Few-Shot Neural Radiance Fields under Unconstrained Illumination

SeokYeong Lee[1,2], JunYong Choi[1,2], Seungryong Kim[2], Ig-Jae Kim[1,3,4], Junghyun Cho[1,3,4]

[1] Korea Institute of Science and Technology, Seoul
[2] Korea University, Seoul
[3] AI-Robotics, KIST School, University of Science and Technology
[4] Yonsei-KIST Convergence Research Institute, Yonsei University
{shapin94, happily, drjay, jhcho}@kist.re.kr seungryong_kim@korea.ac.kr

Abstract

In this paper, we introduce a new challenge for synthesizing novel view images in practical environments with limited input multi-view images and varying lighting conditions. Neural radiance fields (NeRF), one of the pioneering works for this task, demand an extensive set of multi-view images taken under constrained illumination, which is often unattainable in real-world settings. While some previous works have managed to synthesize novel views given images with different illumination, their performance still relies on a substantial number of input multi-view images. To address this problem, we suggest ExtremeNeRF, which utilizes multi-view albedo consistency, supported by geometric alignment. Specifically, we extract intrinsic image components that should be illumination-invariant across different views, enabling direct appearance comparison between the input and novel view under unconstrained illumination. We offer thorough experimental results for task evaluation, employing the newly created NeRF Extreme benchmark—the first in-the-wild benchmark for novel view synthesis under multiple viewing directions and varying illuminations.

Introduction

Neural radiance fields (NeRF) (Mildenhall et al. 2020) have recently made a substantial impact on 3D vision. Through optimizing a multi-layered perceptron (MLP) for mapping 3D point locations to color and volume density, NeRF significantly outperforms prior works (Lombardi et al. 2019; Sitzmann, Zollhöfer, and Wetzstein 2019; Mildenhall et al. 2019) in novel view synthesis.

However, what if *only a few images collected from the internet or mobile phones taken under unconstrained illumination conditions are available?* In most cases, NeRF-based novel view synthesis under such a practical environment is often limited since it 1) requires a massive amount of data for reliable synthesis results, and 2) assumes constrained illumination conditions among input views to encode a view-dependent color. These are key drawbacks for practical usage of NeRF, as they disable view synthesis on images that were casually collected or captured in daily life.

NeRF-W (Martin-Brualla et al. 2021) pioneered view synthesis with inputs under varying illumination, enabling novel view synthesis from internet-collected tourism images (Snavely, Seitz, and Szeliski 2006). Subsequent work (Chen et al. 2022) enables appearance hallucination of the synthesized image given unconstrained image collections, by learning a view-consistent appearance of the scene. However, these works are hindered by the limited number of input images (see Fig. 1). Moreover, previous works that deal with few-shot view synthesis (Yu et al. 2021; Jain, Tancik, and Abbeel 2021; Kim, Seo, and Han 2022; Niemeyer et al. 2022; Deng et al. 2022; Yang, Pavone, and Wang 2023) are often hindered by the illumination variation due to the characteristic of NeRF that learns radiance dependent on viewing direction and illumination.

In this paper, we address the problem of novel view synthesis of scenes *given only sparse input images taken under unconstrained illumination*, for the first time. Our proposed method, dubbed ExtremeNeRF, leverages intrinsic decomposition to mitigate the problem. The color of the scene referred to as albedo, plays an essential role in maintaining consistency regardless of changes in viewing direction or illumination conditions (see Fig. 2).

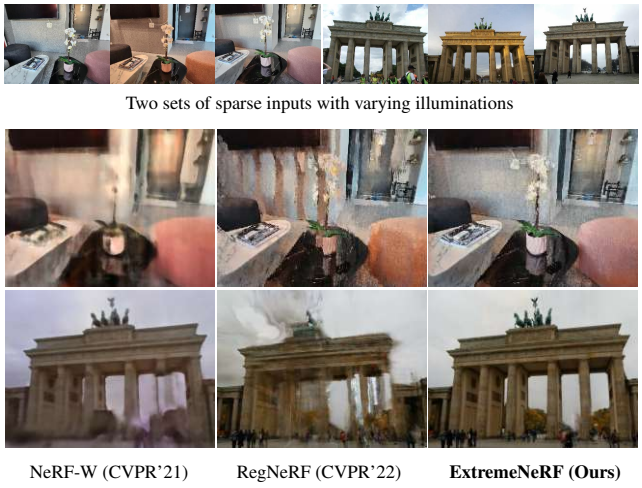

Figure 1: Few-shot view synthesis results on few inputs with varying illuminations. Our ExtremeNeRF demonstrates reliable results in comparison to baseline methods for two specific scenarios: NeRF under varying illuminations (NeRF-W) and few-shot view synthesis (RegNeRF).

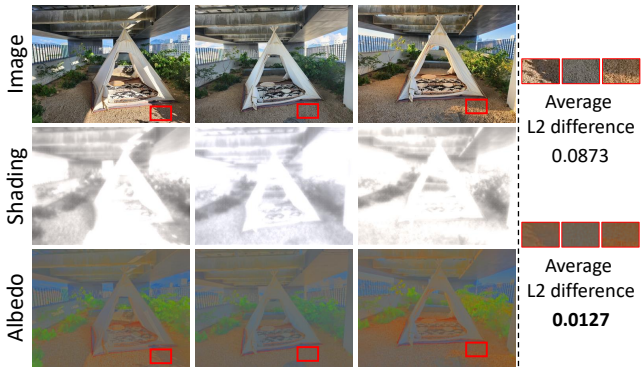

Figure 2: Intrinsic decomposition on multi-view images under varying illumination. Estimated albedo maps exhibit more illumination invariance compared to color maps, resulting in lower differences across multiple views.

Since NeRF often struggles in rendering a large-size patch due to the complexity, it is challenging to infer intrinsic components from the rendered images that are largely dependent on global contexts (Ye et al. 2022). To overcome this, we first extract the global context-aware pseudo-albedo ground truth of the inputs in the offline process. By enforcing a patch-wise module to decompose the same albedo as the pseudo-ground-truth, we then achieve global context-aware intrinsic decomposition during NeRF's optimization with minimum computational costs in an end-to-end manner. This albedo consistency loss is supported by the geometric alignment and depth consistency loss, which provides correspondences between pixels to compare and encourages correct geometry synthesis.

In evaluating our proposed method, we utilize the benchmark datasets (Snavely, Seitz, and Szeliski 2006; Chen et al. 2022) as well as our newly developed NeRF Extreme benchmark. NeRF Extreme represents the first-of-its-kind benchmark for in-the-wild novel view synthesis, capturing scenes under multiple viewing directions and varying illumination.

Related Work

Neural radiance fields. Since the introduction of NeRF (Mildenhall et al. 2020), various extensions have been proposed (Pumarola et al. 2021; Park et al. 2021; Jain et al. 2022; Poole et al. 2022; Yuan et al. 2022b; Kuang et al. 2023). However, NeRF still relies on a massive amount of images taken under consistent illumination. Some of the works investigate ways to synthesize novel views with sparse input views. Yu et al. (Yu et al. 2021) has proved that leveraging knowledge priors leads to better few-shot view synthesis. The following works (Wang et al. 2021; Jain, Tancik, and Abbeel 2021; Wang et al. 2022; Kim, Seo, and Han 2022; Deng et al. 2023) have suggested a variety of priors to improve the performance. Other works have focused on building geometry constraints to address the distortions that arise from sparse input views. Deng et al. (Deng et al. 2022) and Xu et al. (Xu et al. 2022) have presented depth prior-based methods, as (Attal et al. 2021; Roessle et al. 2022; Johari, Lepoittevin, and Fleuret 2022; Yuan et al. 2022a). Some of the other methods (Chen et al. 2021; Johari, Lepoittevin, and Fleuret 2022; Watson et al. 2022; Wynn and Turmukhambetov 2023) utilize implicit geometry priors for the task. Recently, RegNeRF (Niemeyer et al. 2022) suggest depth smoothness constraints enhance the rendered novel view geometry, while FreeNeRF (Yang, Pavone, and Wang 2023) add frequency regularization on it. View synthesis with inputs taken under varying illuminations is covered by some of the previous works (Martin-Brualla et al. 2021; Chen et al. 2022), however, they rely on a massive amount of input images (see Tab. 1) rather than sparse input views.

	RegNeRF	NeRF-W	NeROIC	**Ours**
Varying-illum.	✗	✓	✓	✓
Few-shot	✓	✗	✓	✓
Frontal-facing	✓	✓	✗	✓

Table 1: Our method enables few-shot view synthesis given non-object-centric images taken under varying illumination.

Illumination decomposition. Various frameworks have been developed to tackle the problem of decomposing multiple scene properties including illumination, some of which rely on large datasets of paired images and ground truth information (Li et al. 2020, 2021; Choi et al. 2023). Other approaches, such as those proposed in works like (Li and Snavely 2018; Liu et al. 2020; Das, Karaoglu, and Gevers 2022), have explored methods to address the problem of decomposing illumination-invariant color from the scene. With the help of NeRF, some of the recent works (Boss et al. 2021a,b, 2022; Toschi et al. 2023) include NeROIC (Kuang et al. 2022) deal with a neural decomposition of an image. However, they require massive multi-view sampling of an object, rather than a scene. Decomposing illumination from a scene involves the complex interaction of indirect illumination and scene geometries, aspects that are not extensively addressed in object-level neural decomposition. Other recent works (Ye et al. 2022; Rudnev et al. 2022; Kuang et al. 2023; Yang et al. 2023) deal with NeRF-based inverse rendering of a scene, however, focusing on disentanglement of a scene component rather than view synthesis.

Preliminaries

Neural radiance field. NeRF (Mildenhall et al. 2020) is a view-synthesis framework that maps 5D inputs (3D coordinate and viewing direction of a ray) to color and volume density, denoted by c and σ, respectively. Specifically, with a ray $r_x(t) = o + td_x$, where o, d_x, and t indicate camera origin, ray direction, and scene bound at pixel location $x \in \mathbb{R}^2$, respectively, a view-dependent color $\hat{c}(x) \in [0,1]^3$ can be rendered such that

$$\hat{c}(x) = \int_{t_n}^{t_f} T(t)\sigma(t)c(t)dt, \tag{1}$$

while $T(t) = \exp(-\int_{t_n}^{t} \sigma(s)ds)$ and $\sigma(\cdot), c(\cdot)$ are density and color predictions from the network, respectively. Simi-

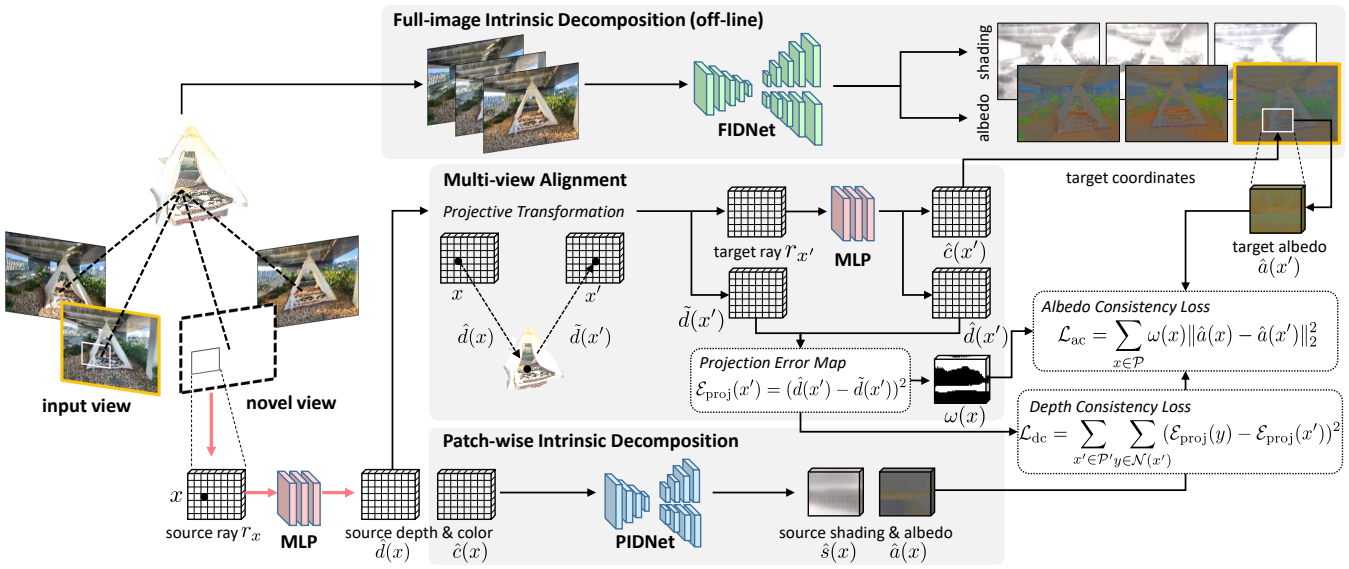

Figure 3: Overall architecture of our ExtremeNeRF. PIDNet extracts intrinsic components from the synthesized patch $\hat{c}(x)$ while enforcing extracted albedo to be identical with the pseudo-albedo ground truth. A weight term $\omega(x)$ and depth consistency loss $\mathcal{L}_{\text{dc}}$ encourage proper correspondence matching between two views. A bold, crimson arrow indicates the inference phase.

larly, a depth value $\hat{d}(x)$ at x can also be rendered as

$$\hat{d}(x) = \int_{t_n}^{t_f} T(t)\sigma(t)t dt. \tag{2}$$

Optimization in NeRF relies on a mean squared error on synthesized color $\hat{c}(x)$ as

$$\mathcal{L}_{\text{color}} = \sum_{x \in \mathcal{S}} \|\hat{c}(x) - c_{\text{gt}}(x)\|_2^2, \tag{3}$$

where $\mathcal{S}$ indicates the set of sampled pixels, and $c_{\text{gt}}(x)$ indicates ground-truth color at x. Since volume density σ is also optimized based on the color consistency across different views, violation of the consistent illumination assumption results in inaccurate geometry.

Intrinsic decomposition. Intrinsic decomposition aims to decompose an image into illumination-invariant color, referred to as albedo, and shading, based on the Lambertian assumption that every observed surface is diffuse. Specifically, a pixel color $c(x)$ is formulated as a multiplication of the albedo ($a(x)$) and the shading ($s(x)$) as follows:

$$\log c(x) = \log a(x) + \log s(x). \tag{4}$$

However, most real-world objects have surfaces whose reflectances vary upon viewing directions and are often lit by colored lights. Thus, Eq. 4 can be rewritten as follows:

$$\log c(x) = \log a(x) + \log s(x) + l + R, \tag{5}$$

which takes light color vector l and non-Lambertian residuals R into account (Li and Snavely 2018).

Real-world image intrinsic decomposition remains a challenging, imperfectly solved task. While recent works (Li and Snavely 2018; Liu et al. 2020; Das, Karaoglu, and Gevers 2022) demonstrate reliable performance, they still face limitations with unseen and challenging cases. Additionally, relying on global context for large-resolution image rendering in NeRF incurs computational and memory expenses.

Method

Overview

The objective of this work is to build an illumination-robust few-shot view synthesis framework by regularizing albedo that should be identical across multi-view images regardless of illumination. Our major challenges are to 1) achieve reliable geometry alignment between different views and 2) decompose the albedo of a rendered view without extensive computational costs.

Instead of directly addressing NeRF-based intrinsic decomposition, we integrate a pre-existing intrinsic decomposition network with NeRF optimization. Our approach involves a few-shot view synthesis framework that employs an offline intrinsic decomposition network, offering global context-aware pseudo-albedo ground truth without the computational overhead. As illustrated in Fig. 3, FIDNet provides pseudo-albedo ground truths for the input images before the start of the training, guiding PIDNet to extract intrinsic components for novel synthesized views based on these pseudo truths and multi-view correspondences. This allows our NeRF to learn illumination-robust few-shot view synthesis through cross-view albedo consistency. Subsequent subsections detail each framework component.

Albedo Estimation

Building upon our hypothesis that albedo aids in view synthesis with varying illumination inputs, it is crucial to decompose intrinsics from both the input and the novel views.

	DTU	PT	NeRD	ReNe	**Ours**
Indoor	✓	✗	✓	✓	✓
Outdoor	✗	✓	✓	✗	✓
Real-world	✓	✓	✓	✓	✓
In-the-wild	✗	✓	✗	✗	✓
Frontal-facing	✗	✓	✗	✗	✓

Table 2: Multi-illumination dataset comparison. Our NeRF Extreme dataset provides in-the-wild, non-object-centric, and varying illumination images taken indoors and outdoors.

Instead of relying on optimization-based methods (Boss et al. 2021a,b, 2022; Ye et al. 2022), which may yield sub-optimal outcomes with limited data (0.06 times less), we propose a concise two-stage intrinsic decomposition pipeline: a full-image and patch-wise intrinsic decomposition network, called FIDNet and PIDNet, respectively. FIDNet, formulated with a pre-trained intrinsic decomposition model, extracts the albedo of the input images - pseudo-albedo ground truths - offline, to guide PIDNet with global contexts. Given the guidance, PIDNet extracts albedo ($\hat{a}(x)$) of the synthesized color patch with the size of S_{patch} at the novel view ($\hat{c}(x)$), minimizing the difference with the pseudo-ground truth, $\mathcal{L}_{\text{ac}}$ (Eq. 8), supported by multi-view alignment process described below.

Geometry Alignment and Regularization

For any 3D point $x^w \in \mathbb{R}^3$ in a world coordinate, a camera projection from x^w to pixel location x can be defined by the inverse of camera-to-world transformation $T \in SE(3)$ and camera intrinsics $K \in \mathbb{R}^{3\times3}$. Likewise, a mapping from pixel location to 3D point can be defined by the inverse operation and $d(x)$, the depth at the pixel location, as:

$$x = KT^{-1}x^w, \quad x^w = Td(x)K^{-1}\bar{x}. \tag{6}$$

Note that $\bar{x} = [x^T, 1]$, a homogeneous representation of x.

Given a pixel x in the novel view, we need the pixel x' in the input view depicting the same 3D point x^w as x for cross-view consistency. If the depth of a given image pixel $d(x)$ is known, x' can be obtained by Eq. 6 as follows:

$$x' = (K'T'^{-1}T)d(x)K^{-1}\bar{x}, \tag{7}$$

where K', T' and K, T indicate camera intrinsics and camera-to-world matrices of the input and novel view, respectively.

Albedo consistency. Based on the pixel correspondence obtained above, we can impose image consistency between inputs and novel views. However, under varying illumination, Eq. 3 cannot regularize view-dependent color as it does under constrained illumination, for its different interactions within illumination (see Fig. 2). To overcome this, we present L-2 normalized albedo consistency loss $\mathcal{L}_{\text{ac}}$ formulated as follows:

$$\mathcal{L}_{\text{ac}} = \sum_{x\in\mathcal{P}} \omega(x)\|\hat{a}(x) - \hat{a}(x')\|_2^2, \tag{8}$$

where $\hat{a}(x)$, $\hat{a}(x')$ indicate the extracted albedo from the novel and the input view, respectively, while $\mathcal{P}$ denotes all the pixels in the novel view. A weight term $\omega(x)$ is described below.

Visibility mask. The projective transformation often utilizes incorrect synthesized depth values. For all cases, a projection error on x', denote by $\mathcal{E}_{\text{proj}}$ can be defined as follow:

$$\mathcal{E}_{\text{proj}}(x') = (\hat{d}(x') - \tilde{d}(x'))^2, \tag{9}$$

where $\hat{d}(x')$ and $\tilde{d}(x')$ indicate rendered depth and projected depth, a byproduct of Eq. 7, repsectively. A projection error $\mathcal{E}_{\text{proj}}$ should be close to zero if there exists neither self-occlusion nor ill-synthesized floating artifacts.

We define visibility mask to exclude invalid cases for multi-view consistency. First, we set the mask as 0 if the projected pixel x' is outside of the field of view. Secondly, we exclude the unrelated pixel pairs caused by the scene geometry (occlusions). To distinguish the projection errors caused by the scene geometry from the ones caused by the floating artifacts, we define a weight term ω as

$$\omega(x) = r_{\text{e}}(1 - (\mathcal{E}_{\text{proj}}(x)/\mathcal{M}_{\text{proj}})), \tag{10}$$

where r_{e} and $\mathcal{M}_{\text{proj}}$ indicate the error rate coefficient and the maximum projection error, respectively. The role of w is to control the weight of the cross-view consistency concerning the amount of projection error. As inaccurate geometry alignments are rectified during optimization while occlusions persist, r_{e} diminishes toward the end criteria, leading to a reduction in the number of pairs that are enforced to maintain cross-view consistency.

Depth consistency. A direct minimization of $\mathcal{E}_{\text{proj}}$ can be counterproductive due to occlusion, by smoothing two unrelated surface depths. Instead, we present a depth consistency loss $\mathcal{L}_{\text{dc}}$ that regularizes the amount of projection error between adjacent pixels. Depth consistency loss $\mathcal{L}_{\text{dc}}$ in the input view can be defined such that

$$\mathcal{L}_{\text{dc}} = \sum_{x'\in\mathcal{P}'} \sum_{y\in\mathcal{N}(x')} (\mathcal{E}_{\text{proj}}(y) - \mathcal{E}_{\text{proj}}(x'))^2, \tag{11}$$

where y indicates one of the 4-neighbor adjacent pixels $\mathcal{N}(x')$ for x'. $\mathcal{P}'$ denotes all the pixels in the input view.

Total Loss

In addition to the loss functions $\mathcal{L}_{\text{ac}}$ and $\mathcal{L}_{\text{dc}}$, we incorporate several constraints for the optimization of NeRF and PIDNet: the depth smoothness loss $\mathcal{L}_{\text{ds}}$ (Niemeyer et al. 2022), the edge-preserving loss $\mathcal{L}_{\text{edge}}$ (Godard, Mac Aodha, and Brostow 2017), the intrinsic smoothness loss $\mathcal{L}_{\text{pid}}$ (Li and Snavely 2018), the chromaticity consistency loss $\mathcal{L}_{\text{chrom}}$ (Ye et al. 2022), and the frequency regularization mask, proposed by FreeNeRF (Yang, Pavone, and Wang 2023). Further details are provided in the supplementary material.

Experiments

In this section, we provide extensive comparisons with the baselines using our newly proposed datasets. Further results and details can be found in the supplementary material.

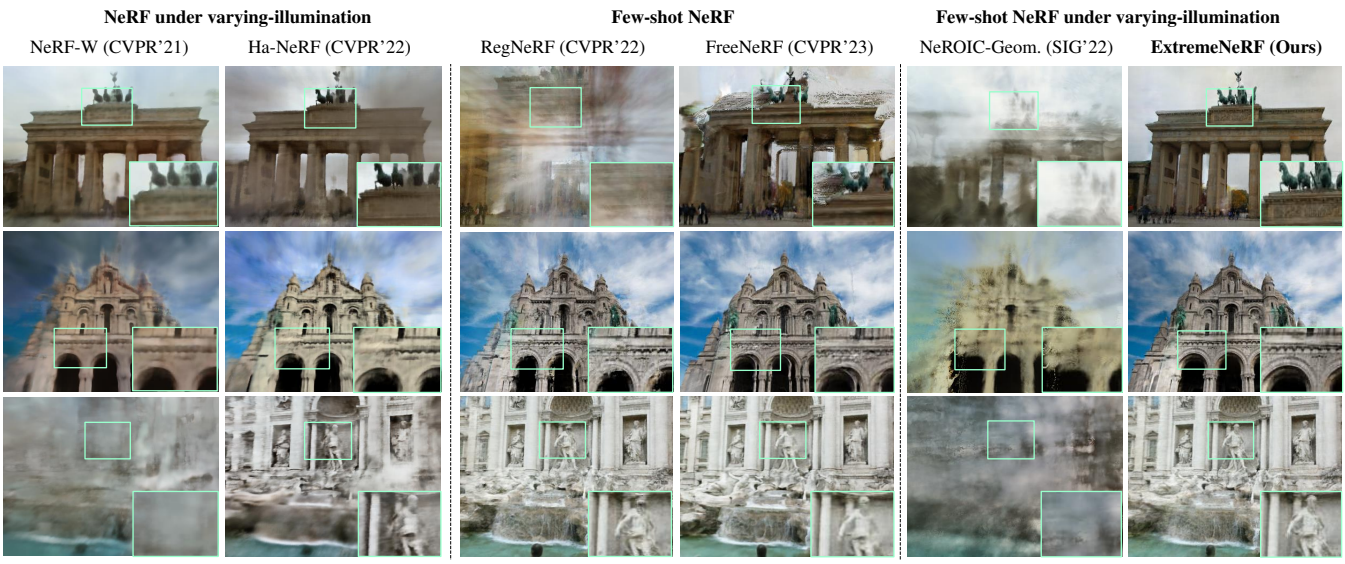

Figure 4: Qualitative comparison on Phototourism $\mathbf{F}^3$ benchmark. Synthesized novel views of 'Brandenburg Gate', 'Sacre Coeur', and 'Trevi Fountain' (from top to bottom), generated by the baselines and our proposed method in 3 view input images.

	Brandenburg Gate			Sacre Coeur			Trevi Fountain		
	SSIM ↑	LPIPS ↓	Abs Rel ↓	SSIM ↑	LPIPS ↓	Abs Rel ↓	SSIM ↑	LPIPS ↓	Abs Rel ↓
NeRF-W (CVPR'21)	0.39	0.59	0.84	0.46	0.52	0.94	0.16	0.64	0.65
Ha-NeRF (CVPR'22)	0.50	0.43	**0.78**	0.46	0.52	0.94	0.39	0.48	**0.52**
RegNeRF (CVPR'22)	0.27	0.56	3.54	0.39	0.44	2.44	0.43	0.37	0.64
FreeNeRF (CVPR'23)	0.31	0.50	4.53	0.32	0.45	3.62	0.45	**0.36**	0.57
NeROIC-Geom. (SIG'22)	0.30	0.63	0.89	0.34	0.66	**0.85**	0.11	0.70	0.79
ExtremeNeRF (Ours)	**0.56**	**0.36**	**0.78**	**0.49**	**0.38**	1.28	**0.57**	**0.36**	0.59

Table 3: Quantitative comparison on Phototourism $\mathbf{F}^3$ in 3 view setting proves that our model succeeded in synthesizing fine geometry details.Bold texts for the best performance, and underline for the 2nd best.

Datasets

In Table 2, Phototourism (PT) and its variants (Snavely, Seitz, and Szeliski 2006; Chen et al. 2022) are the only benchmarks that exhibit both pose and illumination variations. However, these datasets are not suitable for few-shot view synthesis due to their randomness. For extensive experiments, we construct two datasets for the evaluation of few-shot view synthesis under varying illumination.

Phototourism $\mathbf{F}^3$. Phototourism $\mathbf{F}^3$(Frontal Facing Few-shot), a subset of Phototourism (Snavely, Seitz, and Szeliski 2006) dataset, is specifically curated for evaluating few-shot view synthesis under varying illumination. Frontal-facing scenes within similar depth bounds and significant illumination variation in 'Brandenburg Gate', 'Sacre Coeur', and 'Trevi Fountain' are selected for the task. The ground truth depth maps are provided by Phototourism. The rationale behind building a frontal-facing subset can be found in the supplementary material.

NeRF Extreme. To build a benchmark that fully reflects unconstrained environments, we collected multi-view images with varying light sources such as multiple light bulbs and the sun using the the mobile phone camera. We took 40 images per scene - 30 images in the train set and 10 images in the test set. The training sets are captured with at least three different lighting conditions. The camera poses and depth maps are obtained using the COLMAP (Schönberger et al. 2016) and multi-view stereo method (Giang, Song, and Jo 2022), respectively. NeRF Extreme is the first in-the-wild multi-view dataset with varying illumination, whose scenes are not limited to object-centric or outdoor scenes.

Experimental Settings

Baselines. Since there is no previous work that deals with scene-level few-shot view synthesis under varying illumination, we compare our proposed method against three types of baselines. 1) NeRF under varying illumination: NeRF-W (Martin-Brualla et al. 2021), Ha-NeRF (Chen et al. 2022), 2) Few-shot NeRF: RegNeRF (Niemeyer et al. 2022), FreeNeRF (Yang, Pavone, and Wang 2023), and 3) Few-shot NeRF under varying-illumination: NeROIC (Kuang et al. 2022). For NeROIC, we report NeROIC-Geom results as

Figure 5: Qualitative comparison on NeRF Extreme benchmark. A synthesized novel view of 'Cafe', 'Kitchen', and 'Flower' (from top to bottom), generated by the baselines and our proposed method.

	Cafe			Kitchen			Flower		
	SSIM ↑	LPIPS ↓	Abs Rel ↓	SSIM ↑	LPIPS ↓	Abs Rel ↓	SSIM ↑	LPIPS ↓	Abs Rel ↓
NeRF-W (CVPR'21)	0.32	0.55	0.64	0.60	0.47	0.35	0.62	0.42	0.55
Ha-NeRF (CVPR'22)	0.36	0.54	0.62	0.54	0.52	0.37	0.65	0.43	0.70
RegNeRF (CVPR'22)	0.36	0.48	0.66	0.55	0.39	**0.34**	0.58	0.49	0.78
FreeNeRF (CVPR'23)	0.39	0.43	0.90	0.55	0.40	0.81	0.60	0.42	0.86
NeROIC-Geom. (SIG'22)	0.25	0.67	0.80	0.47	0.58	0.49	0.49	0.55	0.54
ExtremeNeRF (Ours)	**0.48**	**0.38**	**0.51**	**0.62**	**0.34**	0.35	**0.67**	**0.40**	**0.49**

Table 4: A quantitative comparison of NeRF Extreme in 3 views demonstrates the superior performance of our ExtremeNeRF.

NeROIC-Full exhibits some divergence. Note that NeROIC is tailored for object-centric scenes, not frontal-facing ones.

For comparison, we used the mean SSIM, LPIPS metric of the synthesized image, and Abs Rel (Absolute Relative Error) of the synthesized depth map. Similar works (Martin-Brualla et al. 2021; Chen et al. 2022; Kuang et al. 2022) have evaluated performance using PSNR after relighting to match the target illumination. However, our main aim is to highlight improved geometry details rather than relighting. Moreover, the baselines struggle with proper relighting in a few-shot setting, making PSNR unsuitable for evaluation.

Implementation details. Our framework is based on the implementation of RegNeRF (Niemeyer et al. 2022). For FIDNet, the official code and model of IIDWW (Li and Snavely 2018) trained with BigTimes dataset are used without fine-tuning. An image size of 300×400 is used for the training, with $S_{\text{patch}} = 32 \times 32$. We train every scene for $70K$ using 4 NVIDIA A100 GPUs.

Computational complexity. Except for a few-shot NeRF, most methods require about 10 to 20 hours of training time to achieve optimal performance on 4 NVIDIA A100 GPUs. For few-shot NeRFs, RegNeRF (Niemeyer et al. 2022), our proposed method, and FreeNeRF (Yang, Pavone, and Wang 2023) take 2,4 and 1.5 hours in 3 views, respectively.

Experimental Results

Comparisons with the baselines. Fig. 4 and Fig. 5 show the qualitative comparison between our ExtremeNeRF and other baseline methods on Phototourism F^3 and NeRF Extreme, respectively. In the 1st and 2nd rows, baselines dealing with varying illumination lack input images, resulting in smoothed geometry details. For few-shot NeRF methods (the 3rd and 4th rows), synthesized geometries face challenges due to inconsistent illumination. Particularly, baselines exhibit higher distortion when confronted with significant illumination variations, as observed in the 'Brandenburg Gate' and 'Cafe' scenes, respectively. The results are further supported by quantitative comparisons in Tab.3 and Tab.4, especially with a large improvement in SSIM and LPIPS (bold texts for the best performance, and underline for the 2nd best). In the case of NeROIC (Kuang et al. 2022), synthesized results from NeROIC-Geom., which is a partially optimized version of the method, are reported.

	Cafe			Kitchen			Flower		
	SSIM ↑	LPIPS ↓	Abs Rel ↓	SSIM ↑	LPIPS ↓	Abs Rel ↓	SSIM ↑	LPIPS ↓	Abs Rel ↓
1-1. w/o AC	0.414	0.42	0.92	0.60	0.35	0.79	0.63	0.41	0.86
1-2. w/ AC - w/o FIDNet	0.413	0.42	0.92	0.60	0.35	0.80	0.63	0.42	0.88
1-3. w/ AC - Albedo MLP	0.413	0.42	0.92	0.59	0.35	0.85	0.64	0.41	0.89
1-4. w/ AC (PIENet)	0.445	0.39	0.89	0.60	0.35	0.79	0.62	0.42	0.88
2-1. w/o DC	0.470	0.39	0.89	**0.62**	0.35	0.81	0.62	0.44	0.88
2-2. w/o Visibility Mask	0.404	0.43	0.92	0.60	0.36	0.80	0.62	0.43	0.89
ExtremeNeRF (Ours)	**0.476**	**0.38**	**0.51**	**0.62**	**0.34**	**0.36**	**0.67**	**0.40**	**0.49**

Table 5: Ablation study of our ExtremeNeRF. Ablation studies on two different groups are provided, in terms of 1) albedo consistency (AC) and 2) depth consistency (DC).

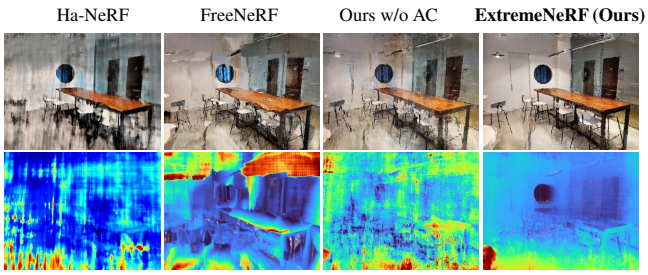

Figure 6: Depth map comparison. Depth maps paired with synthesized images of 'Cafe' scene of NeRF Extreme benchmark are selected (Best viewed in color).

	SSIM ↑	LPIPS ↓	Abs Rel ↓
NeRF-W (CVPR'21)	0.38	0.51	0.79
RegNeRF (CVPR'22)	0.44	0.35	**0.76**
ExtremeNeRF (Ours)	**0.45**	**0.34**	**0.76**

Table 6: Quantitative comparison on the 'fern' scene of the LLFF (Mildenhall et al. 2020) in 3 view settings.

Note that the entire model shows diverged results on frontal-facing scenes. In all cases, our method demonstrates plausible synthesized results with fine geometry details. Additionally, Fig. 6 illustrates that our model exhibits reliable depth synthesis, leading to the expectation of achieving plausible video synthesis performance, even when the Abs Rel score is compatible with each other ('Brandenburg Gate' scene).

Ablation studies. Tab. 5 shows groups of ablation studies to validate the design choices of our work. Each studies related to albedo consistency (1-1 to 1-4) and depth consistency (2-1 to 2-2). The additional ablation studies on the patch size and learned priors can be found in the supplementary material.

The experiments in the first group demonstrate that incorporating albedo consistency between the input and novel views contributes to the regularization of geometry. Quantitative results in 1-2, reveal that removing albedo consistency leads to sub-optimal performance, indicating the importance of this constraint for well-constrained optimization of a novel view. A qualitative comparison of the synthesized maps in Fig. 6 supports the idea that incorporating albedo consistency contributes to reliable depth estimation. In 1-3 and 1-4, we provide empirical evidence that FIDNet serves as a suitable guide for achieving cross-view consistency, rather than MLP that synthesizes albedo. In 1-4, we replace the FIDNet model from IIDWW (Li and Snavely 2018) with the other intrinsic decomposition model (Das, Karaoglu, and Gevers 2022), however, shows degraded performance. Note that FIDNet can be substituted with other models in our framework if they exhibit superior performance.

The experiments in the second group illustrate that depth consistency, when taken into account with proper consideration of scene geometry, contributes to improved geometry. Ablating depth consistency (2-1) and visibility mask (2-2) results in unreliable depths, as enforcing consistency between unrelated surfaces leads to undesirable outcomes.

Comparisons on LLFF. To assess performance on the benchmark with constrained illumination, Table 6 compares results on the 'fern' scene from the LLFF (Mildenhall et al. 2019) dataset. RegNeRF (Niemeyer et al. 2022) shows minor differences compared to our method when illumination is shared among inputs, while NeRF-W (Martin-Brualla et al. 2021) significantly degrades with few-shot inputs.

Conclusion and Further Work

In this paper, we proposed ExtremeNeRF, which can synthesize a novel view in practical environments, where neither a large amount of multi-view images nor consistent illumination is available. By regularizing albedo which should be identical across different views, our method can directly regularize appearance instead of interpolating view-dependent color as vanilla-NeRF did. We have proved that the proposed method outperforms other previous works with new benchmarks in a few-shot view synthesis under an unconstrained illumination environment. Any few-shot NeRF can obtain illumination-robust regularization by utilizing our proposed albedo consistency constraints on their optimization. However, similar to other optimization-based NeRF approaches, relighting a scene given sparse inputs remains a challenge. Further, significant illumination variation may result in noisy input camera poses. Addressing these problems could be a potential direction for our future work.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2023-00227592, Development of 3D object identification technology robust to viewpoint changes, 70%, and No.2020-0-00457, Development of free-form plenoptic video authoring and visualization platform for large space, 30%)

References

Attal, B.; Laidlaw, E.; Gokaslan, A.; Kim, C.; Richardt, C.; Tompkin, J.; and O'Toole, M. 2021. Törf: Time-of-flight radiance fields for dynamic scene view synthesis. *NeurIPS*, 34: 26289–26301.

Boss, M.; Braun, R.; Jampani, V.; Barron, J. T.; Liu, C.; and Lensch, H. 2021a. Nerd: Neural reflectance decomposition from image collections. In *ICCV*, 12684–12694.

Boss, M.; Engelhardt, A.; Kar, A.; Li, Y.; Sun, D.; Barron, J. T.; Lensch, H. P.; and Jampani, V. 2022. SAMURAI: Shape And Material from Unconstrained Real-world Arbitrary Image collections. In *NeurIPS*.

Boss, M.; Jampani, V.; Braun, R.; Liu, C.; Barron, J.; and Lensch, H. 2021b. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. *NeurIPS*, 34: 10691–10704.

Chen, A.; Xu, Z.; Zhao, F.; Zhang, X.; Xiang, F.; Yu, J.; and Su, H. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *ICCV*, 14124–14133.

Chen, X.; Zhang, Q.; Li, X.; Chen, Y.; Feng, Y.; Wang, X.; and Wang, J. 2022. Hallucinated Neural Radiance Fields in the Wild. In *CVPR*, 12943–12952.

Choi, J.; Lee, S.; Park, H.; Jung, S.-W.; Kim, I.-J.; and Cho, J. 2023. MAIR: Multi-view Attention Inverse Rendering with 3D Spatially-Varying Lighting Estimation. *arXiv preprint arXiv:2303.12368*.

Das, P.; Karaoglu, S.; and Gevers, T. 2022. PIE-Net: Photometric Invariant Edge Guided Network for Intrinsic Image Decomposition. In *CVPR*, 19790–19799.

Deng, C.; Jiang, C.; Qi, C. R.; Yan, X.; Zhou, Y.; Guibas, L.; Anguelov, D.; et al. 2023. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *CVPR*, 20637–20647.

Deng, K.; Liu, A.; Zhu, J.-Y.; and Ramanan, D. 2022. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, 12882–12891.

Giang, K. T.; Song, S.; and Jo, S. 2022. CURVATURE-GUIDED DYNAMIC SCALE NETWORKS FOR MULTI-VIEW STEREO. In *International Conference on Learning Representations*.

Godard, C.; Mac Aodha, O.; and Brostow, G. J. 2017. Unsupervised monocular depth estimation with left-right consistency. In *CVPR*, 270–279.

Jain, A.; Mildenhall, B.; Barron, J. T.; Abbeel, P.; and Poole, B. 2022. Zero-shot text-guided object generation with dream fields. In *CVPR*, 867–876.

Jain, A.; Tancik, M.; and Abbeel, P. 2021. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *ICCV*, 5885–5894.

Johari, M. M.; Lepoittevin, Y.; and Fleuret, F. 2022. GeoNeRF: Generalizing NeRF With Geometry Priors. In *CVPR*, 18365–18375.

Kim, M.; Seo, S.; and Han, B. 2022. InfoNeRF: Ray Entropy Minimization for Few-Shot Neural Volume Rendering. In *CVPR*, 12912–12921.

Kuang, Z.; Luan, F.; Bi, S.; Shu, Z.; Wetzstein, G.; and Sunkavalli, K. 2023. Palettenerf: Palette-based appearance editing of neural radiance fields. In *CVPR*, 20691–20700.

Kuang, Z.; Olszewski, K.; Chai, M.; Huang, Z.; Achlioptas, P.; and Tulyakov, S. 2022. NeROIC: neural rendering of objects from online image collections. *ACM Transactions on Graphics*, 41(4): 1–12.

Li, Z.; Shafiei, M.; Ramamoorthi, R.; Sunkavalli, K.; and Chandraker, M. 2020. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *CVPR*, 2475–2484.

Li, Z.; and Snavely, N. 2018. Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 9039–9048.

Li, Z.; Yu, T.-W.; Sang, S.; Wang, S.; Song, M.; Liu, Y.; Yeh, Y.-Y.; Zhu, R.; Gundavarapu, N.; Shi, J.; et al. 2021. Openrooms: An open framework for photorealistic indoor scene datasets. In *CVPR*, 7190–7199.

Liu, Y.; Li, Y.; You, S.; and Lu, F. 2020. Unsupervised learning for intrinsic image decomposition from a single image. In *CVPR*, 3248–3257.

Lombardi, S.; Simon, T.; Saragih, J.; Schwartz, G.; Lehrmann, A.; and Sheikh, Y. 2019. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*.

Martin-Brualla, R.; Radwan, N.; Sajjadi, M. S.; Barron, J. T.; Dosovitskiy, A.; and Duckworth, D. 2021. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 7210–7219.

Mildenhall, B.; Srinivasan, P. P.; Ortiz-Cayon, R.; Kalantari, N. K.; Ramamoorthi, R.; Ng, R.; and Kar, A. 2019. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Trans. on Graphics*, 38(4): 1–14.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 405–421. Springer.

Niemeyer, M.; Barron, J. T.; Mildenhall, B.; Sajjadi, M. S.; Geiger, A.; and Radwan, N. 2022. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *CVPR*, 5480–5490.

Park, K.; Sinha, U.; Barron, J. T.; Bouaziz, S.; Goldman, D. B.; Seitz, S. M.; and Martin-Brualla, R. 2021. Nerfies: Deformable neural radiance fields. In *ICCV*, 5865–5874.

Poole, B.; Jain, A.; Barron, J. T.; and Mildenhall, B. 2022. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*.

Pumarola, A.; Corona, E.; Pons-Moll, G.; and Moreno-Noguer, F. 2021. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 10318–10327.

Roessle, B.; Barron, J. T.; Mildenhall, B.; Srinivasan, P. P.; and Nießner, M. 2022. Dense depth priors for neural radiance fields from sparse input views. In *CVPR*, 12892–12901.

Rudnev, V.; Elgharib, M.; Smith, W.; Liu, L.; Golyanik, V.; and Theobalt, C. 2022. Nerf for outdoor scene relighting. In *ECCV*, 615–631. Springer.

Schönberger, J. L.; Zheng, E.; Pollefeys, M.; and Frahm, J.-M. 2016. Pixelwise View Selection for Unstructured Multi-View Stereo. In *ECCV*.

Sitzmann, V.; Zollhöfer, M.; and Wetzstein, G. 2019. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *NeurIPS*, 32.

Snavely, N.; Seitz, S. M.; and Szeliski, R. 2006. Photo tourism: exploring photo collections in 3D. In *ACM SIGGRAPH*, 835–846.

Toschi, M.; De Matteo, R.; Spezialetti, R.; De Gregorio, D.; Di Stefano, L.; and Salti, S. 2023. ReLight My NeRF: A Dataset for Novel View Synthesis and Relighting of Real World Objects. In *CVPR*, 20762–20772.

Wang, C.; Chai, M.; He, M.; Chen, D.; and Liao, J. 2022. CLIP-NeRF: Text-and-Image Driven Manipulation of Neural Radiance Fields. In *CVPR*, 3835–3844.

Wang, Q.; Wang, Z.; Genova, K.; Srinivasan, P. P.; Zhou, H.; Barron, J. T.; Martin-Brualla, R.; Snavely, N.; and Funkhouser, T. 2021. Ibrnet: Learning multi-view image-based rendering. In *CVPR*, 4690–4699.

Watson, D.; Chan, W.; Martin-Brualla, R.; Ho, J.; Tagliasacchi, A.; and Norouzi, M. 2022. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*.

Wynn, J.; and Turmukhambetov, D. 2023. Diffusionerf: Regularizing neural radiance fields with denoising diffusion models. In *CVPR*, 4180–4189.

Xu, D.; Jiang, Y.; Wang, P.; Fan, Z.; Shi, H.; and Wang, Z. 2022. SinNeRF: Training Neural Radiance Fields on Complex Scenes from a Single Image. *arXiv preprint arXiv:2204.00928*.

Yang, J.; Pavone, M.; and Wang, Y. 2023. FreeNeRF: Improving Few-shot Neural Rendering with Free Frequency Regularization. In *CVPR*, 8254–8263.

Yang, S.; Cui, X.; Zhu, Y.; Tang, J.; Li, S.; Yu, Z.; and Shi, B. 2023. Complementary Intrinsics From Neural Radiance Fields and CNNs for Outdoor Scene Relighting. In *CVPR*, 16600–16609.

Ye, W.; Chen, S.; Bao, C.; Bao, H.; Pollefeys, M.; Cui, Z.; and Zhang, G. 2022. IntrinsicNeRF: Learning Intrinsic Neural Radiance Fields for Editable Novel View Synthesis. *arXiv preprint arXiv:2210.00647*.

Yu, A.; Ye, V.; Tancik, M.; and Kanazawa, A. 2021. pixelnerf: Neural radiance fields from one or few images. In *CVPR*, 4578–4587.

Yuan, Y.-J.; Lai, Y.-K.; Huang, Y.-H.; Kobbelt, L.; and Gao, L. 2022a. Neural Radiance Fields from Sparse RGB-D Images for High-Quality View Synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–16.

Yuan, Y.-J.; Sun, Y.-T.; Lai, Y.-K.; Ma, Y.; Jia, R.; and Gao, L. 2022b. Nerf-editing: geometry editing of neural radiance fields. In *CVPR*, 18353–18364.