GECO-NERF: NEURAL RADIANCE FIELDS WITH GEOMETRIC CONSISTENCY FOR FEW-SHOT NOVEL VIEW SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

We present a novel framework to regularize Neural Radiance Field (NeRF) with geometry-based consistency regularization in a few-shot setting. The proposed approach leverages a rendered depth map at unobserved viewpoint to warp input images to unobserved viewpoint and impose them as pseudo ground-truths to facilitate learning of NeRF. By encouraging consistency at a feature-level instead of using pixel-level reconstruction loss, we regularize the network at semantic and structural levels while allowing view-dependent radiance to model freely after color variations. We also demonstrate an effective method to filter out erroneous warped solutions, along with training strategies to stabilize training during optimization. We show that our model achieves competitive results compared to concurrent few-shot NeRF models.

1 INTRODUCTION

Representing 3D scene as Neural Radiance Field (NeRF) (Mildenhall et al., 2020) has recently proven to be a powerful approach for novel view synthesis and 3D reconstruction (Barron et al., 2021; Jain et al., 2021; Chen et al., 2021).

Despite its impressive performance, NeRF retains a weakness of requiring great number of densely distributed calibrated images for optimization. When limited to sparse observations, NeRF easily overfits to input view images and struggles to produce correct geometry (Zhang et al., 2020). The task that directly addresses this problem, also called a few-shot novel view synthesis, aims to optimize high-fidelity neural radiance field in such sparse scenarios. Previous works attempted to solve this task by utilizing a semantic feature (Jain et al., 2021), entropy minimization (Kim et al., 2022), SfM depth priors (Deng et al., 2022) or normalizing flow (Niemeyer et al., 2022), but their necessity for handcrafted methods and inability to extract local and fine structures limited their applicability.

To alleviate these issues, we for the first time propose a novel regularization technique to enforce a geometric consistency across different views with depth-guided warping and consistency modeling. Based on this, we propose GeCo-NeRF, a novel framework for training robust neural radiance fields in a few-shot setting. Our key insight is that we can leverage geometry captured by NeRF to warp input images to unobserved viewpoints and use them as pseudo ground-truths to facilitate learning of fine details and high-frequency features. By encouraging images rendered at novel views to model warped images with a consistency loss, we can successfully constrain both geometry and appearance to boost fidelity of neural radiance fields even in highly under-constrained few-shot setting. Taking into consideration non-Lambertian nature of given datasets, we propose feature-level regularization loss that captures contextual and structural information while largely ignoring individual color differences. We also generate a consistency mask to prevent inconsistently warped information from harming the network. Finally, we provide coarse-to-fine training strategies for sampling and pose generation to stabilize optimization of the model.

We demonstrate the effectiveness of our method on synthetic and real datasets (Mildenhall et al., 2020; Jensen et al., 2014). Experimental results prove the effectiveness of the proposed model over the latest methods for few-shot novel view synthesis.

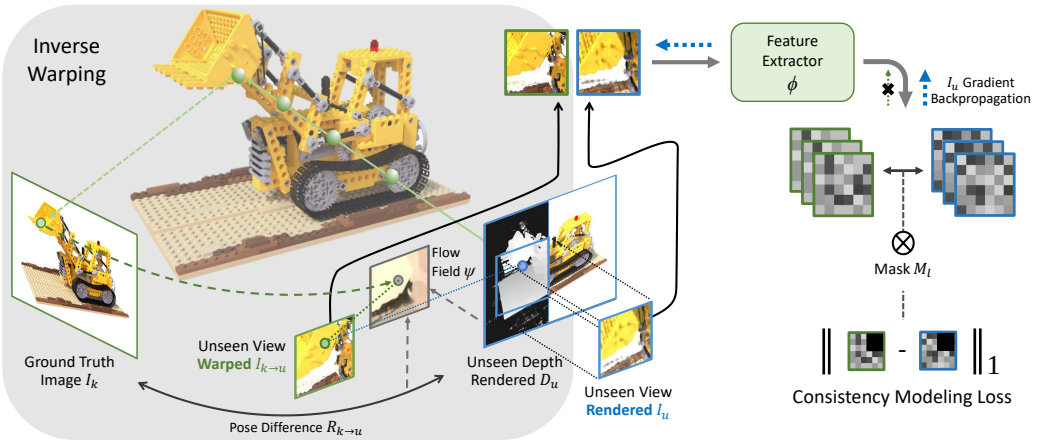


Figure 1: **Illustration of our consistency modeling pipeline for few-shot NerF.** Given a single image I_k and estimated depth map D_u of unobserved viewpoint, we can generate warped image $I_{k\to u}$ at novel viewpoint by establishing geometric correspondence between two viewpoints. Using the warped image as pseudo ground-truth, we cause rendered image of unseen viewpoint to be consistent in structure with warped image, with occlusions taken into consideration.

2 Related Works

Novel view synthesis. Novel view synthesis has been a longstanding problem in computer vision and computer graphics (Sitzmann et al., 2019; Yu et al., 2021b). The task can be defined as generating photo-realistic novel views from given known views. Conventionally, there were various handcrafted approaches. Some methods used direct interpolation of known image or light fields (Buehler et al., 2001; Chaurasia et al., 2013; Levoy & Hanrahan, 1996; Wood et al., 2000; Srinivasan et al., 2017) while other works utilized explicit 3D representations such as meshes or voxels (Buehler et al., 2001; Debevec et al., 1996; Wood et al., 2000; Waechter et al., 2014). More recently, differentiable rendering approaches (Sitzmann et al., 2019; Niemeyer et al., 2020) gained significant attention by its capability to learn neural volumetric representations from solely multi-view images.

Most notable of such approaches is Neural Radiance Field (NeRF) (Mildenhall et al., 2020), where photo-realistic images are rendered by a simple MLP architecture. Sparked by its impressive performance, a variety of follow-up studies based on its continuous neural volumetric representation have been prompted, including dynamic and deformable scenes (Park et al., 2021; Tretschk et al., 2021; Pumarola et al., 2021; Attal et al., 2021), real-time rendering (Yu et al., 2021a; Hedman et al., 2021; Reiser et al., 2021; Müller et al., 2022), self-calibration (Jeong et al., 2021) and generative modeling (Schwarz et al., 2020; Niemeyer & Geiger, 2021; Xu et al., 2021; Deng et al., 2021). Mip-NeRF (Barron et al., 2021) eliminates aliasing artifacts by adopting cone tracing with a single multi-scale MLP. In general, most of these works have difficulty in optimizing a single scene with a few number of images.

Few-shot NeRF. One key limitation of NeRF is its necessity for large number of calibrated views in optimizing neural radiance fields. Some recent works attempted to address this in cases where only few observed views of the scene are available. PixelNeRF(Yu et al., 2021b) conditions a NeRF on image inputs using local CNN features. This conditional model allows the network to learn scene priors across multiple scenes. Stereo radiance fields (Chibane et al., 2021) use local CNN features from input views for scene geometry reasoning and MVSNeRF (Chen et al., 2021) combines cost volume with neural radiance field for improved performance. However, multi-view images of various scenes are essential for these methods for pre-training.

Other works attempt different approach of optimizing a NeRF from scratch in few-shot settings: DSNeRF (Deng et al., 2022) makes use of depth supervision to network to optimize a scene with few images. (Roessle et al., 2021) also utilizes sparse depth prior by extending into dense depth map by depth completion module to guide network optimization. Contrary to aforementioned model using depth prior, DietNeRF (Jain et al., 2021) enforces semantic consistency between rendered images from unseen view and seen images. InfoNeRF (Kim et al., 2022) and RegNeRF (Niemeyer

et al., 2022) constrain geometry without relying on additional modules to generate novel views, as the aforementioned approaches do. RegNeRF (Niemeyer et al., 2022) regularizes the geometry and appearance of patches rendered from unobserved viewpoints. InfoNeRF (Kim et al., 2022) constrains the density's entropy in each ray and ensures consistency across rays in the neighborhood. However, the performance of these methods was still limited.

Monocular depth estimation. Monocular depth estimation aims to predict the dense depth of a single input image (Watson et al., 2021). In the absence of ground truth depth, one alternative is to train depth estimation models using image reconstruction as the supervisory signal (Zhou et al., 2017). A set of images as input, either in the form of stereo pairs (Garg et al., 2016; Wang et al., 2019; Godard et al., 2017) or monocular sequences (Godard et al., 2019; Zhou et al., 2017) is given to the model. By predicting the depth for a given image and projecting it into nearby views, the model is trained by minimizing the image reconstruction error. For instance, sfm learner (Zhou et al., 2017) proposed to leverage multi-view synthesis procedure, and demon (Ummenhofer et al., 2017) proposed to predict depth and camera motion together from unconstrained image pairs. However, this technique was never adopted in a novel view synthesis, which is the topic of this paper.

3 PRELIMINARIES

Neural Radiance Field (NeRF) (Mildenhall et al., 2020) represents a scene as a continuous function f_θ represented by a neural network with parameters θ, where the points are sampled along rays, represented by r, for evaluation by the neural network. Typically, the sampled coordinates $x \in \mathbb{R}^3$ and view direction $d \in \mathbb{R}^2$ are transformed by a positional encoding γ into Fourier features (Tancik et al., 2020) that facilitates learning of high-frequency details. The neural network $f_\theta : \mathbb{R}^{L_x} \times \mathbb{R}^{L_d} \rightarrow [0,1]^3 \times [0,\infty)$ takes as input the transformed coordinate $\gamma(x)$ and viewing directions $\gamma(d)$, and outputs a view-invariant density value $\sigma \in \mathbb{R}$ and a view-dependent color value $c \in \mathbb{R}^3$ such that

$$\{c, \sigma\} = f_\theta(\gamma(x), \gamma(d)). \tag{1}$$

With a ray parametrized as $r(t) = o + td$ from the camera center o through the pixel along direction d, the color is rendered as follows:

$$C(r) = \int_{t_n}^{t_f} T(t)\sigma(r(t))c(r(t), d)dt, \text{ where } T(t) = \exp\left(-\int_{t_n}^{t} \sigma(r(s))ds\right), \tag{2}$$

where $C(r)$ is a predicted color value along the ray $r(t)$ from t_n to t_f, and $T(t)$ denotes an accumulated transmittance along the ray from t_n to t. To optimize the networks f_θ, the observation loss $\mathcal{L}$ enforces the rendered color values to be consistent with ground-truth color value $C'(r)$:

$$\mathcal{L}_{\text{obs}} = \sum_{r \in \mathcal{R}} \|C'(r) - C(r)\|_2^2, \tag{3}$$

with $\mathcal{R}$ representing a batch of training rays.

4 METHODOLOGY

4.1 MOTIVATION AND OVERVIEW

As noted in previous works (Jain et al., 2021; Niemeyer et al., 2022), NeRF's poor performance in a few-shot setting stems from its tendency to overfit to sparse input views, assuming incorrect scene geometry. We observe that this divergent behaviour of NeRF contributes not only to collapse of its global scene geometry, but also its inability to recover fine details and high-frequency scene features. To overcome these issues, we leverage NeRF's predicted geometry to warp input images to unobserved viewpoints and use them as pseudo ground-truths for regularization of optimization process.

Let us denote an image at k-th viewpoint as I_k. In a few-shot novel view synthesis, NeRF is given only a few images $\{I_k\}$ for $k \in \{1, ..., K\}$ with small K, e.g., $K = 3$ or $K = 5$. The objective of novel view synthesis is to train the mapping function f_θ that can be used to recover an image I_u at u-th unseen viewpoint. As we described above, in the few-shot setting, given $\{I_k\}$, directly optimizing

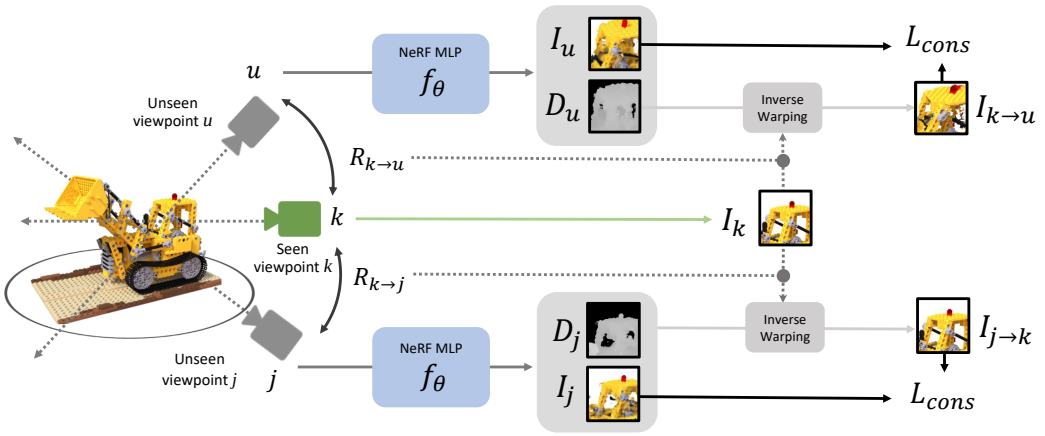


Figure 2: **Illustration of the proposed framework.** GeCo-NeRF regularizes the networks with consistency modeling. Consistency term $\mathcal{L}_{\text{cons}}$ is applied between unobserved viewpoint image and warped observed viewpoint image.

f_θ solely with the pixelwise reconstruction loss is limited by its inability to model view-dependent effects, and thus an additional regularization to encourage the network f_θ to generate consistent appearance and geometry is required.

To achieve this, we propose a novel regularization technique to enforce a geometrical consistency across different views with depth-guided warping and consistency modeling. We focus on the fact that NeRF (Mildenhall et al., 2020) inherently renders not only color image but depth image as well. Combined with known viewpoint difference, the rendered depths can be used to define a geometric correspondence relationship between two arbitrary views.

Specifically, we consider a depth image D_u at unseen viewpoint u. By formulating a warping function $\psi(I_k; D_u, R_{k\to u})$ that warps an image I_k according to the depth D_u and viewpoint difference $R_{k\to u}$, we can encourage a consistency between warped image $I_{k\to u} = \psi(I_k; D_u, R_{k\to u})$ and rendered image I_u at unseen viewpoint u, which in turn improves the few-shot novel view synthesis performance. This framework can overcome the limitations of previous few-shot setting approaches (Mildenhall et al., 2020; Chen et al., 2021; Barron et al., 2021), improving not only global geometry but also high-frequency details and appearance as well.

In the following, we first explain how input images can be warped to unseen viewpoints. Then, we demonstrate how we impose consistency upon the pair of warped image and rendered image for effective regularization, followed by explanation of occlusion handling method and several training strategies that proved crucial for stabilization of NeRF optimization in few-shot scenario.

4.2 Rendered Depth-Guided Warping

In order to address NeRF's poor performance in a few-shot setting, we introduce a novel regularization method to prevent it from overfitting to input viewpoint images and consequently recovering incorrect geometry. We leverage upon our key insight, which is that geometry predicted by few-shot NeRF can be exploited for warping fine details of input images to unobserved viewpoints in a 3D-aware, geometrically consistent manner.

To render an image at novel viewpoints, we first sample a random camera viewpoint, from which corresponding ray vectors are generated in a patch-wise manner. As NeRF outputs density and color values of sampled points along the novel rays, we use recovered density values to render a consistent depth map. Following (Mildenhall et al., 2020), we formulate per-ray depth values as weighted composition of distances traveled from origin. With a ray r corresponding to pixel p parametrized as $r_p(t) = o + td$, the depth rendering is defined similarly to the color rendering:

$$D(p) = \int_{t_n}^{t_f} T(t)\sigma(r_p(t))t dt, \tag{4}$$

(a) GT patch (b) Rendered patch (c) Warped patch (d) Occlusion mask (e) Masked patch

Figure 3: **Visualization of consistency modeling:** (a) ground-truth patch, (b) rendered patch, (c) warped patch, (d) occlusion mask, and (e) warped patch with occlusion masking.

where $D(p)$ is a predicted depth along the ray r_p. We use the rendered depth map D_u to warp input ground-truth image I_k to unseen viewpoint I_u and acquire a warped image $I_{k\to u}$, which is defined as a process such that $I_{k\to u} = \psi(I_k; D_u, R_{k\to u})$. More specifically, pixel location p_u in target unseen viewpoint image is transformed to $p_{u\to k}$ at source viewpoint image by viewpoint difference $R_{u\to k}$ and camera intrinsic parameter K such that

$$p_{u\to k} = KR_{u\to k}D_u(p_u)K^{-1}p_u, \tag{5}$$

where the projected coordinate $p_{u\to k}$ are continuous value. With a differentiable sampler, we extract color values of $p_{u\to k}$ on I_k. More formally, the transforming components process can be written as follows:

$$I_{k\to u}(p_u) = \text{sampler}(I_k; p_{u\to k}), \tag{6}$$

where $\text{sampler}(\cdot)$ is a bilinear sampling operator (Jaderberg et al., 2015).

To overcome the computational bottleneck of full rendering and warping, rays are sampled on a strided grid to make the patch with stride s, which we have set as 2. After the rays undergo volumetric rendering, we upsample the low-resolution depth map back to original resolution with bilinear interpolation. This full-resolution depth map is used for the inverse warping. This way, detailed warped patches of full-resolution can be generated with only a fraction of computational cost that would be required when rendering the original sized ray batch.

4.3 Training Losses

Feature-level consistency modeling. Given the rendered I_u at u-th viewpoint and the warped image $I_{k\to u}$ with depth D_u and viewpoint difference $R_{k\to u}$, we define the consistency between them to encourage additional regularization for globally consistent rendering. One viable option is to naïvely apply NeRF's pixel-wise image reconstruction loss such that

$$\mathcal{L}_C = \|I_{k\to u} - I_u\|. \tag{7}$$

However, we observe that this simple strategy is prone to cause failures in reflectant non-Lambertian surfaces where appearance changes greatly regarding viewpoints (Zhan et al., 2018). To overcome this, we propose a feature-level regularization loss that encourages structural consistency while ignoring view-dependent radiance effects.

Given an image I as an input, we use a convolutional network to extract multi-level feature maps such that $\phi_l(I) \in \mathbb{R}^{H_l\times W_l\times C_l}$ for l-th layer. To measure feature-level consistency between warped image $I_{k\to u}$ and rendered image I_u, we extract their features maps from L layers and compute difference within each feature map pairs that are extracted from the same layer. Formally written, the consistency loss $\mathcal{L}_{\text{cons}}$ is defined as such that

$$\mathcal{L}_{\text{cons}} = \sum_{l=1}^{L} \frac{1}{C_l m_l} \|M_l \odot (\phi_l(I_{k\to u}) - \phi_l(I_u))\|. \tag{8}$$

In accordance with the idea of using the latter as pseudo ground truths, we allow gradient backpropagation to pass only through the rendered image and block it for the warped image. By applying the consistency loss at multiple levels of feature maps, we cause I_u to model after $I_{k\to u}$ both on semantic and structural level. In addition, imperfect and distorted warpings tend to occur due to erroneous predictions in geometry, especially in early stages of optimization, which give the network misleading signals and degrade overall reconstruction quality. To prevent this, we construct

consistency mask M_l, whose sum of non-zero values amount to m_l, to let NeRF ignore regions with geometric inconsistencies. Instead of applying the mask before input into feature extractor network, we directly apply resized masks M_l to the feature maps after using nearest-neighbor down-sampling to make them match the dimensions of l-th layer outputs. Details regarding consistency masks will be further discussed in Sec. 4.4.

For this loss function $\mathcal{L}_{\text{cons}}$, we find l-1 distance function most suited for our task and utilize it to measure consistency across feature difference maps. Empirically, we have discovered that VGG-19 network (Simonyan & Zisserman, 2014) yields best performance in modeling consistencies, likely due to the absence of normalization layers (Johnson et al., 2016) that scale down absolute values of feature differences. Therefore, we employ VGG19 network as our feature extractor network ϕ throughout all of our models.

Our loss function differs from that of DietNeRF (Jain et al., 2021) in that while DietNeRF's consistency loss is limited to regularizing the radiance field in a globally semantic level, our loss combined with warping module is also able to give the network highly rich information on a local, structural level as well. Note that, contrary to DietNeRF giving only high-level feature consistency, our method of using multiple levels of convolutional network for feature difference calculation can be interpreted as enforcing a mixture of both high-level semantic and low-level structural consistency.

Image-aware disparity regularization. We further encourage local depth smoothness on rendered scenes by imposing l-1 penalty on disparity gradient ∂d within randomly sampled patches of input views. In addition, inspired by (Godard et al., 2017), we take into account the fact that depth discontinuities in depth maps are likely to be aligned to gradients of its color image, and introduce an edge-aware term with image gradients ∂I to weight the disparity values. Specifically, following (**?**), we regularize for image-aware depth smoothness

$$\mathcal{L}_{\text{reg}} = |\partial_x d_t^*| e^{-|\partial_x I_t|} + |\partial_y d_t^*| e^{-|\partial_y I_t|}, \tag{9}$$

where $d_t^* = d_t / \overline{d_t}$ is the mean-normalized inverse depth from (**?**) to discourage shrinking of the estimated depth.

We also optimize our model with a combined final loss of original NeRF's pixelwise reconstruction loss $\mathcal{L}_{\text{obs}}$ and two types of aforementioned consistency loss, $\mathcal{L}_{\text{cons}}$ and smoothness $\mathcal{L}_{\text{reg}}$, for unobserved view and observed view, respectively.

4.4 Occlusion Handling

Here we discuss the handling of misplaced warping with a binary occlusion mask M to prevent erroneous projections from harming the regularization process. We generate M by measuring consistency between rendered depth values from target viewpoint and source viewpoint such that

$$M(p_u) = \big[\|D_u(p_u) - D_k(p_{u\to k})\| < \tau\big]. \tag{10}$$

where $[\cdot]$ is Iverson bracket, and $p_{u\to k}$ refers to the corresponding pixel in source viewpoint k for reprojected target pixel p_u of viewpoint u. Here we measure euclidean distance between depth points rendered from target and source viewpoints as a criterion for a threshold masking. If distance between two points are greater than given threshold value τ, we determine two rays as rendering depths of separate surfaces generate mask for the corresponding pixel in unseen viewpoint I_u. The process takes place over every pixel in viewpoint I_u to generate a mask M the same size as rendered pixels.

As explained above, we do not directly apply the mask to image pairs, and resize them with nearest neighbor sampling to match the dimensions of mid-network feature maps and apply the mask to feature-level difference maps, generating M_l. Through this technique, we filter out problematic solutions at feature level and regularize NeRF with only high-confidence image features.

4.5 Training Strategy

In this section, we present novel training strategies to learn the rendering model with the proposed loss functions.

Strided patch sampling. We first present a progressive coarse-to-fine training strategy regarding ray sampling and pose generation to stabilize optimization of our model and facilitate learning of

fine details. We gradually decrease the size of our patches in stages of 180, 120 and 60 as training progresses, which has an effect of letting our model focus on high-level semantics in early stages and low-level fine details in later stages of training.

Progressive camera pose generation. The camera poses for unseen viewpoint regularization are also generated progressively. The camera location and poses are sampled on a half-sphere, with camera direction always directed towards its center. For progressive camera pose sampling, we sample noise value uniformly between an interval of [-β, +β] and add it to the original Euler rotation angles of input view poses. Parameter β grows linearly from 3 to 9 degrees throughout the course of optimization. This design choice can be intuitively understood as stabilizing locations near observed viewpoints at start and propagating this regularization to further locations, where warping becomes more difficult.

Positional encoding frequency annealing. We finally adopt coarse-to-fine frequency annealing strategy previously used by (Park et al., 2021) to regularize our optimization. This strategy forces our network to primarily optimize from coarse, low-frequency details where self-occlusions and fine features are minimized, easing the difficulty of warping process in the beginning stages of training. Following (Park et al., 2021), the annealing equation is $\alpha(t) = \frac{mt}{N}$ where we set hyper-parameter N as $15k$, and after said number of iterations our training progresses with fully open channels at standard setting.

5 EXPERIMENTS

In this section, we evaluate the performance of the proposed method and baselines for novel view synthesis. First, the experimental protocol, baselines, and benchmarks are explained. Next, we compare our model to other baselines on the standard benchmarks trained with few-shot input scenario and show that proposed method improves quality of synthesized novel views. Finally, we analyze components of our method and discuss the effect of the proposed approach on scene geometry and appearance.

5.1 EXPERIMENTAL SETTINGS

Datasets and metrics. We evaluate our model on two kinds of datasets, NeRF-Synthetic (**?**) and LLFF (**?**). NeRF-Synthetic is a realistically rendered 360° synthetic dataset comprised of 8 scenes. We randomly sample 3 viewpoints out of 100 training images in each scene, with 200 testing images for evaluation. We also conduct experiments on LLFF benchmark dataset introduced in (Mildenhall et al., 2019). Following RegNeRF (Niemeyer et al., 2022), we apply standard settings by selecting test set evenly from list of every 8th image and selecting 3 reference views from remaining images. We quantify novel view synthesis quality using PSNR, structural similarity index (SSIM) (**?**), and the LPIPS perceptual metric (**?**) and report the mean value of metrics through all scenes in each datasets. Also, to ease comparison, we report geometric mean of MSE $= 10^{-\text{PSNR}/10}$, $\sqrt{1 - \text{SSIM}}$, and LPIPS.

Baselines. We compare our approach to two baseline models, NeRF(Mildenhall et al., 2020) and mip-NeRF(Barron et al., 2021), and several state-of-the-art models for few-shot novel view synthesis: InfoNeRF(Kim et al., 2022), DietNeRF(Jain et al., 2021), and RegNeRF(Niemeyer et al., 2022).

Implementation details. Our implementation is built on top of the JAX mip-NeRF codebase (Barron et al., 2021). We use Adam optimizer using an exponential learning rate decay. We clip gradients by value at 0.1 and then by norm at 0.1. Our model is trained for 120k iterations for 8 hours on two NVIDIA RTX3090Ti GPUs. We provide more implementation details in supplementary materials.

5.2 COMPARISONS

Qualitative comparisons. Quantitative and qualitative comparison results in Table 1 and Figure 4 demonstrate that our model shows superior performance to baseline mip-NeRF (Barron et al., 2021) and previous SOTA model, RegNeRF (Niemeyer et al., 2022), in 3-view settings. We observe that our warping-based consistency shows high effectiveness in 360° settings, enabling our model to capture fine, narrow details that mip-NeRF and RegNeRF struggle to capture in such sparse view scenarios. Our method also displays higher stability in rendering smooth surfaces and reducing

Table 1: **Quantitative comparison on NeRF-Synthetic (Mildenhall et al., 2020) and LLFF (Mildenhall et al., 2019) dataset.**

Methods	NeRF-Synthetic			LLFF		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF (Mildenhall et al., 2020)	15.93	0.780	0.320	12.93	0.267	0.554
mipNeRF (Barron et al., 2021)	17.71	0.798	0.745	14.62	0.351	0.495
DietNeRF (Jain et al., 2021)	16.06	0.793	0.306	14.94	0.370	0.240
InfoNeRF (Kim et al., 2022)	18.65	0.811	0.230	14.37	0.349	0.457
RegNeRF (Niemeyer et al., 2022)	18.01	0.842	0.747	19.08	0.693	0.336
Ours	**19.23**	**0.866**	**0.723**	**19.15**	**0.693**	**0.288**

(a) Ground-truth (b) mip-NeRF (c) RegNeRF (d) Ours (e) Ours (D)

Figure 4: **Qualitative comparison on NeRF-Synthetic (Mildenhall et al., 2020)** show that in 3-view setting, our method better captures fine details (such as the wire in bottom *mic* scene) and models scenes with less artifacts (background in top *materials* scene) compared to previous methods.

artifacts in background in comparison to previous models, as demonstrated in the qualitative results of the *materials* scene.

Quantitative comparisons. Comparisons in Table 1 and Figure 5 shows our model's competitive results in LLFF dataset, whose PSNR results show large increase in comparison to mip-NeRF baseline and competitive outputs compared to RegNeRF. Observations show that our warping-based consistency modeling successfully removes artifacts that prominently occur between camera and scene in few-shot settings and enables model to accurate capture depth.

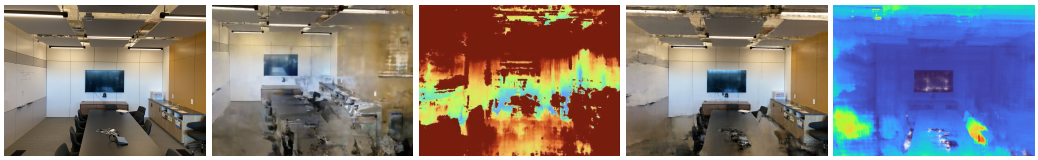
5.3 Analysis

Ablation studies. We validate the design choices in our model by performing an ablation study. Quantitative results are given in Table 2 and qualitative results are shown in Figure 6. We observe that without the consistency modeling loss $\mathcal{L}_{\text{cons}}$, our model suffers a sharp decrease in reconstruction fidelity both quantitatively and qualitatively, easily verified by more cluttered geometry in image (b) of Figure 6. We also validate the inclusion of our occlusion mask $M(p_u)$ by demonstrating the negative effects its absence has on our model: we observe that not only does it negatively effect geometry modeling, but it also causes degradation of appearance as can be denoted by image (d) of Figure 6, resulting washed-out colors throughout the overall scene. Erroneous artifacts pertaining to projections from completely opposite viewpoints were detected in different scenes, supporting the necessity for our occlusion handling method.

Table 2: **Ablation study.**

Models	PSNR↑	SSIM↑
w/o $\mathcal{L}_{\text{Cons}}$	17.83	0.790
w/o Mask	18.81	0.796
w/o $\mathcal{L}_{\text{reg}}$	19.15	0.693
Ours	**19.23**	**0.866**

Plugging in NeRF. We incorporate our proposed regularization technique to vanilla-NeRF (Mildenhall et al., 2020) baseline. Our model is compared to NeRF, DietNeRF and InfoNeRF for fair



(a) Ground-truth (b) mip-NeRF (c) mip-NeRF (D) (d) Ours (e) Ours (D)

Figure 5: **Qualitative results on LLFF (Mildenhall et al., 2019).** Comparison with baseline mip-NeRF shows that our model facilitates learning of coherent depth and geometry in extreme sparse-view settings.

(a) Ground-truth (b) w/o $\mathcal{L}_{\text{cons}}$ (c) w/o Occ. Mask (d) w/o $\mathcal{L}_{\text{reg}}$ (e) Ours

Figure 6: **Qualitative ablation.** Our qualitative ablation results on *Lego* scene shows the contribution of each module in performance of our model in 3-view scenario.

comparison. As described in Table 3, our method achieves large performance gains in terms of all metriscs in comparison to the baseline.

Our method outperforms DietNeRF and InfoNeRF in all metrics, which likewise used vanilla-NeRF baseline for sparse input view scenarios. Results are further described in supplementary materials.

Table 3: **Perceptual Loss Analysis**

Models	PSNR↑	SSIM↑
Photometric Loss	16.93	0.789
Perceptual Loss	**19.23**	**0.866**

Perceptual loss vs. Photometric loss. To justify our usage of perceptual loss for consistency modeling instead of NeRF's standard photometric loss, we conduct an experiment that compares the performance of our model regarding the two losses. Our results show that naively applying photometric loss for consistency modeling leads to divergent behaviors in geometry prediction and general degradation in overall appearance prediction, which justifies our usage of feature-level loss for consistency modeling. This phenomenon can be intuitively explained pixel-level photometric loss causing inherent difficulties within the network by forcing it to ignore view-dependent specular effects . Such consistency modeling between rendered scene and warped scene causes failure in our model to recover coherent appearance 3D geometry, while perceptual loss' feature level consistency modeling and leaves room for such view-dependent effects to be easily modeled by our network.

6 Conclusion

We have presented GeCo-NeRF, a novel approach for optimizing Neural Radiance Fields (NeRF) for few-shot novel view synthesis. We regularize geometry consistency by giving semantic consistency between rendered image and warped image. This approach overcomes limitation of NeRF with sparse inputs, which shows performance degradation with depth ambiguity and many artifacts. With feature consistency loss, we are able to regularize NeRF at unobserved viewpoints to give it beneficial geometric constraint. Further techniques and training strategies we propose prove to have stabilizing effect and facilitate optimization of our network. Our experimental evaluation demonstrates our method's competitiveness results compared to other state of the art baselines.

References

Benjamin Attal, Eliot Laidlaw, Aaron Gokaslan, Changil Kim, Christian Richardt, James Tompkin, and Matthew O'Toole. Törf: Time-of-flight radiance fields for dynamic scene view synthesis. *Advances in neural information processing systems*, 34, 2021.

Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pp. 425–432, 2001.

Gaurav Chaurasia, Sylvain Duchene, Olga Sorkine-Hornung, and George Drettakis. Depth synthesis and local warps for plausible image-based navigation. *ACM Transactions on Graphics (TOG)*, 32 (3):1–12, 2013.

Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14124–14133, 2021.

Julian Chibane, Aayush Bansal, Verica Lazova, and Gerard Pons-Moll. 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*, pp. 7911–7920, 2021.

Paul E Debevec, Camillo J Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 11–20, 1996.

Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. *arXiv preprint arXiv:2112.08867*, 2021.

Ravi Garg, Vijay Kumar Bg, Gustavo Carneiro, and Ian Reid. Unsupervised cnn for single view depth estimation: Geometry to the rescue. In *European conference on computer vision*, pp. 740–756. Springer, 2016.

Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 2017.

Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel J. Brostow. Digging into self-supervised monocular depth prediction. October 2019.

Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5875–5884, 2021.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. *Advances in neural information processing systems*, 28, 2015.

Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5885–5894, 2021.

Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 406–413, 2014.

Yoonwoo Jeong, Seokjun Ahn, Christopher Choy, Anima Anandkumar, Minsu Cho, and Jaesik Park. Self-calibrating neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5846–5854, 2021.

Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016.

Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pp. 31–42, 1996.

Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022.

Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11453–11464, 2021.

Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *CVPR*, 2020.

Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5865–5874, 2021.

Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10318–10327, 2021.

Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14335–14345, 2021.

Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. *arXiv preprint arXiv:2112.03288*, 2021.

Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33: 20154–20166, 2020.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014. URL `http://arxiv.org/abs/1409.1556`.

Vincent Sitzmann, Michael Zollhofer, and Gordon Wetzstein. Scene representation networks: Continuous 3dstructure-aware neural scene representations. In *NeurIPS*, 2019.

Pratul P Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. Learning to synthesize a 4d rgbd light field from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2243–2251, 2017.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *NeurIPS*, 2020.

Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12959–12970, 2021.

Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5038–5047, 2017.

Michael Waechter, Nils Moehrle, and Michael Goesele. Let there be color! large-scale texturing of 3d reconstructions. In *European conference on computer vision*, pp. 836–850. Springer, 2014.

Rui Wang, Stephen M Pizer, and Jan-Michael Frahm. Recurrent neural network for (un-) supervised learning of monocular video visual odometry and depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5555–5564, 2019.

Jamie Watson, Oisin Mac Aodha, Victor Prisacariu, Gabriel Brostow, and Michael Firman. The temporal opportunist: Self-supervised multi-frame monocular depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1164–1174, 2021.

Daniel N Wood, Daniel I Azuma, Ken Aldinger, Brian Curless, Tom Duchamp, David H Salesin, and Werner Stuetzle. Surface light fields for 3d photography. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pp. 287–296, 2000.

Xudong Xu, Xingang Pan, Dahua Lin, and Bo Dai. Generative occupancy fields for 3d surface-aware image synthesis. *Advances in Neural Information Processing Systems*, 34, 2021.

Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. PlenOctrees for real-time rendering of neural radiance fields. In *ICCV*, 2021a.

Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4578–4587, 2021b.

Huangying Zhan, Ravi Garg, Chamara Saroj Weerasekera, Kejie Li, Harsh Agarwal, and Ian Reid. Unsupervised learning of monocular depth estimation and visual odometry with deep feature reconstruction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 340–349, 2018.

Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020.

Tinghui Zhou, Matthew Brown, Noah Snavely, and David G Lowe. Unsupervised learning of depth and ego-motion from video. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1851–1858, 2017.