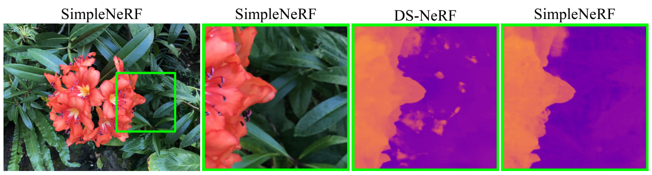
SimpleNeRF: Regularizing Sparse Input Neural Radiance Fields with Simpler Solutions

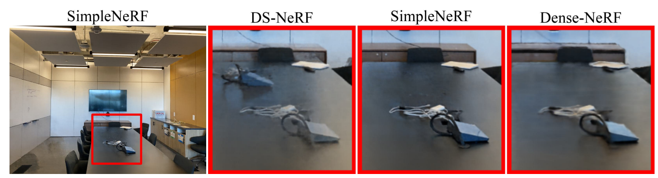
Nagabhushan Somraj
Indian Institute of Science
Bengaluru, Karnataka, India
nagabhushans@iisc.ac.in

Adithyan Karanayil
Indian Institute of Science
Bengaluru, India
adithyanv@iisc.ac.in

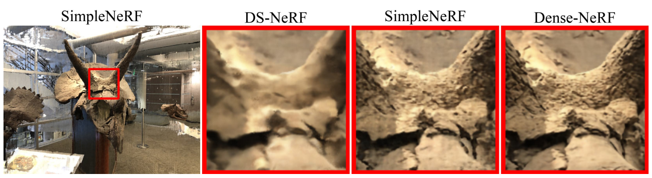
Rajiv Soundararajan
Indian Institute of Science
Bengaluru, India
rajivs@iisc.ac.in



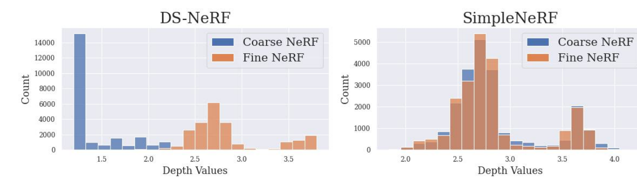
(a) Mitigation of floaters: For a frame from the LLFF flower scene, we show the depths predicted by DS-NeRF and SimpleNeRF. We show the complete frame for reference and focus on a small region to better observe the floaters.



(b) Reducing shape-radiance ambiguity: This is an example from the LLFF room scene. For reference, we show the prediction of Vanilla NeRF trained with dense input views.



(c) Improving Sharpness: The above images correspond to the LLFF horns scene. We enlarge a small region of the frame to better observe the improvement in sharpness.



(d) Improving Sharpness: Histogram of depth values predicted by the coarse and fine NeRF models for the image patch shown in Fig (c).

Figure 1: We show three shortcomings of DS-NeRF when trained with two input views on the LLFF dataset. SimpleNeRF introduces regularizations on DS-NeRF to mitigate these distortions. Notice the floaters in DS-NeRF predictions shown by small orange regions in Fig. (a) which are cleaned by SimpleNeRF. In Fig. (b), we find DS-NeRF suffers from shape-radiance ambiguity and ends up changing the color of the table in different viewpoints to match the observed images. DS-NeRF blends the two input images based on the viewpoint instead of learning correct geometry. This leads to ghosting artifacts in novel viewpoints, which is removed by our model. In Fig (c), we observe SimpleNeRF producing sharper reconstructions than DS-NeRF. Fig (d) shows a possible reason, where the coarse and fine NeRF models in DS-NeRF converge to different depth estimates. This leads to ineffective hierarchical sampling resulting in blurry predictions. We find that SimpleNeRF mitigates this by predicting consistent depth estimates.

ABSTRACT

Neural Radiance Fields (NeRF) show impressive performance for the photo-realistic free-view rendering of scenes. However, NeRFs require dense sampling of images in the given scene, and their performance degrades significantly when only a sparse set of views are available. Researchers have found that supervising the depth estimated by the NeRF helps train it effectively with fewer views. The depth supervision is obtained either using classical approaches or neural networks pre-trained on a large dataset. While the former may provide only sparse supervision, the latter may suffer from generalization issues. As opposed to the earlier approaches, we seek to learn the depth supervision by designing augmented models and training them along with the NeRF. We design augmented models that encourage simpler solutions by exploring the role of positional encoding and view-dependent radiance in training the few-shot NeRF. The depth estimated by these simpler models is used to supervise the NeRF depth estimates. Since the augmented models can be inaccurate in certain regions, we design a mechanism to choose only reliable depth estimates for supervision. Finally, we add a consistency loss between the coarse and fine multi-layer perceptrons of the NeRF to ensure better utilization of hierarchical sampling. We achieve state-of-the-art view-synthesis performance on two popular datasets by employing the above regularizations. The source code for our model can be found on our project page: https://nagabhushansn95.github.io/publications/2023/SimpleNeRF.html

SA Conference Papers '23, December 12–15, 2023, Sydney, NSW, Australia

ACM ISBN 979-8-4007-0315-7/23/12…$15.00
https://doi.org/10.1145/3610548.3618188

CCS CONCEPTS

• **Computing methodologies → Rendering**; **Volumetric models**; *Computer vision*; *Computational photography*; 3D imaging; Reconstruction.

KEYWORDS

neural rendering, novel view synthesis, sparse input NeRF

ACM Reference Format:
Nagabhushan Somraj, Adithyan Karanayil, and Rajiv Soundararajan. 2023. SimpleNeRF: Regularizing Sparse Input Neural Radiance Fields with Simpler Solutions. In *SIGGRAPH Asia 2023 Conference Papers (SA Conference Papers '23), December 12–15, 2023, Sydney, NSW, Australia.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3610548.3618188

1 INTRODUCTION

Neural Radiance Fields (NeRFs) [Mildenhall et al. 2020] show unprecedented levels of performance in synthesizing novel views of a scene by learning a volumetric representation implicitly within the weights of multi-layer perceptrons (MLP). Although NeRFs are very promising for view synthesis, there is a need to improve their design in a wide array of scenarios. For example, NeRFs have been enhanced to learn on unbounded scenes [Barron et al. 2022], handle scenes with highly specular objects [Verbin et al. 2022], and deal with noisy camera poses [Bian et al. 2023]. Yet NeRFs require tens to hundreds of images per scene to learn the scene geometry accurately, and their quality deteriorates significantly when only a few training images are available [Jain et al. 2021]. In this work, we focus on training the NeRF with a sparse set of input images and aim to design novel regularizations for effective training.

Prior work on sparse input NeRFs can be classified into generalized models and regularization based models. Generalized models [Johari et al. 2022] use convolutional neural networks to obtain a latent representation of the scene and use it to condition the NeRF. However, these models require a large dataset for pre-training and may suffer from generalization issues when used to render a novel scene [Niemeyer et al. 2022]. The other thread of work on sparse input NeRFs follows the original NeRF paradigm of training scene-specific NeRFs, and designs novel regularizations to assist NeRFs in converging to a better scene geometry [Zhang et al. 2021]. One popular approach among such models is to supervise the depth estimated by the NeRF. RegNeRF [Niemeyer et al. 2022], DS-NeRF [Deng et al. 2022b] and ViP-NeRF [Somraj and Soundararajan 2023] use simple priors such as depth smoothness, sparse depth or relative depth respectively obtained through classical approaches. On the other hand, DDP-NeRF [Roessle et al. 2022] and SCADE [Uy et al. 2023] pre-train convolutional neural networks (CNN) on a large dataset of scenes to learn a prior on the depth or its probability distribution. These approaches may also suffer from similar issues as the generalized models. This raises the question of whether we can instead learn the depth supervision in-situ without employing any pre-training. We aim to regularize NeRFs by learning augmented models for depth supervision in tandem with the NeRF training.

We follow the popular Occam's razor principle and regularize the NeRF by designing augmented models to choose simpler solutions over complex ones, wherever possible. Some of the key components of the NeRF such as positional encoding, view-dependent radiance, and hierarchical sampling provide powerful capabilities to the NeRF and are designed for the case with dense input views. Existing implementations of these components may be sub-optimal with fewer input views, perhaps due to the highly under-constrained system, causing several distortions. Fig. 1 shows three common distortions, namely, floaters [Roessle et al. 2022], shape-radiance ambiguity [Zhang et al. 2020], and blurred renders for the NeRF model in the few-shot setting. We believe that by simplifying some of the capabilities, one can obtain simpler augmented models that may provide better depth supervision for training the NeRF.

We simplify the capability of the augmented NeRF models with respect to positional encoding and view-dependent radiance. Positional encoding maps two nearby points in 3D space to far-apart points in the encoded space. This allows the NeRF to learn sharp depth discontinuities in 3D space via a smooth function in the encoded space. However, in an under-constrained system, the NeRF learns many undesirable depth discontinuities that can still explain the observed images. This leads to floater artifacts as seen in Fig. 1a. Restricting the ability of the NeRF to learn these sharp discontinuities encourages it to mitigate the floater artifacts. Similarly, the ability of the NeRF to predict view-dependent radiance leads to shape-radiance ambiguity, which we find to be more pronounced in the case of few-shot NeRF. The NeRF can learn to explain the observed images by modifying the color of 3D points in accordance with the input images as seen in Fig. 1b. Restricting the NeRF to predict view-independent radiance only encourages the NeRF to explain the observed images using simpler Lambertian surfaces and can help avoid shape-radiance ambiguity.

We use the depth estimated by the simpler augmented models to supervise the depth estimated by the NeRF model. Note that we use the simplified models as augmentations for depth supervision and not as the main NeRF model since naïvely reducing the capacity of the NeRF may lead to suboptimal solutions [Jain et al. 2021]. For example, the model that predicts view-independent radiance fails if the scene contains specular objects. Further, the simpler augmented models need to be used for supervision only if they explain the observed images accurately. We gauge the reliability of the depths by reprojecting pixels using the estimated depths onto a different nearest train view and comparing them with the corresponding images. We use DS-NeRF [Deng et al. 2022b] as our baseline and design our regularizations on top of it. Our framework can thus be seen as a semi-supervised learning model by considering the sparse depth from a Structure from Motion (SfM) module as providing limited depth labels and the remaining pixels as the unlabeled data. Our approach of using augmented models in tandem with the main NeRF model is perhaps closest to the Dual-Student architecture [Ke et al. 2019] that trains another identical model in tandem with the main model and imposes consistency regularization between the predictions of the two models. In contrast, our augmented models have complementary abilities as compared to the main NeRF model.

Finally, we observe that the coarse and fine NeRFs may converge to different depth estimates when trained with fewer images, as seen in Fig. 1d. This essentially renders the hierarchical sampling ineffective. The resulting model is similar to the one with under-sampled points along the rays. We avoid such degenerate cases by imposing a consistency loss between the depths estimated by the coarse and fine MLPs. We achieve state-of-the-art view synthesis performance on two popular datasets with the above three regularizations. Further, we show that our model learns significantly improved geometry as compared to the prior art. We refer to our model as SimpleNeRF.

We list the main contributions of our work in the following.

- We design two augmented NeRF models that are biased towards simpler solutions and use the depth estimates from these models to regularize the NeRF training. Through the two augmentations, we mitigate incorrect depth discontinuities and shape-radiance ambiguity.
- We design a mechanism to determine whether the depths estimated by the augmented models are accurate and utilize only the accurate estimates to supervise the NeRF.
- We improve the effectiveness of hierarchical sampling by introducing a consistency constraint between the coarse and fine NeRFs. This generates sharper frames.
- We achieve the state-of-the-art performance of few shot NeRFs on two popular datasets.

2 RELATED WORK

Novel view synthesis is a classic problem traditionally solved broadly using image based rendering [Chen and Williams 1993] or light fields [Gortler et al. 1996; Levoy and Hanrahan 1996]. The seminal work by Mildenhall et al. [2020] started a new pathway in neural view synthesis and led to NeRF based models being employed in a wide variety of applications such as 3D editing of scenes [Yuan et al. 2022], gaming [Menapace et al. 2023], extended reality [Deng et al. 2022a] and image reconstruction [Ma et al. 2022; Mildenhall et al. 2022; Pearl et al. 2022], among others. However, many of the above models require dense sampling of input views for a faithful 3D geometry generation. With fewer views, the quality of rendered novel views and the learned 3D geometry degrade significantly introducing severe distortions. This can limit the widespread usage of NeRFs in multiple applications and hence addressing this limitation is of considerable interest.

2.1 Generalized Sparse Input NeRF

Prior work involves various approaches to learning a 3D neural representation with fewer input views. One line of approach attempts to train a generalized model on a large dataset of scenes such that the model can utilize the learned prior to generate a 3D scene representation from the few input images [Chen et al. 2021; Lee et al. 2023; Tancik et al. 2021]. Early pieces of work such as PixelNeRF [Yu et al. 2021], GRF [Trevithick and Yang 2021], and IBRNet [Wang et al. 2021] obtain convolutional features of the input images and additionally condition the NeRF by projecting the 3D points onto the feature grids. MVSNeRF [Chen et al. 2021] and SRF [Chibane et al. 2021] incorporate cross-view knowledge into the features by constructing a cost volume and pair-wise post-processing of the individual frame features respectively. GeoNeRF [Johari et al. 2022] further improves the performance by employing a transformer to effectively reason about the occlusions in the scene. More recent work such as GARF [Shi et al. 2022] and MatchNeRF [Chen et al. 2023] try to provide explicit knowledge about the scene geometry through depth maps and similarity of the projected features respectively. This approach of conditioning the NeRF on learned features is also popular among single image NeRF models [Lin et al. 2023], which can be considered an extreme case of sparse input NeRF. However, the need for pre-training on a large dataset of scenes with multi-view images and issues due to domain shift have motivated researchers to adopt regularization based approaches.

2.2 Regularization based Sparse Input NeRF

A popular approach among regularization based models is to regularize the depth estimated by the NeRF. DS-NeRF [Deng et al. 2022b] uses sparse depth provided by a SfM module to supervise the NeRF estimated depth at sparse keypoints. RegNeRF [Niemeyer et al. 2022] imposes the depth smoothness constraint on the rendered depth maps. ViP-NeRF [Somraj and Soundararajan 2023] instead attempts to regularize the relative depth of objects by obtaining a prior on the visibility of objects. Unlike the above, few other works exploit the advances in depth-estimation using deep neural networks. DDP-NeRF [Roessle et al. 2022] extends DS-NeRF by employing a CNN to convert the sparse depth into dense depth for more supervision. SCADE [Uy et al. 2023] and SparseNeRF [Wang et al. 2023] use the depth map output by single image depth models to constrain the absolute and the relative order of pixel depths respectively. DiffusioNeRF [Wynn and Turmukhambetov 2023] learns the joint distribution of RGBD patches using denoising diffusion models (DDM) and utilizes the gradient of the distribution provided by the DDM to regularize NeRF rendered RGBD patches. Different from the above, our work obtains depth supervision by harnessing the power of learning through augmented models, but without the need for pre-training on a large dataset.

Another line of regularization based approaches hallucinate new viewpoints and regularize the NeRF on different aspects such as semantic consistency [Jain et al. 2021], depth smoothness [Niemeyer et al. 2022], sparsity of mass [Kim et al. 2022] and depth based reprojection consistency [Bortolon et al. 2022; Chen et al. 2022; Kwak et al. 2023; Xu et al. 2022]. More recent works have explored other forms of regularizations. FreeNeRF [Yang et al. 2023] anneals the frequency range of positional encoded NeRF inputs as the training progresses. MixNeRF [Seo et al. 2023] regularizes the NeRF by modeling the volume density along a ray as a mixture of Laplacian distributions. A recent work VDN-NeRF [Zhu et al. 2023], aims to resolve shape-radiance ambiguity, but is designed for training the NeRF with dense input views. However, our regularization is aimed at sparse input NeRF with as few as two views.

3 NERF PRELIMINARIES

We first provide a brief recap of the NeRF and describe the notation required for further sections. To render a pixel $\mathbf{q}$, the NeRF shoots a corresponding ray into the scene and samples N 3D points $\mathbf{p}_1, \mathbf{p}_2, \ldots, \mathbf{p}_N$, where $\mathbf{p}_1$ and $\mathbf{p}_N$ are the closest to and farthest from the camera, respectively. Two MLPs $\mathcal{F}_1, \mathcal{F}_2$ then predict view-independent volume density σ_i and view-dependent color $\mathbf{c}_i$ as

$$\sigma_i, \mathbf{h}_i = \mathcal{F}_1\left(\gamma(\mathbf{p}_i, 0, l_p)\right); \qquad \mathbf{c}_i = \mathcal{F}_2\left(\mathbf{h}_i, \gamma(\mathbf{v}, 0, l_v)\right), \tag{1}$$

where $\mathbf{v}$ is the viewing direction, $\mathbf{h}_i$ is a latent feature of $\mathbf{p}_i$ and

$$\gamma(x, d_1, d_2) = [x, \sin(2^{d_1}x), \cos(2^{d_1}x), \ldots, \sin(2^{d_2-1}x), \cos(2^{d_2-1}x)]$$

is the positional encoding. l_p and l_v are the highest positional encoding frequencies for $\mathbf{p}_i$ and $\mathbf{v}$ respectively. Volume rendering is then applied along every ray to obtain the color for each pixel as $\mathbf{c} = \sum_{i=1}^{N} w_i \mathbf{c}_i$, where the weights w_i are computed as

$$w_i = \exp\left(-\sum_{j=1}^{i-1} \delta_j \sigma_j\right) \cdot \left(1 - \exp\left(-\delta_i \sigma_i\right)\right), \tag{2}$$

and δ_i is the distance between $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$. The expected ray termination length is computed as $z = \sum_{i=1}^{N} w_i z_i$, where z_i is the depth of $\mathbf{p}_i$. z is typically also used as the depth of the pixel $\mathbf{q}$ [Deng et al. 2022b]. $\mathcal{F}_1$ and $\mathcal{F}_2$ are trained using the mean squared error loss (MSE) as $\mathcal{L}_{\text{color}} = \|\mathbf{c} - \hat{\mathbf{c}}\|^2$, where $\hat{\mathbf{c}}$ is the true color of $\mathbf{q}$.

NeRF circumvents the need for the dense sampling of 3D points by employing two sets of MLPs, a coarse NeRF and a fine NeRF, both trained using $\mathcal{L}_{\text{color}}$. The coarse NeRF is trained with a coarse stratified sampling, and the fine NeRF with dense sampling around object surfaces, where object surfaces are coarsely localized based on the predictions of the coarse NeRF.

4 METHOD

Our key idea is to employ simpler NeRF models to obtain better depth supervision in certain regions of the scene. We describe our regularizations based on the simpler solutions in Sec. 4.1. We explain our approach to selecting reliable depth estimates for supervision in Sec. 4.2. Finally, Sec. 4.3 describes our solution to mitigate the sub-optimal utilization of hierarchical sampling. Fig. 2 shows the overall architecture of our model.

4.1 Regularization with Simpler Solutions

Our regularization consists of simplifying the NeRF model with respect to the positional encoding and view-dependent radiance capabilities to obtain better depth supervision. Both positional encoding and view-dependent radiance are elements designed to increase the capability of the NeRF to explain complex phenomena. For example, the former helps in learning thin, repetitive objects against a farther background and the latter for specular objects. However, when training with sparse views, the fewer constraints coupled with the higher capacity of the NeRF lead to solutions that overfit the observed images and learn implausible scene geometries.

We rethink ways of employing positional encoding and view-dependent radiance such that the NeRF utilizes the higher capability only when needed. The challenge here is that it is not known apriori where one needs to employ the higher capability of the NeRF. Our solution here is to use the higher capability NeRF as the main model and employ lower capability NeRFs as augmentations to provide guidance on where to use simpler solutions. Since the scene geometry is mainly learned by the coarse NeRF, we add the augmentations only to the coarse NeRF. We employ two augmentations, one each for regularizing positional encoding and view-dependent radiance, which we describe in the following subsections. In the remainder of this paper, we refer to the two augmentations as points (Sec. 4.1.1) and views (Sec. 4.1.2) augmentations respectively.

4.1.1 Undesirable Depth Discontinuities. The positional encoding maps two nearby points in $\mathbb{R}^3$ to two farther away points in $\mathbb{R}^{3(2l_p+1)}$ allowing the NeRF to learn sharp discontinuities between the two points in $\mathbb{R}^3$ as a smooth function in $\mathbb{R}^{3(2l_p+1)}$. However, this is mainly required at depth edges, and most regions of natural scenes are typically smooth in depth. With sparse input views, the positional encoding causes the NeRF to learn depth discontinuities even in smooth regions due to incorrect matching of pixels between the input views. This gives rise to "floaters" [Barron et al. 2022] where a part of an object is broken away from it and "floats" freely in space. We reduce the depth discontinuities by reducing the highest positional encoding frequency for $\mathbf{p}_i$ to $l_p^{\text{ap}} < l_p$ as

$$\sigma_i, \mathbf{h}_i = \mathcal{F}_1^{\text{ap}}(\gamma(\mathbf{p}_i, 0, l_p^{\text{ap}})), \tag{3}$$

where $\mathcal{F}_1^{\text{ap}}$ is the MLP of the augmented model. The main model is more accurate where depth discontinuities are required and the augmented model is more accurate where discontinuities are not required. We determine the above using a ternary mask m_{ap} as we explain in Sec. 4.2. We supervise the depth predicted by the main NeRF using that of the augmented model, for those pixels for which the depth estimated by the augmented model is reliable, by setting $m_{\text{ap}} = 1$. Similarly, we also determine the pixels for which the depth estimated by the main model is more accurate than that of the augmented model, where we set $m_{\text{ap}} = -1$. In such pixels, we supervise the augmented model with the depth estimated by the main model, which helps the augmented model to improve further and provide better depth supervision for the main model. For pixels where the depth estimated by both the main and augmented models are unreliable, we set $m_{\text{ap}} = 0$. If z_c and z_{ap} are the depths estimated by the coarse NeRF of the main and the augmented models respectively, we impose the depth supervision as

$$\mathcal{L}_{\text{ap}} = \mathbb{1}_{\{m_{\text{ap}}=1\}} \odot \|z_c - \mathcal{T}(z_{\text{ap}})\|^2 + \mathbb{1}_{\{m_{\text{ap}}=-1\}} \odot \|\mathcal{T}(z_c) - z_{\text{ap}}\|^2, \tag{4}$$

where $\odot$ denotes element-wise product, $\mathbb{1}$ is the indicator function and $\mathcal{T}$ is the stop-gradient operator.

Since color tends to have more discontinuities than depth in regions such as textures, we include the remaining high-frequency positional encoding components of $\mathbf{p}_i$ in the input for $\mathcal{F}_2$ as

$$\mathbf{c}_i = \mathcal{F}_2^{\text{ap}}(\mathbf{h}_i, \gamma(\mathbf{p}_i, l_p^{\text{ap}}, l_p), \gamma(\mathbf{v}_i, 0, l_v)). \tag{5}$$

Note that $\mathbf{h}_i$ already includes the low-frequency positional encoding components of $\mathbf{p}_i$.

4.1.2 Shape-Radiance Ambiguity. The ability of the NeRF to predict view-dependent radiance helps it learn non-Lambertian surfaces. With fewer images, the NeRF can simply learn any random geometry and change the color of 3D points in accordance with the input viewpoint to explain away the observed images [Zhang et al. 2020] To bias the NeRF against this, we disable the view-dependent radiance in the second augmented NeRF model to output color based on $\mathbf{p}_i$ alone. If z_{av} is the depth estimated by the augmented model, we impose the depth supervision as

$$\mathcal{L}_{\text{av}} = \mathbb{1}_{\{m_{\text{av}}=1\}} \odot \|z_c - \mathcal{T}(z_{\text{av}})\|^2 + \mathbb{1}_{\{m_{\text{av}}=-1\}} \odot \|\mathcal{T}(z_c) - z_{\text{av}}\|^2, \tag{6}$$

where m_{av} is a ternary mask indicating where the depths estimated by the augmented and the main model are reliable. We note that while the augmented model is more accurate in Lambertian regions, the main model is better equipped to handle specular objects.

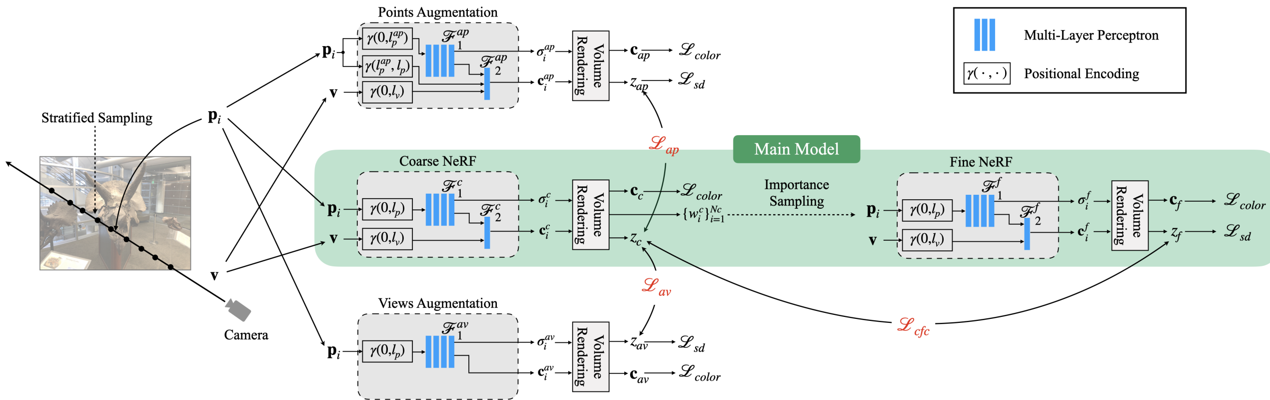


Figure 2: Architecture of SimpleNeRF. We train two augmented NeRF models in tandem with the NeRF to obtain depth supervision. In points augmentation, we reduce the positional encoding frequencies input to $\mathcal{F}_1$ and concatenate them to the input of $\mathcal{F}_2$. For views augmentation, we ask $\mathcal{F}_1$ to output both volume density and color based on position alone. We add depth supervision losses $\mathcal{L}_{\text{ap}}$ and $\mathcal{L}_{\text{av}}$ between the coarse NeRFs of the main and augmented models and a consistency loss $\mathcal{L}_{\text{cfc}}$ between the coarse and fine NeRFs of the main model. During inference, only the Main Model is employed.

Table 1: Quantitative results on LLFF dataset.

	2 views			3 views			4 views		
Model	**LPIPS ↓**	**SSIM ↑**	**PSNR ↑**	**LPIPS ↓**	**SSIM ↑**	**PSNR ↑**	**LPIPS ↓**	**SSIM ↑**	**PSNR ↑**
RegNeRF	0.3056	0.5712	18.52	0.2908	0.6334	20.22	0.2794	0.6645	21.32
FreeNeRF	**0.2638**	0.6322	19.52	0.2754	0.6583	20.93	0.2848	0.6764	21.91
DS-NeRF	0.3106	0.5862	18.24	0.3031	0.6321	20.20	0.2979	0.6582	21.23
DDP-NeRF	0.2851	0.6218	18.73	0.3250	0.6152	18.73	0.3042	0.6558	20.17
ViP-NeRF	0.2768	0.6225	18.61	0.2798	0.6548	20.54	0.2854	0.6675	20.75
SimpleNeRF	0.2688	**0.6501**	**19.57**	**0.2559**	**0.6940**	**21.37**	**0.2633**	**0.7016**	**21.99**

4.2 Determining Reliable Depth Estimates

We follow similar procedures to determine the masks m_{ap} and m_{av} and hence explain the mask computation using a generic variable m_a to denote either of the masks. Given the depths z_c and z_a estimated by the main and augmented models respectively for pixel $\mathbf{q}$, we reproject a $k \times k$ patch around $\mathbf{q}$ to the nearest training view using both z_c and z_a. We compute the MSE in intensities between the reprojected patch and the corresponding patch in the training image and choose the depth corresponding to lower MSE as the reliable depth. To filter out the cases where both the main and augmented models predict incorrect depth, we define a threshold e_τ and mark the depth to be reliable if its corresponding MSE is also less than e_τ. If e_c and e_a are the reprojection MSE corresponding to z_c and z_a respectively, we compute the mask as

$$m_a = \begin{cases} 1 & \text{if } (e_a \leq e_c) \text{ and } (e_a \leq e_\tau) \\ -1 & \text{if } (e_c < e_a) \text{ and } (e_c \leq e_\tau) \\ 0 & \text{otherwise .} \end{cases} \tag{7}$$

For specular regions, the reprojected patches may not match in intensities leading to m_a being zero. This implies supervision for fewer pixels and not supervision with incorrect depth estimates. We note that although both the views augmentation and the depth reliability estimation work only in the non-specular regions, there is no redundancy in the model. The views augmentation model may still make errors in the non-specular regions, and hence it is necessary to determine the reliability of its depth estimates.

4.3 Hierarchical Sampling

Since multiple solutions can explain the observed images in the few-shot setting, the coarse and fine MLP may converge to different depth estimates for a given pixel as shown in Fig. 1c. Thus, dense sampling may not be employed around the region where fine NeRF predicts the object surface, which is equivalent to using only the coarse sampling for the fine NeRF. This can lead to blur in rendered images as seen in Fig. 1d. To prevent such inconsistencies, we drive the two NeRFs to be consistent in their solutions by imposing an MSE loss between the depths predicted by the two NeRFs. If z_c and z_f are the depths estimated by the coarse and fine NeRFs respectively, we define the coarse-fine consistency loss as

$$\mathcal{L}_{\text{cfc}} = \mathbb{1}_{\{m_{\text{cfc}}=1\}} \odot \|z_c - \mathcal{T}(z_f)\|^2 + \mathbb{1}_{\{m_{\text{cfc}}=-1\}} \odot \|\mathcal{T}(z_c) - z_f\|^2, \tag{8}$$

where the mask m_{cfc} is determined as described in Sec. 4.2.

Table 2: Quantitative results on RealEstate-10K dataset.

	2 views			3 views			4 views		
Model	**LPIPS ↓**	**SSIM ↑**	**PSNR ↑**	**LPIPS ↓**	**SSIM ↑**	**PSNR ↑**	**LPIPS ↓**	**SSIM ↑**	**PSNR ↑**
RegNeRF	0.4129	0.5916	17.14	0.4171	0.6132	17.86	0.4316	0.6257	18.34
FreeNeRF	0.5036	0.5354	14.70	0.4635	0.5708	15.26	0.5226	0.6027	16.31
DS-NeRF	0.2709	0.7983	26.26	0.2893	0.8004	26.50	0.3103	0.7999	26.65
DDP-NeRF	0.1290	0.8640	27.79	0.1518	0.8587	26.67	0.1563	0.8617	27.07
ViP-NeRF	0.0687	0.8889	32.32	0.0758	0.8967	31.93	0.0892	0.8968	31.95
SimpleNeRF	**0.0635**	**0.8942**	**33.10**	**0.0726**	**0.8984**	**33.21**	**0.0847**	**0.8987**	**32.88**

Table 3: Evaluation of depth estimated by different models with two input views. The reference depth is obtained using NeRF with dense input views.

	LLFF		RealEstate-10K	
model	**MAE ↓**	**SROCC ↑**	**MAE ↓**	**SROCC ↑**
DS-NeRF	0.2074	0.7230	0.7164	0.6660
DDP-NeRF	0.2048	0.7480	0.4831	0.7921
ViP-NeRF	0.1999	0.7344	0.3856	0.8446
SimpleNeRF	**0.1420**	**0.8480**	**0.3269**	**0.9215**

4.4 Overall Loss

We impose the pixel color reconstruction loss on the main model and the augmented models as

$$\mathcal{L}_{\text{color}} = \|\mathbf{c}_c - \hat{\mathbf{c}}\|^2 + \|\mathbf{c}_f - \hat{\mathbf{c}}\|^2 + \|\mathbf{c}_{\text{ap}} - \hat{\mathbf{c}}\|^2 + \|\mathbf{c}_{\text{av}} - \hat{\mathbf{c}}\|^2, \quad (9)$$

where the subscripts c, f, 'ap', and 'av' denote the outputs of the coarse NeRF, fine NeRF, points augmentation model, and the views augmentation model respectively. We also include the sparse depth loss on the models as,

$$\mathcal{L}_{\text{sd}} = \|z_f - \hat{z}\|^2 + \|z_{\text{ap}} - \hat{z}\|^2 + \|z_{\text{av}} - \hat{z}\|^2, \quad (10)$$

where $\hat{z}$ is the sparse depth given by the SfM model. We do not impose $\mathcal{L}_{\text{sd}}$ on the coarse NeRF of the main model following Deng et al. [2022b]. Our final loss is a combination of all the losses as

$$\mathcal{L} = \lambda_1\mathcal{L}_{\text{color}} + \lambda_2\mathcal{L}_{\text{sd}} + \lambda_3\mathcal{L}_{\text{ap}} + \lambda_4\mathcal{L}_{\text{av}} + \lambda_5\mathcal{L}_{\text{cfc}} \quad (11)$$

5 EXPERIMENTS

5.1 Evaluation Setup

We compare the performance of sparse input NeRF models on LLFF [Mildenhall et al. 2019] and RealEstate-10K [Zhou et al. 2018] datasets with 2, 3, and 4 input views. We assume the camera parameters are known for the input images, since in applications such as robotics or extended reality, external sensors or pre-calibrated set of cameras may provide the camera poses. We follow prior work [Somraj and Soundararajan 2023] to choose the train and test images. We provide more details in the supplementary.

We quantitatively evaluate the predicted frames from various models using peak signal to noise ratio (PSNR), structural similarity (SSIM) [Wang et al. 2004] and LPIPS [Zhang et al. 2018] measures. We employ depth mean absolute error (MAE) and spearman rank order correlation coefficient (SROCC) to evaluate the models on their ability to predict absolute and relative depth in novel views. We train a NeRF model with dense input views and use its depth predictions as pseudo ground truth. On the LLFF dataset, we normalize the predicted depths by the median ground truth depth, since the scenes have different depth ranges. Following RegNeRF [Niemeyer et al. 2022], we evaluate the predictions only in the regions of interest. We mask out the regions of test frames which are not visible in the train frames. To determine such regions, we use the depth estimated by a NeRF trained with dense input views and compute the visible region mask through reprojection error in depth. We provide more details on the mask computation in the supplementary along with the unmasked evaluation scores for reference.

5.2 Comparisons

We evaluate the performance of our model against various sparse input NeRF models on both datasets. We compare with DS-NeRF [Deng et al. 2022b], DDP-NeRF [Roessle et al. 2022] and RegNeRF [Niemeyer et al. 2022] which regularize the depth estimated by the NeRF. Further, we include two recent models, FreeNeRF [Yang et al. 2023] and ViP-NeRF [Somraj and Soundararajan 2023], among the comparisons. We train the models on both datasets using the codes provided by the respective authors. Implementation details of our model are provided in the supplement.

5.3 Results

Tabs. 1 and 2 show the view-synthesis performance of SimpleNeRF and other prior art on LLFF and RealEstate-10K datasets. We find that SimpleNeRF achieves state-of-the-art performance on both datasets in most cases. The higher performance of all the models on the RealEstate-10K dataset is perhaps due to the scenes being simpler. Hence, the performance improvement is also smaller as compared to the LLFF dataset. On RealEstate-10K, we observe that all the models struggle on one of the five scenes as compared to the other scenes. Excluding this scene, with two input views, SimpleNeRF improves SSIM over ViP-NeRF from 0.9596 to 0.9685, which we believe is a significant improvement at such high quality regime. In Tab. 2, we show the average performance on all five scenes and show the per-scene performance of various models in the supplementary.

Fig. 3 shows predictions of various models on an example scene from the RealEstate-10K dataset, where we observe that SimpleNeRF is the best in reconstructing the novel view. Figs. 5 to 11 show more comparisons on both datasets. Further, SimpleNeRF improves significantly in estimating the depth of the scene as seen in Tab. 4

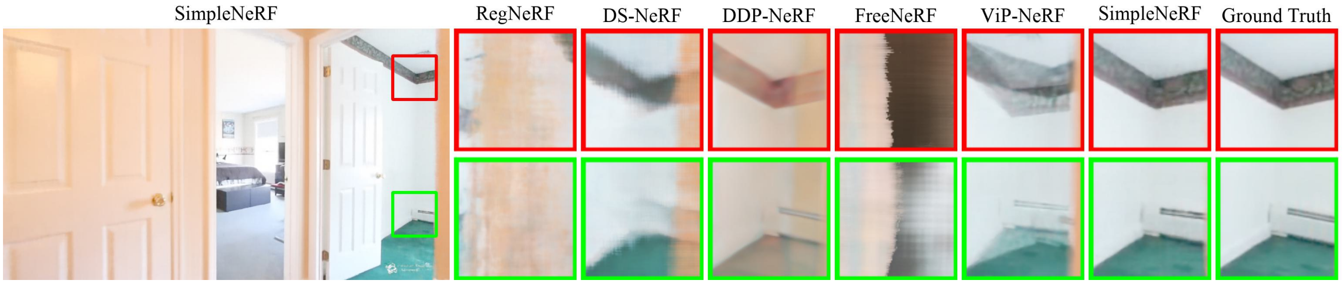


Figure 3: Qualitative examples on RealEstate-10K dataset with three input views. SimpleNeRF predictions are closest to the ground truth among all the models. In particular, DDP-NeRF predictions have a different shade of color and ViP-NeRF suffers from shape-radiance ambiguity creating ghosting artifacts.

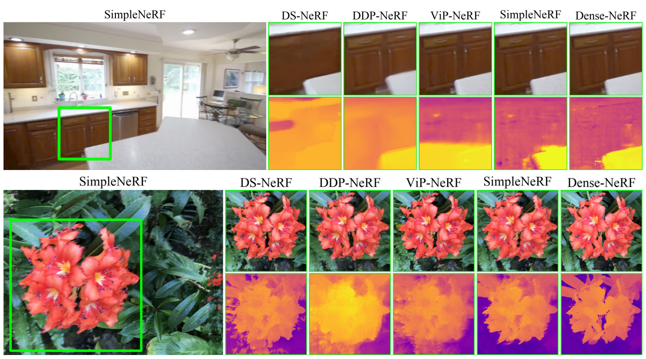


Figure 4: Estimated depth maps on RealEstate-10K and LLFF datasets with two input views. In both examples, the two rows show the predicted images and the depths respectively. We find that SimpleNeRF is significantly better at estimating the scene depth. Also, DDP-NeRF synthesizes the left table edge at a different angle due to incorrect depth estimation.

and Fig. 4. SimpleNeRF also performs significantly better in estimating relative depth, even outperforming ViP-NeRF which uses a prior based on relative depth. Estimating better geometry may be more crucial in downstream applications such as 3D scene editing. We provide video comparisons in the supplementary.

5.3.1 Ablations. We test the importance of each of the components of our model components by disabling them one at a time. We disable the points and views augmentations and coarse-fine consistency loss individually. When disabling $\mathcal{L}_{\text{cfc}}$, we additionally add augmentations to the fine NeRF since the knowledge learned by coarse NeRF may not propagate to the fine NeRF. We also analyze the need to supervise with only the reliable depth estimates by disabling the mask and stop-gradients in $\mathcal{L}_{\text{ap}}$, $\mathcal{L}_{\text{av}}$, and $\mathcal{L}_{\text{cfc}}$. Tab. 4 shows a quantitative comparison between the ablated models.

We observe that each of the components is crucial and disabling any of them leads to a drop in the performance. Further, using all the depths for supervision instead of only the reliable depths leads to a significant drop in performance. Finally, disabling $\mathcal{L}_{\text{cfc}}$ also leads to a drop in performance in addition to increasing the training time by almost 2× due to the inclusion of augmentations for the fine NeRF.

Table 4: Ablation experiments on both datasets with two input views.

	RealEstate-10K		LLFF	
model	**LPIPS ↓**	**MAE ↓**	**LPIPS ↓**	**MAE ↓**
SimpleNeRF	**0.0635**	**0.33**	**0.2688**	**0.14**
w/o points augmentation	0.0752	0.38	0.2832	0.15
w/o views augmentation	0.0790	0.39	0.2834	0.15
w/o coarse-fine consistency	0.0740	0.42	0.3002	0.19
w/o reliable depth	0.0687	0.45	0.3020	0.22
w/ identical augmentations	0.0777	0.40	0.2849	0.15
w/ smaller n/w as points aug	0.0740	0.38	0.2849	0.15

We conduct two additional experiments to further analyze the impact of the augmentation we design. Firstly, we analyze if there is a need to design simpler augmentations by replacing our novel augmentations with identical replicas of the NeRF as augmentations. In the second experiment, we use a smaller network, by reducing the number of layers from eight to four, for the points augmentation instead of using fewer positional encoding frequencies. Tab. 4 shows that in both the above cases, the performance drops to that of SimpleNeRF without the points augmentation or lower. In particular, reducing positional encoding frequencies has a higher impact than reducing the number of network layers perhaps because the network with fewer layers may still be capable of learning floaters on account of using all the positional encoding frequencies. In the supplement, we analyze how the performance of SimpleNeRF varies as l_p^{ap} varies.

5.4 Limitations

Our approach to determining reliable depth estimates for supervision depends on the reprojection error, which may be high for specular objects even if the depth estimates are correct. It may be helpful to explore approaches to determine the reliability of depth estimates without employing the reprojection error. The shape of the reprojected patches to determine reliable depth estimates may change significantly if the input viewpoints are diverse. This can lead to incorrect mask estimation. Further, our model requires accurate camera poses of the sparse input images. Finally, the use of augmented models adds to computational and memory overhead during training by about 1.5 times.

Figure 5: Qualitative examples on the RealEstate-10K dataset with two input views. While DDP-NeRF predictions contain blurred regions, ViP-NeRF predictions are color-saturated in certain regions of the door. SimpleNeRF does not suffer from these distortions and synthesizes a clean frame. For reference, we also show the input images.

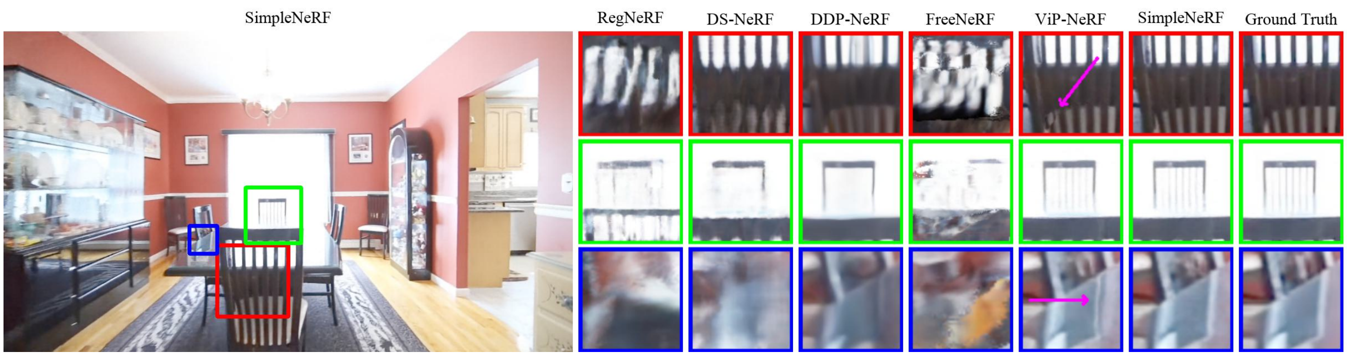


Figure 6: Qualitative examples on the RealEstate-10K dataset with four input views. We find that SimpleNeRF and ViP-NeRF perform the best among all the models. However, ViP-NeRF predictions contain minor distortions as pointed out by the magenta arrow, which is rectified by SimpleNeRF.

6 CONCLUSION

We address the problem of few-shot NeRF by obtaining depth supervision through simpler augmented models that are trained in tandem with the NeRF. We design two augmentations that learn simpler solutions and help the main NeRF model mitigate floater artifacts and shape-radiance ambiguity. By imposing a consistency loss between the coarse and fine NeRFs, we ensure better application of hierarchical sampling leading to sharper predictions. SimpleNeRF achieves state-of-the-art performance on two commonly used datasets in synthesizing novel views as well as estimating better scene geometry. In future, we plan to explore the role of simpler augmentations for newer models such as instant-ngp [Müller et al. 2022] and related applications such as neural surfaces [Yariv et al. 2021] to train them with sparse input views.

ACKNOWLEDGMENTS

This work was supported in part by a grant from Qualcomm. The first author was supported by the Prime Minister's Research Fellowship awarded by the Ministry of Education, Government of India. The authors would also like to thank Shankhanil Mitra for valuable discussions.

REFERENCES

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. 2023. NoPe-NeRF: Optimising Neural Radiance Field with No Pose Prior. (June 2023).

Matteo Bortolon, Alessio Del Bue, and Fabio Poiesi. 2022. Data augmentation for NeRF: a geometric consistent solution based on view morphing. *arXiv e-prints*, Article arXiv:2210.04214 (2022), arXiv:2210.04214 pages. arXiv:2210.04214

Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. 2021. MVSNeRF: Fast Generalizable Radiance Field Reconstruction from Multi-View Stereo. *arXiv e-prints*, Article arXiv:2103.15595 (March 2021), arXiv:2103.15595 pages. arXiv:2103.15595

Di Chen, Yu Liu, Lianghua Huang, Bin Wang, and Pan Pan. 2022. GeoAug: Data Augmentation for Few-Shot NeRF with Geometry Constraints. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Shenchang Eric Chen and Lance Williams. 1993. View Interpolation for Image Synthesis. In *Proceedings of the Computer Graphics and Interactive Techniques (SIGGRAPH)*. https://doi.org/10.1145/166117.166153

Yuedong Chen, Haofei Xu, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. 2023. Explicit Correspondence Matching for Generalizable Neural Radiance

Fields. *arXiv e-prints*, Article arXiv:2304.12294 (2023), arXiv:2304.12294 pages. arXiv:2304.12294

Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. 2021. Stereo Radiance Fields (SRF): Learning View Synthesis for Sparse Views of Novel Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. 2022b. Depth-Supervised NeRF: Fewer Views and Faster Training for Free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Nianchen Deng, Zhenyi He, Jiannan Ye, Budmonde Duinkharjav, Praneeth Chakravarthula, Xubo Yang, and Qi Sun. 2022a. FoV-NeRF: Foveated Neural Radiance Fields for Virtual Reality. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 28, 11 (2022), 3854–3864. https://doi.org/10.1109/TVCG.2022.3203102

Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. 1996. The Lumigraph. In *Proceedings of the Computer Graphics and Interactive Techniques (SIGGRAPH)*. https://doi.org/10.1145/237170.237200

Ajay Jain, Matthew Tancik, and Pieter Abbeel. 2021. Putting NeRF on a Diet: Semantically Consistent Few-Shot View Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. 2022. GeoNeRF: Generalizing NeRF With Geometry Priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhanghan Ke, Daoye Wang, Qiong Yan, Jimmy Ren, and Rynson W.H. Lau. 2019. Dual Student: Breaking the Limits of the Teacher in Semi-Supervised Learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Mijeong Kim, Seonguk Seo, and Bohyung Han. 2022. InfoNeRF: Ray Entropy Minimization for Few-Shot Neural Volume Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Minseop Kwak, Jiuhn Song, and Seungryong Kim. 2023. GeCoNeRF: Few-shot Neural Radiance Fields via Geometric Consistency. *arXiv e-prints*, Article arXiv:2301.10941 (2023), arXiv:2301.10941 pages. arXiv:2301.10941

SeokYeong Lee, JunYong Choi, Seungryong Kim, Ig-Jae Kim, and Junghyun Cho. 2023. ExtremeNeRF: Few-shot Neural Radiance Fields Under Unconstrained Illumination. *arXiv e-prints*, Article arXiv:2303.11728 (2023), arXiv:2303.11728 pages. arXiv:2303.11728

Marc Levoy and Pat Hanrahan. 1996. Light Field Rendering. In *Proceedings of the Computer Graphics and Interactive Techniques (SIGGRAPH)*. https://doi.org/10.1145/237170.237199

Kai-En Lin, Yen-Chen Lin, Wei-Sheng Lai, Tsung-Yi Lin, Yi-Chang Shih, and Ravi Ramamoorthi. 2023. Vision Transformer for NeRF-Based View Synthesis From a Single Input Image. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*.

Li Ma, Xiaoyu Li, Jing Liao, Qi Zhang, Xuan Wang, Jue Wang, and Pedro V. Sander. 2022. Deblur-NeRF: Neural Radiance Fields From Blurry Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Willi Menapace, Aliaksandr Siarohin, Stéphane Lathuilière, Panos Achlioptas, Vladislav Golyanik, Elisa Ricci, and Sergey Tulyakov. 2023. Plotting Behind the Scenes: Towards Learnable Game Engines. *arXiv e-prints*, Article arXiv:2303.13472 (2023), arXiv:2303.13472 pages. arXiv:2303.13472

Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P. Srinivasan, and Jonathan T. Barron. 2022. NeRF in the Dark: High Dynamic Range View Synthesis From Noisy Raw Images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. 2019. Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines. *ACM Transactions on Graphics (TOG)* 38, 4 (July 2019), 1–14. https://doi.org/10.1145/3306346.3322980

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Transactions on Graphics (ToG)* 41, 4 (2022), 1–15.

Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. 2022. RegNeRF: Regularizing Neural Radiance Fields for View Synthesis From Sparse Inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Naama Pearl, Tali Treibitz, and Simon Korman. 2022. NAN: Noise-Aware NeRFs for Burst-Denoising. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Barbara Roessle, Jonathan T. Barron, Ben Mildenhall, Pratul P. Srinivasan, and Matthias Nießner. 2022. Dense Depth Priors for Neural Radiance Fields From Sparse Input Views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Seunghyeon Seo, Donghoon Han, Yeonjin Chang, and Nojun Kwak. 2023. MixNeRF: Modeling a Ray with Mixture Density for Novel View Synthesis from Sparse Inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yue Shi, Dingyi Rong, Bingbing Ni, Chang Chen, and Wenjun Zhang. 2022. GARF: Geometry-Aware Generalized Neural Radiance Field. *arXiv e-prints*, Article arXiv:2212.02280 (2022), arXiv:2212.02280 pages. arXiv:2212.02280

Nagabhushan Somraj and Rajiv Soundararajan. 2023. ViP-NeRF: Visibility Prior for Sparse Input Neural Radiance Fields. In *Proceedings of the ACM Special Interest Group on Computer Graphics and Interactive Techniques (SIGGRAPH)*. https://doi.org/10.1145/3588432.3591539

Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P. Srinivasan, Jonathan T. Barron, and Ren Ng. 2021. Learned Initializations for Optimizing Coordinate-Based Neural Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Alex Trevithick and Bo Yang. 2021. GRF: Learning a General Radiance Field for 3D Representation and Rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Mikaela Angelina Uy, Ricardo Martin-Brualla, Leonidas Guibas, and Ke Li. 2023. SCADE: NeRFs from Space Carving with Ambiguity-Aware Depth Estimates. (June 2023).

Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. 2022. Ref-NeRF: Structured View-Dependent Appearance for Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. https://doi.org/10.1109/CVPR52688.2022.00541

Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. 2023. SparseNeRF: Distilling Depth Ranking for Few-shot Novel View Synthesis. *arXiv e-prints*, Article arXiv:2303.16196 (2023), arXiv:2303.16196 pages. arXiv:2303.16196

Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P. Srinivasan, Howard Zhou, Jonathan T. Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. 2021. IBRNet: Learning Multi-View Image-Based Rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Jamie Wynn and Daniyar Turmukhambetov. 2023. DiffusioNeRF: Regularizing Neural Radiance Fields with Denoising Diffusion Models. *arXiv e-prints*, Article arXiv:2302.12231 (2023), arXiv:2302.12231 pages. arXiv:2302.12231

Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. 2022. SinNeRF: Training Neural Radiance Fields on Complex Scenes from a Single Image. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Jiawei Yang, Marco Pavone, and Yue Wang. 2023. FreeNeRF: Improving Few-shot Neural Rendering with Free Frequency Regularization. (June 2023).

Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. 2021. Volume Rendering of Neural Implicit Surfaces. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelNeRF: Neural Radiance Fields From One or Few Images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. 2022. NeRF-Editing: Geometry Editing of Neural Radiance Fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

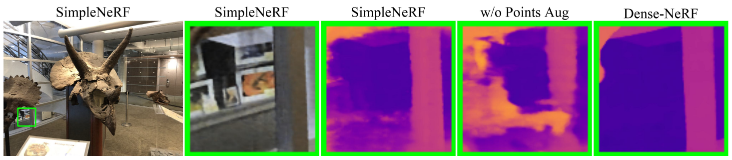
Jason Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. 2021. NeRS: Neural Reflectance Surfaces for Sparse-view 3D Reconstruction in the Wild. In *Proceedings of the Advances in Neural Information Processing Systems (NeurIPS)*.

Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. 2020. NeRF++: Analyzing and Improving Neural Radiance Fields. *arXiv e-prints*, Article arXiv:2010.07492 (2020), arXiv:2010.07492 pages. arXiv:2010.07492

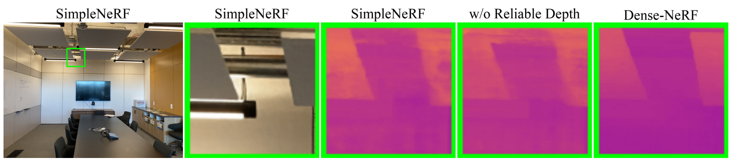
Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. 2018. Stereo Magnification: Learning View Synthesis Using Multiplane Images. *ACM Transactions on Graphics (TOG)* 37, 4 (July 2018).

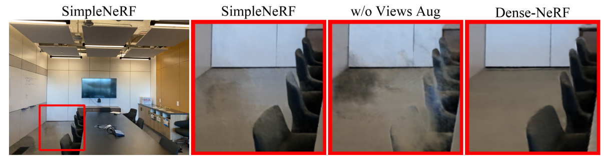
Bingfan Zhu, Yanchao Yang, Xulong Wang, Youyi Zheng, and Leonidas Guibas. 2023. VDN-NeRF: Resolving Shape-Radiance Ambiguity via View-Dependence Normalization. *arXiv e-prints*, Article arXiv:2303.17968 (2023), arXiv:2303.17968 pages. arXiv:2303.17968



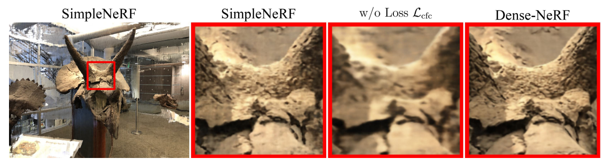
(a) Without Points Augmentation: The ablated model introduces floaters that are significantly reduced by using the points augmentation model.



(c) Without Reliability of Depth Supervision: The points augmentation model struggles to learn sharp depth discontinuities at true depth edges. Supervising the main model using such depths without determining their reliability causes the main model to learn incorrect depth. As a result, the ablated model fails to learn sharp depth discontinuities at certain regions.



(b) Without Views Augmentation: The ablated model suffers from shape-radiance ambiguity and produces ghosting artifacts.



(d) Without Coarse-Fine Consistency: We observe that while SimpleNeRF predictions are sharper, the ablated model without coarse-fine consistency loss, $\mathcal{L}_{\text{cfc}}$ produces blurred renders. This is similar to Fig. 1c, where we observe DS-NeRF also produce blurred renders.

Figure 7: Qualitative examples for ablated models on the LLFF dataset with two input views. We also show the outputs of the dense-input NeRF for reference.

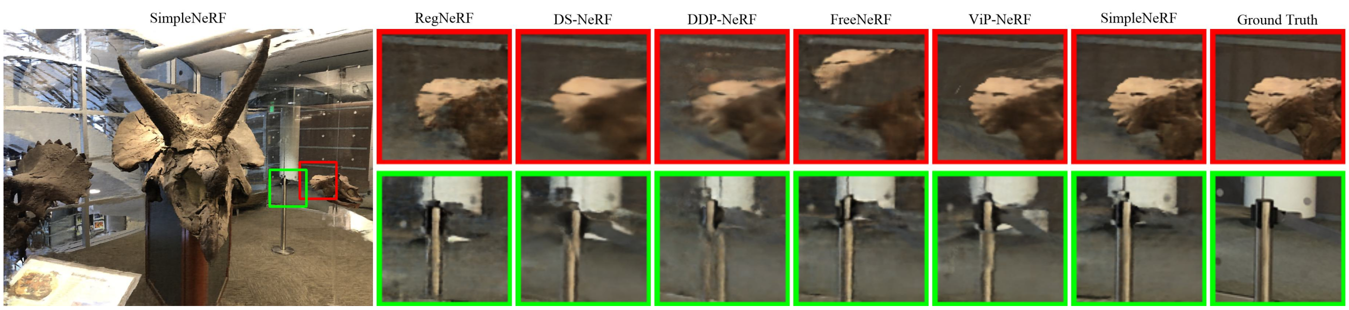


Figure 8: Qualitative examples on the LLFF dataset with two input views. DDP-NeRF and ViP-NeRF synthesize frames with broken objects in the second row and FreeNeRF breaks the object in the first row, due to incorrect depth estimations. SimpleNeRF produces sharper frames devoid of such artifacts.

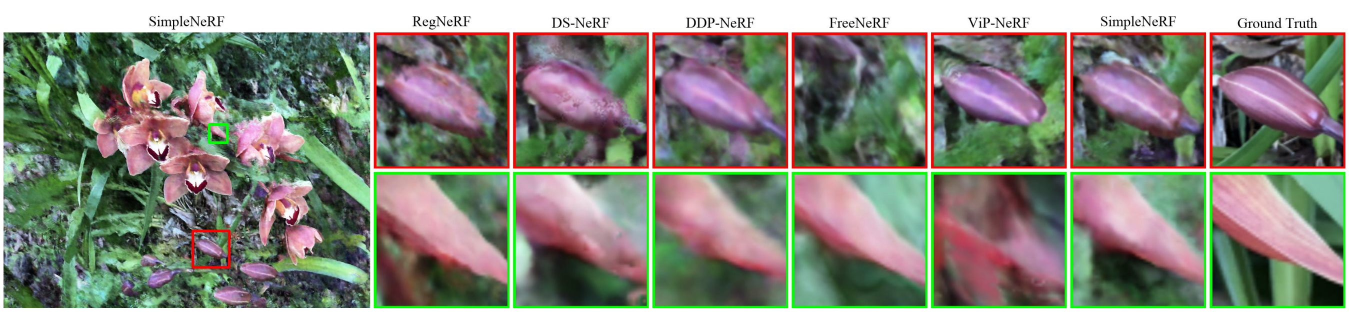


Figure 9: Qualitative examples on the LLFF dataset with three input views. In the first row, the orchid is displaced out of the cropped box in the FreeNeRF prediction, due to incorrect depth estimation. ViP-NeRF and RegNeRF fail to predict the complete orchid accurately and contain distortions at either ends. In the second row, ViP-NeRF prediction contains severe distortions. SimpleNeRF reconstructs the best among all the models in both examples.

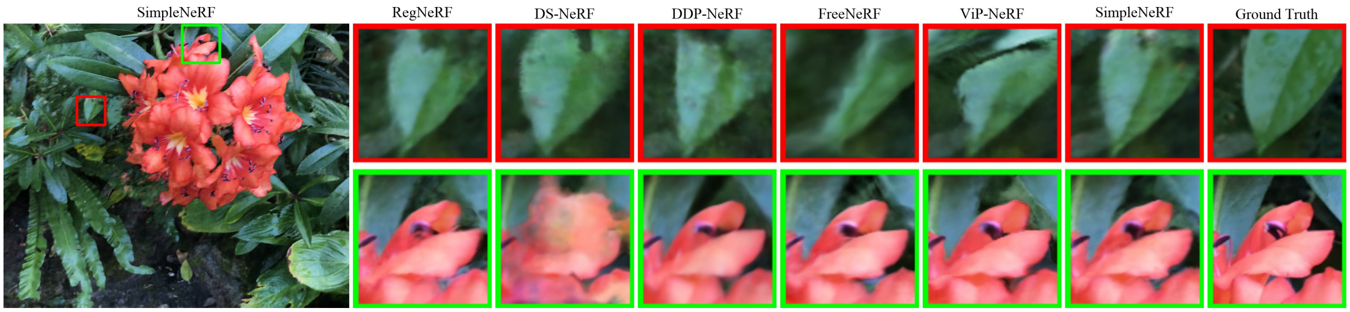


Figure 10: Qualitative examples on the LLFF dataset with four input views. In the first row, we find that ViP-NeRF, FreeNeRF, and DDP-NeRF struggle to reconstruct the shape of the leaf accurately. In the second row, DS-NeRF introduces floaters. SimpleNeRF does not suffer from such artifacts and reconstructs the shapes better.

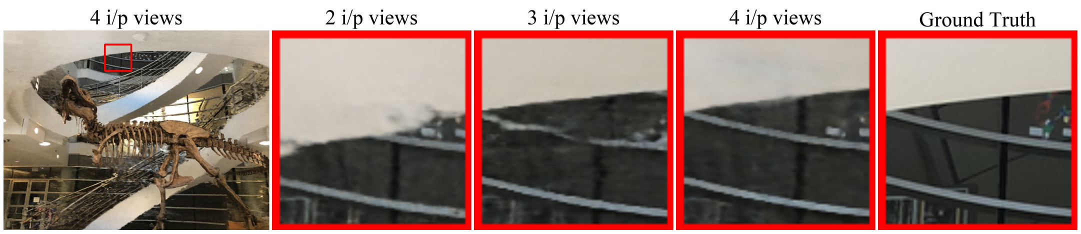


Figure 11: Qualitative examples on the LLFF dataset with two, three, and four input views. We observe errors in depth estimation with two input views causing a change in the position of the roof. While this is corrected with three input views, there are a few shape distortions in the metal rods. With four input views, even such distortions are corrected.