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

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

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 smoothes 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\degree 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.
\textsuperscript{†}Corresponding author (shengfenghe@smu.edu.sg).
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 \( x \in \mathbb{R}^3 \) and the associated 2D viewing directions \( d \in \mathbb{R}^2 \), NeRF maps them into a volume density \( \sigma \in \mathbb{R} \) and an RGB value \( c \in \mathbb{R}^3 \), such that:

\[
(c, \sigma) = f(\gamma(x), \gamma(d)),
\]

where the \( \gamma \) is a positional encoding that projects \( x \) and \( d \) into a higher dimensional feature space [43]. With a ray parameterized as \( r_p(t) = o + td_p \), cast from the camera’s optical center \( o \) along direction \( d_p \), the expected color \( \hat{C}(r_p) \) of pixel \( p \) is rendered as follows:

\[
\hat{C}(r_p) = \int_{t_n}^{t_f} T(t)\sigma(r_p(t))c(r_p(t), d_p)\,dt,
\]

where \( t_n \) and \( t_f \) are the near and far bounds of the ray for sampling, and \( T(t) = \exp\left(-\int_{t_n}^{t} \sigma(r(s))\,ds\right) \) denotes the cumulative transparency along the ray from \( t_n \) to \( t_f \).

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 \( I_i^t \) 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 \( I_{s\leftrightarrow r}^{i,j} \) 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 \( \hat{D}_c \), where the depth value on each pixel \( p \) is accumulated by the density along each ray \( r_p \) omitted from the camera:

\[
\hat{D}_c(p) = \int_{t_n}^{t_f} T(t)\sigma(r_p(t))\,dt.
\]

Subsequently, we obtain the warped image \( I_{s\leftrightarrow r}^{i,j} \) through a cross-view transformation \( H_{s\leftrightarrow r} \), which deforms each pixel \( p_r \) of \( I_i^t \) 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.

Warped images $I_{w\leftarrow r}$

Estimated Depth $D_{w\leftarrow r}$

Depth estimation

Warp

Novel view $s$

Unreliable voxel

Reliable voxel

Smoothing voxel

Input view $r$

Voxel density $

\sum_{p \in R_{rel}} \| I^{i,j}_{s\leftarrow r}(p) - I^i_s(p) \|^2.

(8)

For $R_{warp}$, we propose to leverage the relative depth prior of the warped $I^{i,j}_{s\leftarrow r}$ 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

\begin{align*}
p_{s\leftarrow r} &= H_{s\leftarrow r}(p_r) \\
&= f_{s\leftarrow w}(f_{w\leftarrow r}(p_r)),
\end{align*}

(3)

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

\begin{align*}
f_{w\leftarrow r}(p_r) &= D^i_r(p_r)T^{-1}_rK^{-1}_r p_r, \\
f_{s\leftarrow w}(p_w) &= K_sT_s(p_w),
\end{align*}

(4)

(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 $I^{i,j}_{s\leftarrow r}$. 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 $I^i_r$ and $I^{i,j}_{s\leftarrow r}$. To achieve this, we obtain the pseudo depth $D^i_s$ of view $j$ by rendering from the radiance fields. Following Eq. 4, we map the pixel of $I^i_r$ and $I^{i,j}_{s\leftarrow r}$ into the same coordinate and obtain the correlation map $M_{cor}$ by comparing the distance between each pixel pair $(p_r, p_{s\leftarrow r})$:

\begin{align*}
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}
\end{align*}

(6)

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

\begin{align*}
M_{s\leftarrow r} = M_{cor} \cup M_{warp}.
\end{align*}

(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 $I^i_s$ on view $j$ via a reconstruction loss defined as:

\begin{align*}
L_{rel} = \sum_{p \in R_{rel}} \| I^{i,j}_{s\leftarrow r}(p) - I^i_s(p) \|^2.
\end{align*}

(8)
the voids occurring in $D_{se-r}$. Subsequently, a relative depth regularization loss [24] is introduced to constrain the geometric consistency on the rendered depth $D_r$, which is defined as:

$$L_{unr} = \sum_{p \in R_{unr}} \sum_{\hat{p} \in N(D_{se-r})} h(p, \hat{p}),$$

where

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

Here $p$ denotes any pixel within the unreliable region $R_{unr}$; $\hat{p}$ represents each pixel within a neighborhood $N(D_{se-r})$ of $p$, where $N(D_{se-r})$ is obtained by calculating pixels that have close depth values with $D_{se-r}$ in the warped depth map; $D^p_s$, $D^\hat{p}_s$ and $D^p_{se-r}$, $D^\hat{p}_{se-r}$ denote the depth values of $p$ and $\hat{p}$ within the rendered depth map $D_r$, and estimated depth map $D_{se-r}$, 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 $D^p_{se-r} > D^\hat{p}_{se-r}$, or $D^\hat{p}_s > D^p_s$ but $D^\hat{p}_{se-r} < D^p_{se-r}$, then $h(p, \hat{p})$ penalizes the depth difference $|D^p_s - D^\hat{p}_s|$ 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 $L_{unr}$:

$$L_{bgc} = \lambda_{rel} L_{rel} + \lambda_{unr} L_{unr}. \quad \text{(11)}$$

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 $f_v$ and density $\sigma$ 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 $f_v$ and $\sigma$ between a given voxel $v$ with its six adjacent points $V$ as follows:

$$L(v) = \sum_{v \in V} \Delta f_v(v, \hat{v}) + \Delta \sigma(v, \hat{v}), \quad \text{(12)}$$

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 $\rho(v)$ for each voxel in the grid. Specifically, given a warped image, we cast $O$ rays $r$ through each pixel of the reliable regions $R_{rel}$. For each voxel $v$, we obtain a reliability score by accumulating the number of rays that pass this voxel, denoted by $S(v)$. The maximum number of times being passed by rays is denoted as $S(v)_{max}$. Then the reliability-aware smooth factor is defined as $\rho(v) = \frac{S(v)}{S(v)_{max}}$. Finally, we formulate the reliability-aware voxel smoothing losses on $f_v$ and $\sigma$ as:

$$L_{f_v} = \sum_v \sum_{\hat{v} \in V} (1 + e^{-\rho(v)}) \Delta f_v(v, \hat{v}),$$
$$L_{\sigma} = \sum_v \sum_{\hat{v} \in V} (1 + e^{-\rho(v)}) \Delta \sigma(v, \hat{v}). \quad \text{(13)}$$

The final regularization loss is defined as follows:

$$L_{ras} = \lambda_f L_{f_v} + \lambda_d L_{\sigma}. \quad \text{(14)}$$

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 $v$ with a reliability weight $w_v = 1 + \rho(v)$, 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:

$$L_{ds} = \frac{1}{|R|} \sum_{r \in R} \sum_{(x, y) \in D} \|d(x, y) - d(x, y + 1)\|^2_2 + \|d(x, y) - d(x + 1, y)\|^2_2, \quad \text{(15)}$$
ReV oRF (Ours) GT

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

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 underlined, respectively. (*) denotes the time cost of pre-training.

<table>
<thead>
<tr>
<th>$\text{Methods}$</th>
<th>(\text{PSNR} \uparrow)</th>
<th>(\text{SSIM} \uparrow)</th>
<th>(\text{LPIPS} \downarrow)</th>
<th>(\text{Training Time})</th>
</tr>
</thead>
<tbody>
<tr>
<td>NeRF [25]</td>
<td>15.93</td>
<td>0.780</td>
<td>0.320</td>
<td>2 hrs</td>
</tr>
<tr>
<td>PixelNeRF [48]</td>
<td>16.09</td>
<td>0.738</td>
<td>0.390</td>
<td>3-4 days* + 10 hrs</td>
</tr>
<tr>
<td>DietNeRF [14]</td>
<td>16.06</td>
<td>0.793</td>
<td>0.306</td>
<td>19 hrs</td>
</tr>
<tr>
<td>3DGS [17]</td>
<td>17.55</td>
<td>0.701</td>
<td>0.250</td>
<td>3 mins</td>
</tr>
<tr>
<td>InfoNeRF [18]</td>
<td>18.62</td>
<td>0.811</td>
<td>0.230</td>
<td>4 hrs</td>
</tr>
<tr>
<td>VGOS [40]</td>
<td>18.91</td>
<td>0.825</td>
<td>0.205</td>
<td>3 mins</td>
</tr>
<tr>
<td>GeCoNeRF [19]</td>
<td>19.78</td>
<td>0.880</td>
<td>0.185</td>
<td>&gt; 2 hrs</td>
</tr>
<tr>
<td>Ours</td>
<td>\textbf{20.72}</td>
<td>\textbf{0.848}</td>
<td>\textbf{0.179}</td>
<td>7 mins</td>
</tr>
</tbody>
</table>

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 ReV oRF is formulated as:

\[
L_{\text{total}} = L_{\text{rgb}} + L_{\text{bgc}} + L_{\text{rs}} + \lambda_{\text{ds}} L_{\text{ds}}. \tag{16}
\]

4. Experiments

In this section, we demonstrate the superiority of the proposed ReV oRF 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 ReV oRF and gradually improve the geometric details. During the whole training period, we set the values of \(\lambda_{\text{rel}}\) and \(\lambda_{\text{unr}}\) in Eq. 11 as \(10^{-1}\) and \(10^{-2}\), respectively. The values of \(\lambda_{d}, \lambda_f\) in Eq. 14, and \(\lambda_{ds}\) 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 \(\theta\) and azimuthal angle \(\phi\) 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.
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...
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.

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

<table>
<thead>
<tr>
<th>Lrs</th>
<th>Lds</th>
<th>Lunr</th>
<th>Lrel</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
<th>LPIPS↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>17.19</td>
<td>0.767</td>
<td>0.223</td>
</tr>
</tbody>
</table>

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 underlined, respectively.

<table>
<thead>
<tr>
<th>Methods</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
<th>LPIPS↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRF [7]</td>
<td>17.07</td>
<td>16.75</td>
<td>17.39</td>
</tr>
<tr>
<td>PixelNeRF [48]</td>
<td>16.17</td>
<td>17.03</td>
<td>18.92</td>
</tr>
<tr>
<td>DVGO [38]</td>
<td>16.60</td>
<td>21.25</td>
<td>22.89</td>
</tr>
<tr>
<td>VGOS [40]</td>
<td>19.26</td>
<td>22.21</td>
<td>23.04</td>
</tr>
<tr>
<td>Ours</td>
<td>19.35</td>
<td>21.55</td>
<td>22.39</td>
</tr>
</tbody>
</table>

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


[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


[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


[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


[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


[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
