$$

Bilateral Geometric Consistency Loss. According to the obtained unreliability mask $M_{s\leftarrow r}$, we categorize the warped novel view image into reliable regions R_{rel} and unreliable regions R_{unr}. Subsequently, we propose a bilateral geometric consistency loss to facilitate the self-training. Since the contents within R_{rel} are considered reliable, we explicitly constrain the appearance of rendered image I_s^j on view j via a reconstruction loss defined as:

$$L_{rel} = \sum_{p\in\mathcal{R}_{rel}} \|I_{s\leftarrow r}^{i,j}(p) - I_s^j(p)\|^2. \tag{8}$$

For R_{unr}, we propose to leverage the relative depth prior of the warped $I_{s\leftarrow r}^{i,j}$ to improve the geometric consistency. Specifically, we extract the depth map $D_{s\leftarrow r}$ with a powerful pretrained depth estimation model DPT [33]. By analyzing the semantics of the surrounding context, we could inpaint

the voids occurring in $D_{s\leftarrow r}$. Subsequently, a relative depth regularization loss [24] is introduced to constrain the geometric consistency on the rendered depth D_s, which is defined as:

$$L_{unr} = \sum_{p\in R_{unr}} \sum_{\hat{p}\in N(D^p_{s\leftarrow r})} h(p,\hat{p}), \tag{9}$$

where

$$h(p,\hat{p}) = \begin{cases} \max(|D^{\hat{p}}_s - D^p_s| - m, 0) & \text{if } (D^{\hat{p}}_{s\leftarrow r} - D^p_{s\leftarrow r}) \\ & \quad \times (D^{\hat{p}}_s - D^p_s) < 0, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Here p denotes any pixel within the unreliable region R_{unr}; $\hat{p}$ represents each pixel within a neighborhood $N(D^p_{s\leftarrow r})$ of p, where $N(D^p_{s\leftarrow r})$ is obtained by calculating pixels that have close depth values with $D^p_{s\leftarrow r}$ in the warped depth map; $D^p_s, D^{\hat{p}}_s$ and $D^p_{s\leftarrow r}, D^{\hat{p}}_{s\leftarrow r}$ denote the depth values of p and $\hat{p}$ within the rendered depth map D_s and estimated depth map $D_{s\leftarrow r}$, respectively. The function $h(p,\hat{p})$ penalizes inconsistent relative ordering between the depth values of p and $\hat{p}$ in the two depth maps. Specifically, if $D^p_s < D^{\hat{p}}_s$ but $D^p_{s\leftarrow r} > D^{\hat{p}}_{s\leftarrow r}$, or $D^p_s > D^{\hat{p}}_s$ but $D^p_{s\leftarrow r} < D^{\hat{p}}_{s\leftarrow r}$, then $h(p,\hat{p})$ penalizes the depth difference $|D^{\hat{p}}_s - D^p_s|$ beyond a threshold m, to prevent the depth values from shifting dramatically.

Our bilateral geometric consistency loss is then defined as a weighted sum of L_{rel} and L_{unr}:

$$L_{bgc} = \lambda_{rel} L_{rel} + \lambda_{unr} L_{unr}. \tag{11}$$

This loss enables us to thoroughly explore the information from both reliable and unreliable regions, facilitating the learning of cross-view consistency. Note that we also apply Eq. 9 on input views I_r as a depth regularization between the rendered depth and D_r for fully exploring the depth prior.

3.4. Unreliability for Voxel Feature Regularization

In this section, we incorporate the unreliability on regularizing the feature of each position of the voxel grid. With a proper design of reliability-aware voxel smoothing and reliability-aware learning adjustment, we further improve the quality of the rendered image, avoiding suboptimal scene reconstruction.

Reliability-aware Voxel Smoothing. Employing voxelized feature representations [38, 39] can significantly improve the training and rendering speed of NeRF, by storing the RGB features f_c and density σ in a voxel grids. To facilitate the learning of voxel representation, DVGO [38, 39] propose a differentiable voxel smoothing loss, which regularizes the difference of f_c and σ between a given voxel $\mathbf{v}$ with its six adjacent points V as follows:

$$L(\mathbf{v}) = \sum_{\hat{\mathbf{v}}\in V} \Delta_{f_c}(\mathbf{v},\hat{\mathbf{v}}) + \Delta_\sigma(\mathbf{v},\hat{\mathbf{v}}), \tag{12}$$

where $\Delta(\cdot,\cdot)$ denotes an error metric for the difference (*e.g.*, L_1, L_2, or Huber loss).

However, under the few-shot scenario, learning on the sparse input views can easily overfit the view-specific image, which could lead to a degenerated voxel grid that may contain fluctuant features. To address this issue, we propose to regularize the voxel features for balanced and smooth learning. By taking the unreliability of each synthesized novel view image into consideration, we mitigate the influence of unreliable regions while promoting learning in more reliable areas. We design a reliability-aware smooth factor $\rho(\mathbf{v})$ for each voxel in the grid. Specifically, given a warped image, we cast O rays $\mathbf{r}$ through each pixel of the reliable regions R_{rel}. For each voxel $\mathbf{v}$, we obtain a reliability score by accumulating the number of rays that pass this voxel, denoted by $S(\mathbf{v})$. The maximum number of times being passed by rays is denoted as $S(\mathbf{v})_{max}$. Then the reliability-aware smooth factor is defined as $\rho(v) = \frac{S(\mathbf{v})}{S(\mathbf{v})_{max}}$. Finally, we formulate the reliability-aware voxel smoothing losses on f_c and σ as:

$$\begin{aligned} L_{f_c} &= \sum_{\mathbf{v}} \sum_{\hat{\mathbf{v}}\in V} (1 + e^{-\rho(\mathbf{v})}) \Delta_{f_c}(\mathbf{v},\hat{\mathbf{v}}), \\ L_\sigma &= \sum_{\mathbf{v}} \sum_{\hat{\mathbf{v}}\in V} (1 + e^{-\rho(\mathbf{v})}) \Delta_\sigma(\mathbf{v},\hat{\mathbf{v}}). \end{aligned} \tag{13}$$

The final regularization loss is defined as follows:

$$L_{rs} = \lambda_f L_{f_c} + \lambda_d L_\sigma. \tag{14}$$

In this case, the unreliable regions will have smoother supervision during training, mitigating the inconsistency caused by potential overfitting.

Reliability-guided Learning Adjustment. To further facilitate the learning of geometric and appearance information, we apply a reliability-guided learning strategy to dynamically prioritize the learning towards the reliable zones, while eliminating the false supervision of unreliable regions at the beginning of training. Concretely, we adjust the importance of each voxel $\mathbf{v}$ with a reliability weight $\mathbf{w_v} = 1 + \rho(\mathbf{v})$, to control the gradients of each voxel during the back-propagation.

3.5. Optimization of ReVoRF

To avoid dramatic variation of the unreliable depth, we further adopt a depth smoothness loss function [30] as an extra regularization for better relative depth supervision:

$$\begin{aligned} L_{ds} = \frac{1}{|R|} \sum_{r\in R} \sum_{(x,y)\in D} & \|d(x,y) - d(x,y+1)\|_2^2 \\ & + \|d(x,y) - d(x+1,y)\|_2^2, \end{aligned} \tag{15}$$

Diet-NeRF [14] InfoNeRF [18] VGOS [40] ReVoRF (Ours) GT

Figure 3. 4-views reconstructions on Realistic Synthetic 360° [27]. ReVoRF enables more consistent reconstruction with detailed appearance.

	Realistic Synthetic 360° dataset			
Methods	PSNR↑	SSIM↑	LPIPS↓	Training Time↓
NeRF [25]	15.93	0.780	0.320	2 hrs
PixelNeRF [48]	16.09	0.738	0.390	3-4 days* + 10 hrs
DietNeRF [14]	16.06	0.793	0.306	19 hrs
3DGS [17]	17.55	0.701	0.250	**3 mins**
InfoNeRF [18]	18.62	0.811	0.230	4 hrs
VGOS [40]	18.91	0.825	0.205	**3 mins**
GeCoNeRF [19]	<u>19.78</u>	**0.880**	0.185	$>$ 2 hrs
Ours	**20.72**	<u>0.848</u>	**0.179**	<u>7 mins</u>

Table 1. Quantitative comparison for 4-views setting in the Realistic Synthetic 360° dataset [25]. The best and the second-best results are highlighted in **bold** and <u>underlined</u>, respectively. (*) denotes the time cost of pre-training.

where R represents the set of rays emanating from the sampled views, D refers to the depth patch that is centered around r, and $d(x, y)$ is of the depth value in position (x, y).

The final objective of ReVoRF is formulated as:

$$L_{total} = L_{rgb} + L_{bgc} + L_{rs} + \lambda_{ds} L_{ds}. \tag{16}$$

4. Experiments

In this section, we demonstrate the superiority of the proposed ReVoRF through extensive experiments. The details of experiment settings are discussed in Sec. 4.1. Analysis on comparison experiments and ablation study are performed in Sec. 4.2 and Sec. 4.3, respectively.

4.1. Experiment Settings

Datasets. The experiments are conducted on inward-facing scenes from the Realistic Synthetic 360° dataset [27] and forward-facing scenes from the LLFF dataset [26].

Realistic Synthetic 360° comprises path-traced images from 8 synthetic scenes, which are characterized by their complex geometry and realistic rendering of non-Lambertian materials. Each scene is represented by 400 images, rendered by inward-facing virtual cameras positioned at varying viewpoints. We adhere to the protocol established by InfoNeRF [18] and randomly select 4 views from 100 training images as sparse inputs. The model's performance is then evaluated on a set of 200 testing images.

LLFF consists of 8 real-world scenes captured with a handheld cellphone, featuring 20 to 62 forward-facing images per scene. These scenes encompass a range of complex environments. In line with the standard protocol [27], we reserve 1/8 of these images for testing purposes. The remaining images are used for training, from which we randomly sample three views for input into our model.

Implementation Details. Following DVGO [38, 39], we adopt a coarse-to-fine optimization scheme to stabilize the training of ReVoRF and gradually improve the geometric details. During the whole training period, we set the values of λ_{rel} and λ_{unr} in Eq. 11 as 10^{-1} and 10^{-2}, respectively. The values of λ_d, λ_f in Eq. 14, and λ_{ds} in Eq. 15 are set as $5{\cdot}10^{-4}$, $5{\cdot}10^{-5}$, and $5{\cdot}10^{-5}$ in the coarse stage, and decreased to $5{\cdot}10^{-5}$, 10^{-5}, and $5{\cdot}10^{-5}$ in the fine stage. The warping poses collected for the Realistic Synthetic 360° dataset [25] and the LLFF dataset [21] are different. For Realistic Synthetic 360°, we randomly vary the polar angle θ and azimuthal angle ϕ in the range of $[5°, 10°]$ based on the input view, and subsequently warp each input sparse view to its four neighboring views defined by $\{(\theta, \phi), (-\theta, \phi), (\theta, -\phi), (-\theta, -\phi)\}$. For the LLFF dataset, the warped views are obtained by randomly interpolating between every adjacent input view. To speed up the training stage and improve the quality of depth supervision, the warping is performed periodically, which updates the warped depth maps and RGB images every 1000 training steps.

(a) RegNeRF [30] (b) SparseNeRF [12] (c) VGOS [40] (d) ReVoRF (Ours) (e) GT

Figure 4. Comparisons on the LLFF dataset [26] in 3-view setting. The red and blue boxes denote compared regions. Our approach achieves better results in reconstructing fine details with enhanced clarity. Please zoom in for details.

	NeRF LLFF dataset			
Methods	PSNR↑	SSIM↑	LPIPS↓	Training Time↓
PixelNeRF [48]	16.17	0.438	0.512	3-4 days* + 10 hrs
SRF [7]	17.07	0.436	0.529	2-3 days* + 43mins
MVSNeRF [5]	17.88	0.584	<u>0.327</u>	1-2 days* + 10mins
Mip-NeRF [2]	14.62	0.351	0.495	14 hrs
DietNeRF [14]	14.94	0.370	0.496	18 hrs
3DGS [17]	13.05	0.407	0.388	13 mins
RegNeRF [30]	19.08	0.587	0.336	4 hrs
VGOS [40]	<u>19.35</u>	0.620	0.432	**5 mins**
SparseNeRF [12]	**19.86**	<u>0.624</u>	0.328	$>$ 2 hrs
Ours	19.26	**0.644**	**0.316**	<u>11 mins</u>

Table 2. Quantitative comparison for 3-views setting on LLFF [21]. The best and the second-best results are highlighted in **bold** and <u>underlined</u>, respectively. (*) signifies the pre-training time.

Evaluation Metrics. To assess the effectiveness of our method, we employ several established metrics, including PSNR (Peak Signal-to-Noise Ratio) for assessing image reconstruction accuracy, SSIM (Structural Similarity Index Measure) [44] for evaluating changes in luminance and contrast that affect structural integrity, and LPIPS (Learned Perceptual Image Patch Similarity) [51], which uses deep learning to approximate human visual perception. These metrics provide a comprehensive analysis of our model's performance, covering aspects of accuracy, perceptual quality, and structural fidelity in the reconstructed images.

4.2. Comparisons

On the Realistic Synthetic 360° dataset [25], we compare our method with state-of-the-art approaches, including RegNeRF [30], DietNeRF [14], infoNeRF [18], VGOS [40],PixelNeRF [48], and GeCoNeRF [19], in a 4-view setting. On the LLFF dataset [21], we implement our method in a 3-view setting and compare with SRF [7], MVSNeRF [5], mip-NeRF [2], DietNeRF [14], RegNeRF [30],VGOS [40], SparseNeRF [12] and GeCoNeRF [19]. We adopt the reported results from VGOS [40], sparseNeRF [12], and GeCoNeRF [19]. Besides, we also compare with the advanced reconstruction method 3DGS [17].

Qualitative Experiments. Fig. 3 compares our approach with some recent methods on the Realistic Synthetic 360° dataset [25]. Diet-NeRF [14] performs poorly in the setting of 4 views, while infoNeRF [18] and VGOS [40] are inferior to our method in terms of both geometric shapes and detail resolution. Our approach demonstrates superior performance in both geometry and details.

Fig. 4 shows a qualitative comparison on a scene from the LLFF dataset [21]. While all methods can recover the overall structure of the scene, our approach excels at the quality of details as shown in the magnified regions. Our method incorporates smoothness while retaining fine details, achieving the most natural results.

Quantitative Experiments. Table 1 shows quantitative results from different methods on the Realistic Synthetic 360° dataset [25]. In terms of training time, our method is at least an order of magnitude faster than all other methods except for VGOS [40]. Although our method is slightly lower than VGOS [40], it significantly enhances the PSNR, LPIPS [51], and SSIM [44] of the rendered images. Our method achieves state-of-the-art accuracy in PSNR and LPIPS [51]. Additionally, despite not utilizing a pre-trained model or perceptual loss for high-level semantic information extraction, our method still achieves the second-best performance in perceived SSIM [44].

Table 2 shows a quantitative comparison on the LLFF dataset [21]. Our method achieves the highest scores in both SSIM [44] and LPIPS [51], indicating that our images exhibit the best performance in terms of human perceptual reconstruction. We also achieve the third-highest PSNR, which, together with our state-of-the-art performance in SSIM [44] and LPIPS [51], demonstrates that our approach has made improvements in certain aspects of image rendering.

4.3. Ablation Study

Our ablation study is segmented into five distinct groups with Table 3, with DVGO [38, 39] serving

Baseline | Baseline+L_{rel} | Baseline+L_{rel}+L_{unr} | Baseline+L_{rel}+L_{unr}+L_{ds} | Full Model

Figure 5. Visualizations of the ablation on Chair scene from the Realistic Synthetic 360° [27] dataset in 4 views setting. With the proposed losses, our methods could gradually improve the cross-view consistency and reduce the noise compared with the baseline.

L_{rs}	L_{ds}	L_{unr}	L_{rel}	PSNR↑	SSIM↑	LPIPS↓
				17.19	0.767	0.223
			✓	17.79	0.780	0.243
		✓	✓	19.01	0.805	0.228
	✓	✓	✓	<u>19.23</u>	<u>0.811</u>	<u>0.220</u>
✓	✓	✓	✓	**20.72**	**0.848**	**0.179**

Table 3. Ablation study on the Realistic Synthetic 360° dataset [25] in the 4-view setting. The best and the second-best results are highlighted in **bold** and <u>underlined</u>, respectively.

Methods	PSNR↑			SSIM↑			LPIPS↓		
	3-view	6-view	9-view	3-view	6-view	9-view	3-view	6-view	9-view
SRF [7]	17.07	16.75	17.39	0.436	0.438	0.465	0.529	0.521	0.503
PixelNeRF [48]	16.17	17.03	18.92	0.438	0.473	0.535	0.512	0.477	0.430
MVSNeRF [5]	17.88	19.99	20.47	0.584	0.660	0.695	0.327	0.264	0.244
DVGO [38]	16.60	21.25	<u>22.89</u>	0.560	<u>0.704</u>	<u>0.746</u>	<u>0.422</u>	**0.246**	<u>0.228</u>
VGOS [40]	**19.35**	<u>21.55</u>	22.39	<u>0.620</u>	0.671	0.692	0.432	0.328	0.325
Ours	<u>19.26</u>	**22.21**	**23.04**	**0.644**	**0.720**	**0.753**	**0.316**	<u>0.269</u>	**0.225**

Table 4. Comparison of 3, 6, and 9 input views on LLFF [21]. The best and the second-best results are highlighted in **bold** and <u>underlined</u>, respectively.

as the baseline. We incrementally introduce our proposed contributions along with various regularization methods to enhance the model's rendering quality. The groups are delineated as: baseline, baseline+L_{rel}, baseline+L_{rel}+L_{unr}, baseline+L_{rel}+L_{unr}+L_{ds}, and baseline+L_{rel}+L_{unr}+L_{ds}+L_{rs}. The ablation results reveal that each incremental contribution positively impacts the rendering quality in various aspects. After the addition of L_{rel}, our PSNR increased by 0.6. Subsequent inclusion of L_{unr} led to a further rise in PSNR by **1.22**. With the incorporation of L_{ds}, the PSNR went from 19.01 to 19.23. Finally, after adding L_{rs}, our PSNR peaked at **20.72**, SSIM [44] reached **0.848**, and LPIPS [51] arrived at **0.179**, marking a significant enhancement.

Besides, we explore the potential of ReVoRF under the settings where more input views are available. We report extra results for 6-view and 9-view settings in Table 4, demonstrating that increased input views generally enhance performance. ReVoRF maintains superiority across these settings.

Fig. 5 presents the visualization of our ablation study on the Chair scene. We incorporated L_{rel} to enhance the texture quality of the model, which, however, introduced some noise artifacts. To mitigate these artifacts, L_{unr} and L_{ds} were subsequently integrated. These adjustments successfully reduced noise but at the cost of blurring the geometric structures in the process. The issue of maintaining geometric consistency while eliminating noise was addressed through the implementation of L_{rs}. We observe that our method effectively prevents the collapse of new views caused by overfitting due to a limited number of viewpoints.

5. Conclusion, Limitation, and Future Work

In this paper, we address the challenge of view deformation by discerning reliable and unreliable areas, subsequently introducing a bilateral geometric consistency regularization. This approach maximizes the use of reliable regions while delicately exploring the depth in unreliable areas, applying a more lenient constraint to these zones. Further extending our method into voxel space, we transform 2D reliable areas into 3D space through a reliability-aware voxelization smoothing process. Our method, when applied to various datasets, has proven to be highly precise, significantly bolstering geometric consistency and demonstrating its efficacy in intricate 3D reconstruction tasks.

Our method shares a common limitation of the voxel-based method: the tendency to produce smoothed results, leading to a loss in fine details. Besides, the exceptionally challenging context for NeRF with sparse input also limits its application in more complex scenes, such as large-scale scene reconstruction. For future work, we aim to refine the voxelization technique to better preserve details, potentially exploring hybrid models that combine voxel-based methods with alternative geometric representations for a more detailed reconstruction.

Acknowledgement. The work is supported by the Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2023B1515020097), Singapore MOE Tier 1 Funds (MSS23C002), and the NRF Singapore under the AI Singapore Programme (No. AISG3-GV-2023-011).

References

[1] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In *NeurIPS*, volume 33, pages 14927–14937, 2020. 3

[2] 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 *ICCV*, pages 5855–5864, 2021. 1, 2, 7

[3] Matteo Bortolon, Alessio Del Bue, and Fabio Poiesi. Vm-nerf: Tackling sparsity in nerf with view morphing. In *ICIAP*, pages 63–74. Springer, 2023. 2, 3

[4] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *CVPR*, pages 130–141, 2023. 2

[5] 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. In *ICCV*, pages 14124–14133, 2021. 2, 3, 7, 8

[6] Di Chen, Yu Liu, Lianghua Huang, Bin Wang, and Pan Pan. Geoaug: Data augmentation for few-shot nerf with geometry constraints. In *ECCV*, pages 322–337. Springer, 2022. 1

[7] Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. In *CVPR*, pages 7911–7920, 2021. 3, 7, 8

[8] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, pages 12882–12891, 2022. 1

[9] Yilun Du, Yinan Zhang, Hong-Xing Yu, Joshua B. Tenenbaum, and Jiajun Wu. Neural radiance flow for 4d view synthesis and video processing. In *ICCV*, 2021. 2

[10] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *CVPR*, pages 5501–5510, 2022. 2

[11] Jun Gao, Tianchang Shen, Zian Wang, Wenzheng Chen, Kangxue Yin, Daiqing Li, Or Litany, Zan Gojcic, and Sanja Fidler. Get3d: A generative model of high quality 3d textured shapes learned from images. *NeurIPS*, 35:31841–31854, 2022. 2

[12] Zhaoxi Chen Guangcong, Chen Change Loy, and Ziwei Liu. Sparsenerf: Distilling depth ranking for few-shot novel view synthesis. In *ICCV*, 2023. 1, 7

[13] Wenbo Hu, Yuling Wang, Lin Ma, Bangbang Yang, Lin Gao, Xiao Liu, and Yuewen Ma. Tri-miprf: Tri-mip representation for efficient anti-aliasing neural radiance fields. In *ICCV*, pages 19774–19783, 2023. 1, 2

[14] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *ICCV*, pages 5885–5894, 2021. 1, 2, 3, 6, 7

[15] Yutao Jiang, Yang Zhou, Yuan Liang, Wenxi Liu, Jianbo Jiao, Yuhui Quan, and Shengfeng He. Diffuse3d: Wide-angle 3d photography via bilateral diffusion. In *ICCV*, pages 8998–9008, 2023. 1

[16] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In *NeurIPS*, pages 5574–5584, 2017. 3

[17] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), July 2023. 6, 7

[18] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *CVPR*, pages 12912–12921, 2022. 1, 2, 3, 6, 7

[19] Min-Seop Kwak, Jiuhn Song, and Seungryong Kim. Geconerf: Few-shot neural radiance fields via geometric consistency. In *ICML*. JMLR.org, 2023. 1, 2, 3, 6, 7

[20] 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 Autom. Lett.*, 7(4):12070–12077, 2022. 3

[21] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *CVPR*, pages 6498–6508, 2021. 2, 6, 7, 8

[22] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields. In *ICCV*, pages 5773–5783, 2021. 2

[23] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, pages 7210–7219, 2021. 2, 3

[24] Alican Mertan, Damien Jade Duff, and Gözde Ünal. Siralama sorunu olarak nispi derinlik tahmini relative depth estimation as a ranking problem. In *2020 28th Signal Processing and Communications Applications Conference (SIU)*, pages 1–6. IEEE, 2020. 5

[25] B Mildenhall, PP Srinivasan, M Tancik, JT Barron, R Ramamoorthi, and R Ng. Nerf: Representing scenes

as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 2, 3, 6, 7, 8

[26] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: practical view synthesis with prescriptive sampling guidelines. *ACM TOG*, 38(4):29:1–29:14, 2019. 6, 7

[27] 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 *ECCV*, volume 12346, pages 405–421, 2020. 6, 8

[28] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG*, 41(4):1–15, 2022. 1, 2

[29] Radford M. Neal. *Bayesian learning for neural networks*. PhD thesis, University of Toronto, Canada, 1995. 3

[30] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, 2022. 5, 7

[31] Xuran Pan, Zihang Lai, Shiji Song, and Gao Huang. Activenerf: Learning where to see with uncertainty estimation. In *ECCV*, volume 13693, pages 230–246, 2022. 3

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763. PMLR, 2021. 3

[33] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 2020. 4

[34] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In *NeurIPS*, volume 31, 2018. 3

[35] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *CVPR*, pages 16632–16642, 2023. 2

[36] Jianxiong Shen, Antonio Agudo, Francesc Moreno-Noguer, and Adria Ruiz. Conditional-flow nerf: Accurate 3d modelling with reliable uncertainty quantification. In *ECCV*, volume 13663, pages 540–557, 2022. 3

[37] Jianxiong Shen, Adria Ruiz, Antonio Agudo, and Francesc Moreno-Noguer. Stochastic neural radiance fields: Quantifying uncertainty in implicit 3d representations. In *3DV*, pages 972–981, 2021. 3

[38] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *CVPR*, pages 5459–5469, 2022. 2, 5, 6, 7, 8

[39] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Improved direct voxel grid optimization for radiance fields reconstruction. *arXiv preprint arXiv:2206.05085*, 2022. 2, 5, 6, 7

[40] Jiakai Sun, Zhanjie Zhang, Jiafu Chen, Guangyuan Li, Boyan Ji, Lei Zhao, and Wei Xing. Vgos: Voxel grid optimization for view synthesis from sparse inputs. In *IJCAI*, pages 1414–1422, 8 2023. 3, 6, 7, 8

[41] Feitong Tan, Sean Fanello, Abhimitra Meka, Sergio Orts-Escolano, Danhang Tang, Rohit Pandey, Jonathan Taylor, Ping Tan, and Yinda Zhang. Volux-gan: A generative model for 3d face synthesis with hdri relighting. In *ACM TOG*, pages 1–9, 2022. 2

[42] 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. In *CVPR*, pages 8248–8258, 2022. 1

[43] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, volume 33, pages 7537–7547, 2020. 3

[44] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 7, 8

[45] Jamie Wynn and Daniyar Turmukhambetov. DiffusioNeRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models. In *CVPR*, 2023. 3

[46] Dongyu Yan, Jianheng Liu, Fengyu Quan, Haoyao Chen, and Mengmeng Fu. Active implicit object reconstruction using uncertainty-guided next-best-view optimization. *IEEE Robotics Autom. Lett.*, 8(10):6395–6402, 2023. 3

[47] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *CVPR*, pages 8254–8263, 2023. 1, 3

[48] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, pages 4578–4587, 2021. 1, 2, 3, 6, 7, 8

[49] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. Nerf-editing: geometry editing of neural radiance fields. In *CVPR*, pages 18353–18364, 2022. 2

[50] 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 *CVPR*, pages 18376–18386, 2022. 1, 2, 3

[51] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 7, 8

[52] Chenxi Zheng, Bangzhen Liu, Xuemiao Xu, Huaidong Zhang, and Shengfeng He. Learning an interpretable stylized subspace for 3d-aware animatable artforms. *IEEE TVCG*, 2024. 2

[53] Chenxi Zheng, Bangzhen Liu, Huaidong Zhang, Xuemiao Xu, and Shengfeng He. Where is my spot? few-shot image generation via latent subspace optimization. In *CVPR*, pages 3272–3281, 2023. 2

[54] Yang Zhou, Hanjie Wu, Wenxi Liu, Zheng Xiong, Jing Qin, and Shengfeng He. Single-view view synthesis with self-rectified pseudo-stereo. *IJCV*, 131(8):2032–2043, 2023. 2