ExtremeNeRF: Few-shot Neural Radiance Fields Under Unconstrained Illumination

SeokyYeong 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

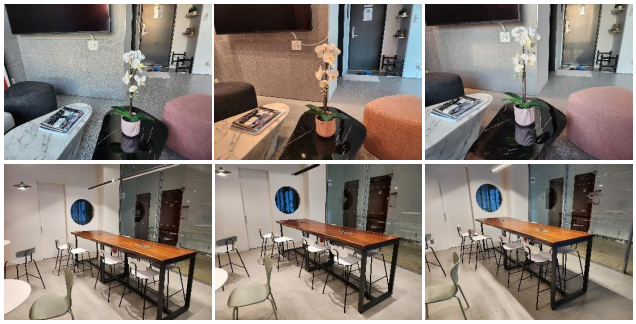
Abstract

In this paper, we propose a new challenge that synthesizes a novel view in a more practical environment, where the number of input multi-view images is limited and illumination variations are significant. Despite recent success, neural radiance fields (NeRF) require a massive amount of input multi-view images taken under constrained illuminations. To address the problem, we suggest ExtremeNeRF, which utilizes occlusion-aware multi-view albedo consistency, supported by geometric alignment and depth consistency. 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 provide extensive experimental results for an evaluation of the task, using the newly built NeRF Extreme benchmark, which is the first in-the-wild novel view synthesis benchmark taken under multiple viewing directions and varying illuminations. The project page is at https://seokyeong94.github.io/ExtremeNeRF/

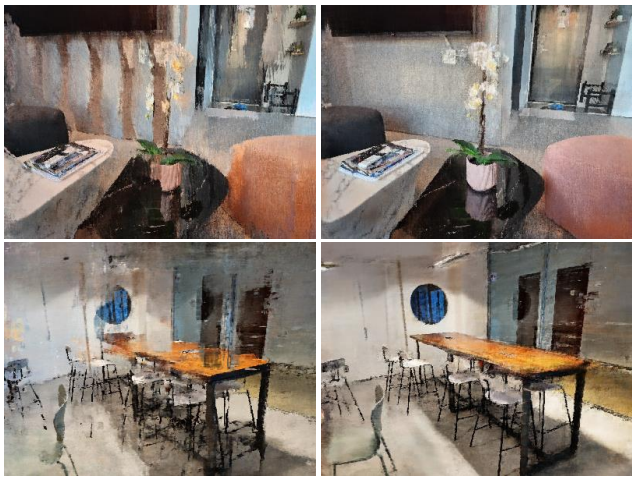
1. Introduction

Neural radiance fields (NeRF) [38] have recently had significant impacts on 3D scene reconstruction and novel view synthesis, and the following works have steadily improved the performance of NeRF in various aspects, such as generalization ability [29, 63, 57, 10, 9, 54, 32, 44], representation ability [2, 24, 66, 22, 4, 49, 27], and practicality [42, 43, 58, 18, 26, 23, 63, 61, 39, 52, 35, 6, 65].

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT)(No.2020-0-00457, 50%) and KIST Institutional Program(Project No.2E32301, 50%).



Sparse inputs with varying illuminations



RegNeRF [41] **ExtremeNeRF (Ours)**

Figure 1: **Few-shot view synthesis results on inputs with varying illuminations.** Our ExtremeNeRF shows reliable novel view synthesis results compared to the state-of-the-art method, RegNeRF [41], in an extremely challenging environment where only 3 view images taken under varying illuminations are available.

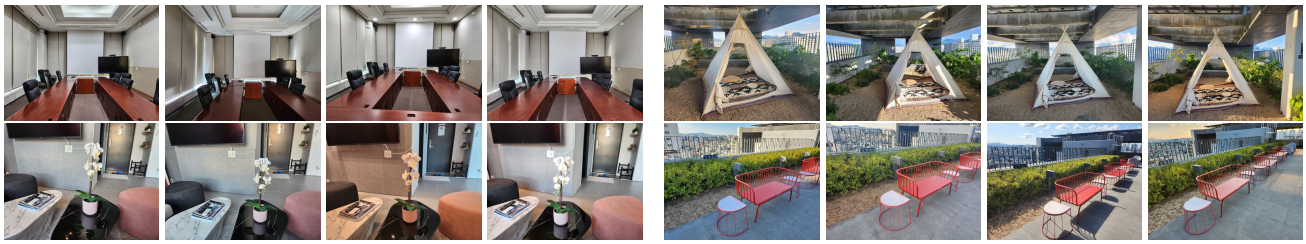


Figure 2: **Examples of NeRF Extreme dataset.** Each scene consists of 10 test images with mild-lighting conditions and 30 train images with at least 3 lighting variations - including variations in color/direction/intensity of lights.

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.

Many studies [57, 56, 14, 63, 41, 61, 50, 16, 33, 30, 64, 55] have succeeded in reducing the number of input images by using prior knowledge or by applying regularization techniques for the task, but they adhere to the assumption of constrained illumination. Fig. 1 shows that the state-of-the-art few-shot view synthesis method [41] fails to synthesize a plausible result. In addition, some works [36, 6, 5, 51, 7] have attempted to overcome the constrained illumination assumption. However, all of the methods have limited practicality in that a sufficient amount of inputs (80 to 200 images) or foreground masks for an object are required.

In this paper, we address the problem of few-shot view synthesis under unconstrained illumination conditions, whose application is not limited to object-centric scenes. Our proposed method, dubbed ExtremeNeRF, can directly regularize the appearances of the unknown view by enforcing consistency among intrinsic components between input and rendered views, which should be independent of viewing direction and illumination. 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 [62]. 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 projective transformation-based geometric alignment and depth consistency loss, which provides correspondences between pixels to compare and encourages correct geometry synthesis.

A newly built NeRF Extreme dataset, which is the first in-the-wild multi-view dataset that has both indoor and outdoor scenes taken under varying illumination, is also proposed (see Fig. 2). Our ExtremeNeRF provides plausible few-shot view synthesis and video rendering results, given about 0.06 times fewer inputs with varying illumination.

2. Related Work

Few-shot NeRF. After Yu et al. [63] has proved that leveraging knowledge priors leads to better few-shot view synthesis, the following works [57, 26, 56, 30, 59, 13] have suggested a variety of priors to improve the performance. Especially, Deng et al. [14] and Xu et al. [61] have presented depth prior based methods, as [46, 15, 29, 3, 64]. Some of the other methods [54, 29, 10, 45, 9, 59] utilize implicit geometry priors for the task. Similar to several previous works [61, 68, 11, 55], our method is based on the projective transformation between different views, but we focus more on view synthesis under varying illumination, not covered by the aforementioned methods.

NeRF under varying illumination. Unlike vanilla NeRF which requires images from strictly-controlled environments, some papers have attempted to handle images taken under challenging real-world environments, e.g., varying illumination. Some works [67, 5, 51, 7, 62, 6, 31] have tried to factorize the scene components based on a NeRF-like framework, but exhaustive computational costs are required. NeRF-W [36] has succeeded in enabling view synthesis of internet photos taken under varying illumination using photometric embeddings. However, the target illumination is limited to the sun. More recently, Tancik et al. [52] presented a way to synthesize a large-scale city scene captured under varying lighting conditions. Boss et al. [6, 5, 7] and Kuang et al. [31] suggest a method to decompose object-centric images for inverse-rendering, allowing for novel view synthesis and relighting. NeROIC [31], in particular, is applicable in few-shot environments but the application is still limited to object-centric images.

Multi-view and/or multi-illumination datasets. Understanding the physical characteristics of a real-world scene

	LLFF [37]	DTU [1]	PT [28]	RTMV [53]	NeRD [5]	SAMURAI [6]	**Ours**
Mult.-illum.	✗	✓	✓	✗	✓	✓	✓
Indoor	✓	✓	✗	✓	✓	✓	✓
Outdoor	✓	✗	✓	✓	✓	✓	✓
Real-world	✓	✓	✓	✓	✓	✓	✓
In-the-wild	✓	✗	✓	✗	✗	✓	✓
Object-centric	✗	✓	✗	✗	✓	✓	✗

Table 1: **Multi-view dataset comparison.** Our NeRF Extreme provides in-the-wild, non-object-centric, and varying illumination images taken from indoor and outdoor scenes.

requires multiple images with varying conditions. However, most of the existing datasets have their focus on varying single attributes like viewing direction or illumination instead of both. The well-known view synthesis benchmarks [37, 53] have a rich variation of viewing directions yet lack lighting variation. On the other hand, the datasets that are collected for the purpose of lighting-estimation [40, 17] or intrinsic decomposition [34] have enough lighting variations but are not collected in a multi-view setup. Some works [1, 36, 5, 6] have built datasets with variation in both attributes, but still have limitations in that their application is limited to outdoor scene collections [28] or object-centric scenes [1, 5, 6] (see Tab. 1).

Intrinsic image decomposition. As Weiss et al. [60] have proved, multiple images with varying illumination can provide rich guidance to intrinsic decomposition. Li et al. [34] has succeeded in disentangling intrinsic components without supervision, by enforcing multiple images with varying illumination to have identical albedo. Das et al. [12] has shown state-of-the-art performance by leveraging cross-color-ratio, the illumination-invariant image gradients in an intrinsic decomposition. We use [34] to extract the pseudo-albedo ground truth of the inputs before the start of the training phase. Recently, Ye et al. [62] have suggested a NeRF framework that enables intrinsic decomposition for a purpose of scene editing. However, did not address the problem of few-shot view synthesis under varying illumination.

3. Preliminaries

Neural radiance fields (NeRF). NeRF [38] is a volume rendering-based 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, respectively, a view-dependent color $\hat{c}(x)$ 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. Similarly, a depth value $\hat{d}(x)$ at x can also be rendered as

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

In vanila-NeRF [38], view synthesis is done by optimizing mean squared error on synthesized color 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.

RegNeRF. RegNeRF [41] uses a depth smoothness regularization of a novel view synthesized patch for few-shot view synthesis, defined as follows:

$$\mathcal{L}_{\text{ds}} = \sum_{x \in \mathcal{S}} \sum_{l \in \mathcal{N}(x)} (\hat{d}(l) - \hat{d}(x))^2, \tag{4}$$

where $\hat{d}(x)$ indicates synthesized depth value at x, and l indicates one of the 4-neighbor adjacent pixels $\mathcal{N}(x)$.

This regularization has succeeded in enhancing the ruined geometry of the synthesized scene. However, RegNeRF [41] also relies on the assumption that input images should share consistent illumination conditions as [38].

4. Methodology

4.1. Overview

The objective of this work is to build an illumination-robust few-shot view synthesis framework by regularizing intrinsic components that should be identical across multi-view images regardless of illumination. However, there exist three major challenges in comparing the albedo extracted from the novel synthesized view and those of input images: 1) Since our proposed method requires pixel-to-pixel correspondence between different views, a geometric alignment is needed to select the pixel to compare. 2) There always exists a non-intersecting or occluded region between the novel view and the input view images, which should be considered during cross-view regularization. 3) The key feature for successful intrinsic decomposition is the global context of a scene. However, the NeRF-based structure has difficulties in rendering full-resolution images since continuous ray sampling upon the entire image requires massive computational and memory costs.

To address these problems, we suggest a few-shot view synthesis framework that utilizes an offline intrinsic decomposition network, providing global context-aware pseudo-albedo ground truth without the computational overhead. As illustrated in Fig. 3, FIDNet [34] provides pseudo-albedo ground truths for the input images before the start of

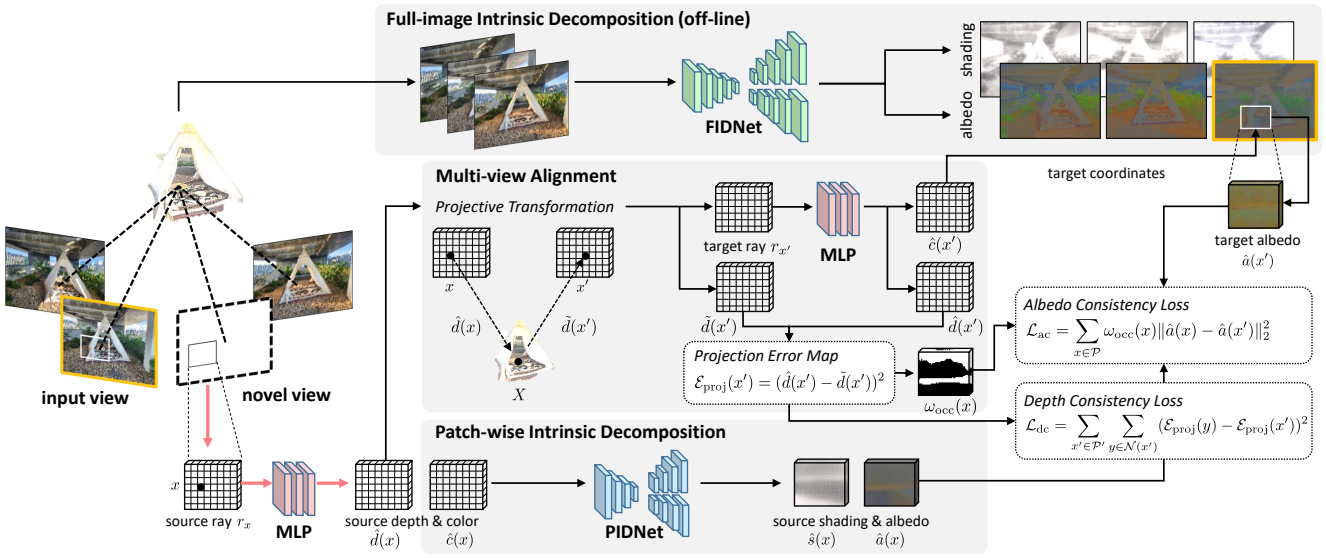


Figure 3: **Overall architecture of our ExtremeNeRF.** PIDNet extracts intrinsic components from the synthesized patch $\hat{c}(x)$ while optimizing albedo consistency $\mathcal{L}_{\text{ac}}$, which enforces extracted albedo to be identical with the pseudo-albedo ground truth. An occlusion-aware weight term $\omega_{\text{occ}(x)}$ and depth consistency loss $\mathcal{L}_{\text{dc}}$ encourage proper correspondence matching between two views. A bold, crimson arrow indicates the inference phase while green and black arrows indicate the offline and online training phases, respectively.

the training. Then PIDNet learns to extract intrinsic components of the synthesized patch, given the pseudo-albedo ground truth of the corresponding patch depicting the same 3D surface. As a result, our NeRF learns illumination-robust few-shot view synthesis.

In the following subsections, we introduce each component of our framework in detail.

4.2. Intrinsic Consistency Regularization

Geometric alignment. As illustrated in Fig. 3, a pair of images is selected from a set of inputs and randomly generated novel views, for every iteration. In order to get a pixel correspondence between the two, we use a projective transformation, similar to several concurrent works [68, 61]. Given a pixel x in the novel view, we need the corresponding image pixel x' in the input view depicting the same 3D point. If the depth of a given image pixel is known, x' can be obtained using projective transformation as follows:

$$x' = (KT'^{-1}T)d(x)K^{-1}x, \tag{5}$$

where K indicates camera intrinsics and T and T' indicate camera-to-world matrices of the novel view and input view. Both K and T are given for the calibrated images.

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. 4). To overcome this, we present $L2$ normalized albedo consistency loss $\mathcal{L}_{\text{ac}}$ formulated as follows:

$$\mathcal{L}_{\text{ac}} = \sum_{x\in\mathcal{P}} \omega_{\text{occ}}(x)\|\hat{a}(x) - \hat{a}(x')\|_2^2, \tag{6}$$

where $\hat{a}(x)$ and $\hat{a}(x')$ indicate the extracted albedo at x and x' from the novel view and input view, respectively. $w_{\text{occ}}(x)$ indicates the weight term to consider inaccurate correspondences coming from occlusions or out-of-region pixels, while $\mathcal{P}$ denotes all the pixels in the novel view. Details of $w_{\text{occ}}(x)$ are in the following.

Occlusion handling. Eq. 5 described above byproducts $\tilde{d}(x')$, a depth value at pixel x' in the input view. $\hat{d}(x')$, a synthesized depth value at x' should be identical to $\tilde{d}(x')$ if there exists neither self-occlusion nor ill-synthesized floating artifacts. For all cases, a projection error on x', denote by $\mathcal{E}_{\text{proj}}$ can by defined as

$$\mathcal{E}_{\text{proj}} = (\hat{d}(x') - \tilde{d}(x'))^2. \tag{7}$$

However, a problem exists in that both inaccurate correspondence and occlusion cause large projection errors. In order to minimize $\mathcal{E}_{\text{proj}}$ that came from inaccurate correspondences while protecting the pixel pairs with occlusion, we define an occlusion-aware weight term w_{occ}, on a pixel x' of input view, as

$$\omega_{\text{occ}} = r_{\text{e}}(1 - (\mathcal{E}_{\text{proj}}/\mathcal{M}_{\text{proj}})), \tag{8}$$

where r_{e} and $\mathcal{M}_{\text{proj}}$ indicates the error rate coefficient and the maximum value of $\mathcal{E}_{\text{proj}}(x)$, respectively. By using

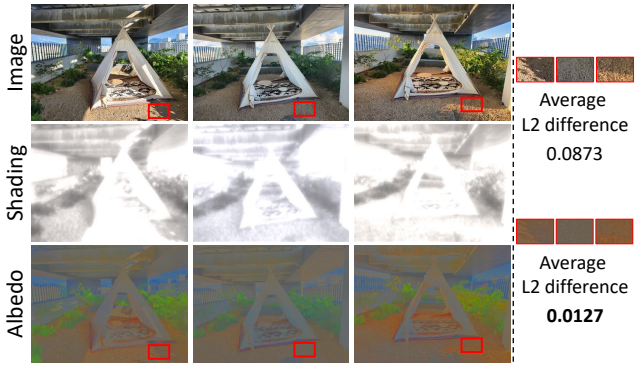


Figure 4: **Examples of intrinsic decomposition on NeRF Extreme.** Estimated albedo maps can provide appearances that are more illumination-invariant than input color images, as shown in a lower difference across multi-views.

w_{occ}, the input-novel view pairs with occlusion that are likely to have large projection errors will not be enforced to have albedo consistency. r_{e} decays from 1 to the end criteria (0.5), reducing the number of pairs that are enforced to have the same albedo, since inaccurate correspondence pairs will decrease while training.

Depth consistency. Geometric alignment in our setup may utilize incorrect synthesized depth values that are commonly observed in novel view synthesis. In order to prevent the model to enforce consistency between pixels with inaccurate correspondence, and to efficiently correct ill-synthesized scene geometry, we present a depth consistency loss $\mathcal{L}_{\text{dc}}$. A direct minimization of $\mathcal{E}_{\text{proj}}$, however, can be counterproductive due to occlusion, by smoothing two unrelated surface depths. Through experimentation, we have discovered that total variation normalization on projection error can better regularize the scene geometry, successfully reducing floating artifacts without suffering adverse effects from occlusion. Depth consistency loss $\mathcal{L}_{\text{dc}}$ in the input view can be defined such that

$$\mathcal{L}_{\text{dc}} = \sum_{x' \in \mathcal{P}'} \sum_{y \in \mathcal{N}(x')} (\mathcal{E}_{\text{proj}}(y) - \mathcal{E}_{\text{proj}}(x'))^2, \qquad (9)$$

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. Details of the projection error cases and depth regularization can be found in the supplementary materials.

4.3. Albedo Estimation

As discussed previously, a successful intrinsic decomposition inevitably requires the global context of a scene, while NeRF struggles with handling large-resolution inputs. Moreover, intrinsic rendering methods [36, 5, 7, 6, 62] cannot be used in our case due to the lack of input data (uses 0.06 times fewer data). To address these challenges without imposing a computational burden or requiring supervision, we propose a two-stage intrinsic decomposition pipeline: a full-image and patch-wise intrinsic decomposition network called FIDNet and PIDNet. Before training begins, FIDNet extracts the intrinsic components of the input images - pseudo-albedo ground truths - offline, in order to provide guidance to PIDNet with global contexts. During training, PIDNet extracts patch-wise intrinsic components ($\hat{a}(x)$) of the synthesized color patch at the novel view ($\hat{c}(x)$). Given the pseudo-albedo ground truth provided by FIDNet, PIDNet is trained to minimize $\mathcal{L}_{\text{ac}}$ (Eq. 15). Specifically, FIDNet [34] uses a shared encoder and 2 decoders, one for log-scale albedo and the other for shading images. The network also predicts a 3-dimensional light color c as a side output. PIDNet follows the architecture with a shallower structure.

4.4. Total Loss Functions

In addition to albedo consistency loss and depth consistency loss, an edge-preserving loss $\mathcal{L}_{\text{edge}}$ [21], an intrinsic smoothness loss $\mathcal{L}_{\text{pid}}$ [34], and chromaticity consistency loss $\mathcal{L}_{\text{chrom}}$ [62] are also used with color consistency loss $\mathcal{L}_{\text{color}}$ and depth smoothness loss $\mathcal{L}_{\text{ds}}$. An edge-preserving loss $\mathcal{L}_{\text{edge}}$ gives the constraint that gradients of the novel synthesized view (i.e. edge) should be identical to the one of the input view. A patch-wise intrinsic smoothness loss $\mathcal{L}_{\text{pid}}$ is formulated in the same way as depth consistency loss. A patch-wise chromaticity consistency loss $\mathcal{L}_{\text{chrom}}$ gives the constraint that the chromaticity of the input patch and the extracted albedo is the same.

5. Experiments

In this section, we introduce our newly built dataset, NeRF Extreme, the first in-the-wild multi-view dataset with varying illumination, whose scenes are not limited to object-centric ones. After that, the experimental settings and results will follow. Details of the dataset statistics and experimental settings are in the supplementary material.

5.1. NeRF Extreme Dataset

To build a view synthesis benchmark that fully reflects unconstrained environments such as mobile phone images captured under casual conditions, we collected multi-view images with a variety of light sources such as multiple light bulbs and the sun. For indoor scenes, we varied the illumination by turn-on/off light sources, and closing/opening curtains. For outdoor scenes, scenes are captured at different times with different sunlights (see Fig. 2). All the images are taken in the wild using the off-the-shelves camera on the mobile phone. Similar to LLFF [37], all the camera poses are obtained using the COLMAP [48] structure-from-motion framework. Depth maps are also included in the dataset, which are obtained by the recent multi-view stereo method [33].

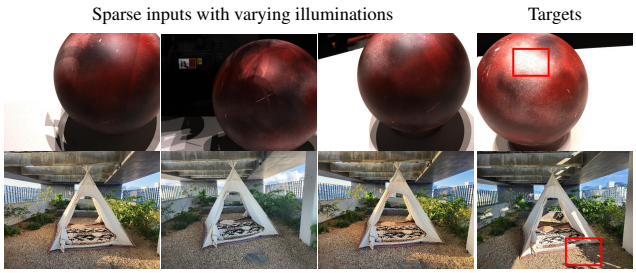


Figure 5: **Ill-posed color synthesis examples.** Given inputs with varying illuminations, surface characteristics and illuminations that are not observed in inputs (red boxes) are impossible to be inferred by any few-shot NeRF methods.

Our dataset consists of 10 scenes, 6 of which are indoor and 4 are outdoor. We took 40 images per scene with a resolution of $3,000 \times 4,000$, with 30 images in the train set and 10 images in the test set. The training images of each scene are captured with at least 3 different lighting conditions. To make the dataset more widely useful, we captured the test scenes under mild lighting conditions, rather than extremely low or high contrasts or intensities, similar to images in typical NeRF benchmark datasets [37, 1].

5.2. Experimental Settings

Implementation details. Our framework is based on JAX [8] implementation of RegNeRF [41]. For FIDNet and albedo consistency measure, official code and publicly available model of IIDWW [34] trained with BigTimes dataset [34] is used without fine-tuning.

Datasets. In addition to our proposed dataset, NeRF Extreme, the experimental results on DTU [1] are provided. For DTU, we adhere to the evaluation protocols that are used by the other previous works [41, 26, 63] except for the lighting variations. For NeRD [5] dataset, only weak comparisons are available for its inadequacy in few-shot view synthesis. More explanations and experimental results are in the supplementary material. Similarly, experiments on SAMURAI are not available for its incapability of camera poses, as mentioned in their paper [6]. In Sec. 5.5, we provide experimental results on LLFF [37] dataset, in addition.

Comparison algorithms. Since none of the previous works are dealing with the suggested problem, both few-shot view synthesis methods and NeRF-based inverse rendering methods, capable of view synthesis under varying illumination are weak baselines for us. For few-shot view synthesis, mip-NeRF [4] and RegNeRF [41], which are baseline and the state-of-the-art few-shot NeRF, respectively, are compared. In the case of view synthesis under varying illumination, however, comparisons with most of the baselines are unavailable. For NeRF-W [36], neither the implementation nor the pre-processed data are publicly available. Plus, it can only handle outdoor scenes. NeROIC [31] and the other works [5, 7, 6] cannot deal with in-the-wild scenes like our NeRF Extreme, as their inputs are limited to object-centric scenes paired with foreground masks. This limitation only allows for weak comparisons. In the supplementary material, we provide an additional comparison with the baselines.

5.3. Evaluation Metrics

The problem of evaluating a novel view synthesis under varying illumination has rarely been discussed since inferring consistent colors and synthesizing a view with unseen illumination is an ill-posed problem. Fig. 5 illustrates the situations where scene appearances are impossible to be synthesized using the information provided by the inputs. Although our method succeeds in synthesizing the plausible novel view, the physical characteristics like surface reflectance (1st row) or unknown illumination (2nd row) that are not observed in input images cannot be synthesized. Similar works [36, 6] have dealt with the problem by evaluating typical metrics like PSNR and SSIM on the synthesized image after relighting to fit the target illumination. In this paper, we suggest evaluation methods in addition to PSNR and SSIM, that can compare the underlying characteristics of the scene regardless of illumination.

CCRD. Geverse et al. [19] has presented the cross-color-ratio (CCR), an illumination-invariant image gradient that only depends on albedo transitions. Using the log-scale CCR image proposed by [12], we use the cross-color-ratio difference (CCRD), a metric that can measure the maintenance of underlying consistent color from the synthesized image to the inputs, as $L1$ difference between CCR images of the ground truth and the synthesized view.

Absolute relative error. In order to compare the quality of the synthesized depth, we utilize Absolute Relative Error(Abs Rel), which is commonly used in the depth estimation task. For DTU [1], provided ground-truth depth maps are used for the measurement. In the case of our NeRF Extreme, a depth map estimated by [20] is used as a pseudo-ground-truth. The estimated depth maps were publicly available within our dataset.

5.4. Experimental Results

Qualitative comparison. Fig. 6 and Fig. 7 show the qualitative comparison results between our ExtremeNeRF and other few-shot view synthesis baselines on NeRF Extreme and light-varying DTU [1] with 3 input view setting. Our method outperforms other baseline methods [4, 41] in synthesizing geometry and eliminating distortions that come from varying illumination inputs. Especially, for depth maps, our method shows a large improvement compared to the baselines, even when the baseline succeeds in synthesizing a color image with competitive quality (Fig. 8), enabling

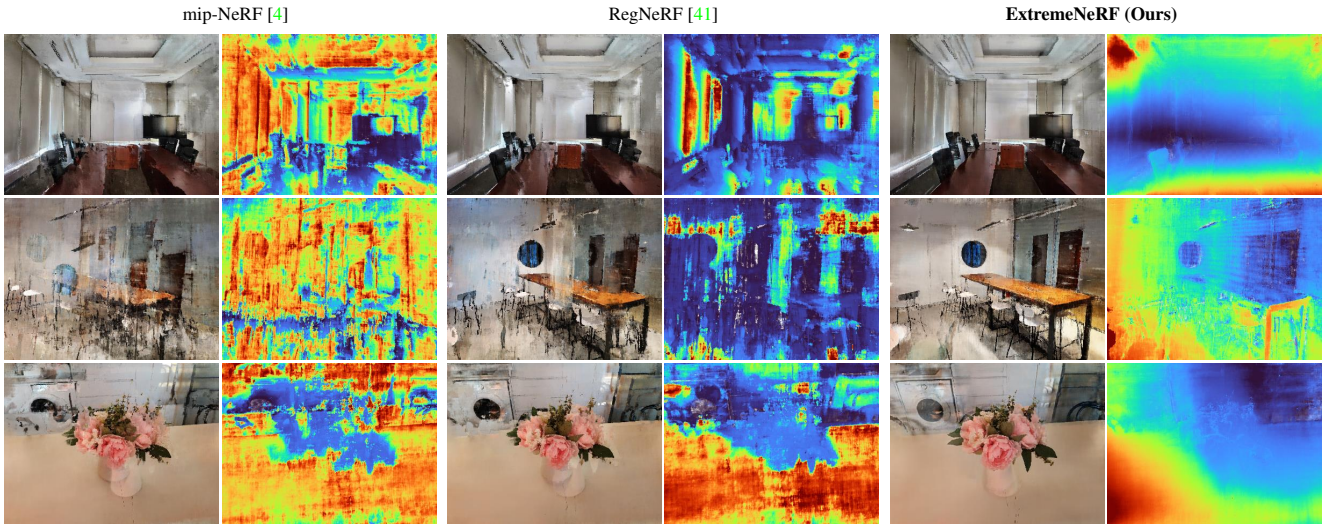


Figure 6: **Qualitative comparison on NeRF Extreme.** A synthesized novel view and corresponding depth map are generated by the baselines and our proposed method with 3 view input images. Our proposed method shows plausible synthesis results compared to the other baseline methods (Best viewed in color).

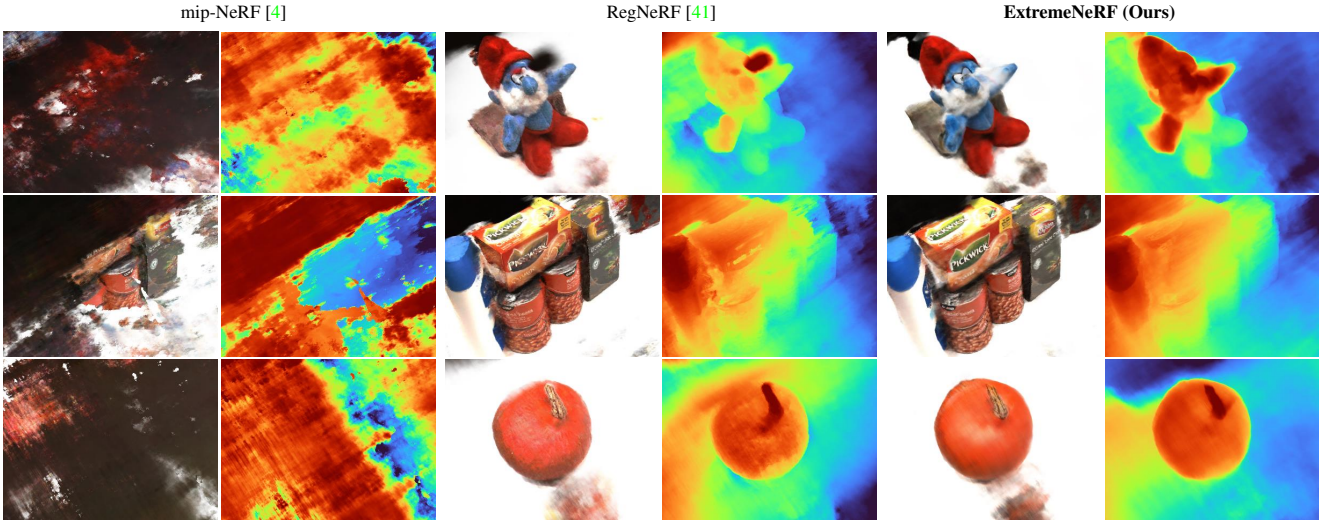


Figure 7: **Qualitative comparison on light-varying DTU [1].** Given light-varying 3 view input images, our proposed method shows reliable synthesis qualities both for the color and the depth maps, while RegNeRF [41] shows floating artifacts and ill-synthesized depths (Best viewed in color).

further applications such as video rendering.

Quantitative comparison. Tab. 2 and 3 shows a quantitative comparison between the methods on 3 view settings. For all cases within NeRF Extreme, our method outperforms others. Our method shows the lowest CCRD and Abs Rel with a large difference indicating that our ExtremeNeRF can maintain the underlying physical properties of the target scene even with sparse inputs with varying illumination. In the case of light-varying DTU [1], performance degradation can be found, compared to the results of NeRF Extreme. However, despite the absence of image global contexts, our method shows the lowest CCRD, proving robustness against illumination changes. More qualitative and quantitative results are in the supplementary material.

5.5. Ablation Studies

In this section, ablation studies on various components of our method are provided. We found that cross-view consistency regularization neglecting illumination has an adverse effect on the underlying color consistency and depth map. Further, the depth map shows improvement by not enforcing albedo consistency for cases suspected of occlusion.

Depth consistency regularization. The 2nd rows of Tab. 4 shows ablation on depth consistency term $\mathcal{L}_{\text{dc}}$. The

	CCRD ↓		Abs Rel ↓		PSNR ↑		SSIM ↑		LPIPS ↓	
	3-view	6-view	3-view	6-view	3-view	6-view	3-view	6-view	3-view	6-view
mip-NeRF [4]	18.27	17.43	0.32	0.30	12.92	13.83	0.26	0.30	0.55	0.49
RegNeRF [41]	18.13	16.75	0.31	0.32	14.36	15.02	0.32	0.37	0.51	0.44
ExtremeNeRF (Ours)	16.97	16.44	0.26	0.26	14.86	15.18	0.36	0.40	0.44	0.42

Table 2: **Quantitative comparison on NeRF Extreme.** We compare our quantitative result with the few-shot view synthesis baselines that can be operated in the same experimental setting as ours. For every case, our model shows the best synthesizing performance for each metric.

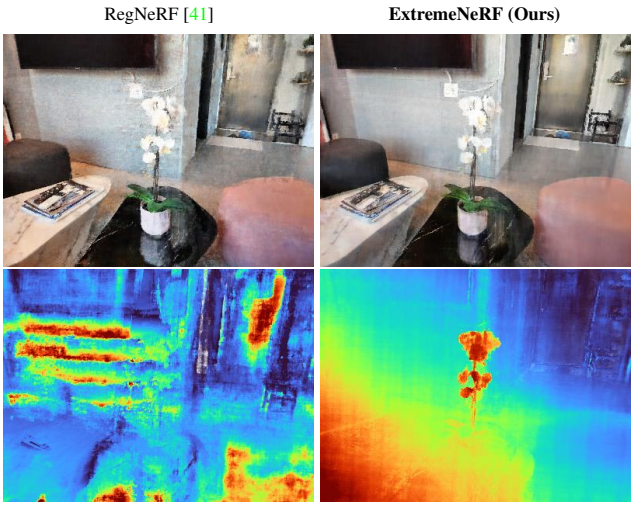


Figure 8: **Examples of rendered depths with corresponding rendered colors.** Even though a synthesized result of the previous work [41] often shows the competitive results as us in 6 view settings, corresponding depth information is highly distorted compared to ours.

	CCRD ↓	Abs Rel ↓	PSNR ↑	SSIM ↑	LPIPS ↓
mip-NeRF [4]	19.78	0.35	7.62	0.20	0.66
RegNeRF [41]	17.91	0.27	14.89	0.59	0.34
ExtremeNeRF (Ours)	17.72	0.29	14.49	0.58	0.36

Table 3: **Quantitative comparison on light-varying DTU [1] with 3 views.** Our method has strength in maintaining the underlying color characteristics of the target scene, as shown in lower CCRD.

	CCRD ↓	Abs Rel ↓	PSNR ↑	SSIM ↑	LPIPS ↓
RegNeRF [41]	18.13	0.31	14.36	0.32	0.51
RegNeRF w/ $\mathcal{L}_{\text{dc}}$	17.47	0.29	14.77	0.36	0.46
Ours w/o $\mathcal{L}_{\text{ac}}$	17.00	0.29	14.86	0.37	0.43
Ours w/o ω_{occ}	17.37	0.30	14.76	0.36	0.44
Ours ($r_e = 1.0$)	17.10	0.30	15.00	0.37	0.43
Ours ($r_e = 0.7$)	17.00	0.27	14.83	0.36	0.45
Ours ($r_e = 0.5$)	16.97	0.26	14.86	0.36	0.44

Table 4: **Ablation study of our ExtremeNeRF.**

results prove that our depth consistency term is not only beneficial for the depth map but also for the underlying color consistency. This is due to a decrease in the number of ill-synthesized pixels. Additionally, we ablate our

	CCRD ↓	Abs Rel ↓	PSNR ↑	SSIM ↑	LPIPS ↓
RegNeRF [41]	15.63	0.32	15.92	0.35	0.38
RegNeRF w/ $\mathcal{L}_{\text{dc}}$	15.45	0.32	16.01	0.36	0.37

Table 5: **Ablation study for depth consistency term on LLFF [37].**

depth consistency term on the view synthesis benchmark with constrained illumination [37] (Tab. 5). A lower CCRD proves that our depth consistency term helps to maintain underlying color consistency even when the unconstrained illumination assumption is absent.

Albedo consistency regularization. The ablation on $\mathcal{L}_{\text{ac}}$ (the 3rd rows of Tab. 4) was performed by enforcing view-dependent color consistency across views instead of albedo, similar to Eq. 3. Interestingly, the experimental results show the 2nd lowest CCRD, since the synthesized color neglects illumination and viewing direction - i.e. similar to albedo. However, the marginal decrease in Abs Rel proves that it fails to regularize depth compared to ours.

Occlusion handling. Occlusion handling in our method is in the role of preventing consistency regularization between the pixels with occlusions. The last 4 rows in Tab. 4 show ablation studies on occlusion handling. The ablation on ω_{occ} proves that consistency regularization across views without ω_{occ} can only bring minor changes in performance, especially for depth. In the case of the error rate coefficient r_e, the Abs Rel of the synthesized depth has increased, if r_e is kept at 1.0 without decreasing it. This is due to the enforced consistency between the pixel pairs with occlusions. Note that the best error rate coefficient r_e may vary depending on the target scene geometry. In our experiments, decaying r_e from 1 to 0.5 shows the best results on our NeRF Extreme dataset, which is likely to have many occlusions.

6. Conclusion

In this paper, we propose ExtremeNeRF, which can synthesize a novel view image in more practical environments, where neither a massive amount of multi-view images nor constrained illumination is available. By regularizing intrinsic components which should be identical across different views, our method can directly regularize appearance

instead of interpolating view-dependent color as vanilla-NeRF did. Supported by geometry alignment and depth consistency, our pipeline enables intrinsic decomposition while considering global contexts with marginal computational overhead. We have proved with our multi-view, varying-illumination dataset that the proposed method outperforms other previous works in a few-shot view synthesis under an unconstrained illumination environment, with extended applicability on non-object-centric scenes.

Limitations. More practical view synthesis under unconstrained environments requires the consideration of scenes with unknown camera poses and transient components. It should be addressed in the future, as it has rarely been discussed in few-shot, varying illumination settings.

Appendix

A. Additional Experimental Results

A.1. Comparisons with NeROIC

As mentioned in the paper, we would like to emphasize that none of the previous works deal with the same problem as us, which is a few-shot views synthesis of a non-object-centric scene under unconstrained illumination. In the paper, we provide comparisons with the baselines of Barron et al. [4] and Niemeyer et al. [41]. While NeROIC [31] and other related works such as NeuralPIL [7], NeRD [5], and SAMURAI [6] deal with view synthesis under varying illumination, they have several differences from us, which we describe as follows.

- All of these works can only handle object-centric, 360 scenes paired with foreground masks, but our proposed method is targeting non-object-centric, forward-facing scenes without additional masks.
- Previous works [31, 7, 5, 6] are an inverse rendering framework, which aims to decompose images into their geometry, material, and illumination. Therefore, they require massive computational costs with multi-stage training. Our proposed work is the first end-to-end few-shot synthesis framework that handles inputs with unconstrained illuminations with minimum computational costs (3.5 hours in a 3-view setting).
- Boss et al. [5] and following works [7, 6] does not assume few-shot settings. Especially, SAMURAI [6] assumes unknown camera poses of inputs, unlike other previous works and our proposed method.

In the following, we additionally provide comparisons with NeROIC, on our NeRF Extreme dataset. Note that NeROIC is the only baseline that performs few-shot view synthesis with varying illumination inputs, which shows better performance compared to other baselines (detailed experimental results are on their paper, [31]) under unconstrained illumination conditions. Comparisons with NeRF-W [36] are not available since 1) it is targeting outdoor scenes and 2) neither the implementation nor the pre-processed data are publicly available.

	PSNR ↑	SSIM ↑	LPIPS ↓
NeROIC [31]-Geom	13.06	0.25	0.64
Extreme NeRF (Ours)	14.86	0.36	0.44

Table 6: **Quantitative comparison on NeRF Extreme.** Note that results from NeROIC-Full are not included for their diverged outputs.

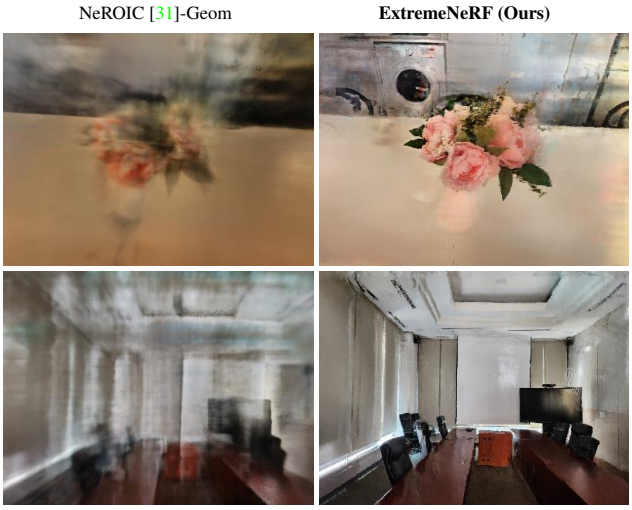


Figure 9: **Qualitative comparison with NeROIC [31] on NeRF Extreme.** A synthesized novel view by NeROIC [31]-Geom and our proposed method in 3-view setting. NeROIC-Geom shows unclear, distorted results compared to ours. Note that synthesized results of NeROIC-Full are not included since the network diverged.

Comparison on NeRF Extreme. We compare our model with NeROIC on our NeRF Extreme dataset in the 3-view setting, Tab. 6 and Fig. 9 show the quantitative and qualitative comparison results, respectively. NeROIC-Geom refers to the geometry network of NeROIC while NeROIC-Full refers to the full rendering network of NeROIC. Both NeROIC-Geom and NeROIC-full show over-smoothed or diverged results on our dataset, while our proposed ExtremeNeRF shows plausible view synthesis results. The results demonstrate the fact that none of the previous works can successfully perform few-shot view synthesis under unconstrained illumination if the target scene is the non-object-centric one. See Sec. B for more details about NeROIC.

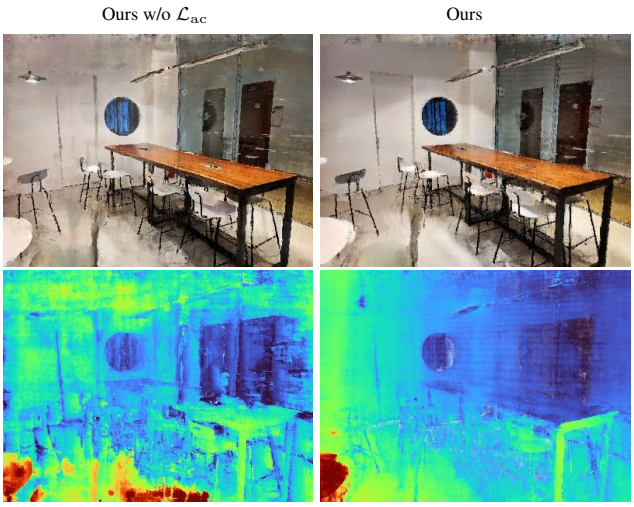


Figure 10: **Qualitative comparison on ablation studies.** Ablation on our cross-view albedo consistency loss results in degraded synthesizing performance, especially in the depth map.

Figure 11: **Relighting results.**

A.2. Ablation Studies

As shown in Tab. 4 of the paper, our model with ablation on each component show a larger difference in the quality of the underlying depth. For better analysis, we provide additional qualitative comparison which visualizes the difference in the quality of the synthesized depth. Fig. 10 shows the synthesized color and depth map with and without cross-view albedo consistency regularization. Our model with albedo consistency regularization shows a better performance in synthesizing depth compared to the one without albedo consistency. It also supports the result described in Fig. 8 of the paper, that neglecting illumination can result in distorted depth in few-shot view synthesis under unconstrained illumination.

A.3. Relighting

Unlike the other previous works that tried to achieve inverse rendering of a scene given inputs with unconstrained illumination, our proposed method focuses on a few-shot view synthesis task for practical usage. The architectural choice results in better computational efficiency, however, has its limitation in that it cannot explicitly decompose the illumination of the target scene. In this supplementary material, we additionally provide relighting results of our proposed method, inspired by the image formation model we use during intrinsic decomposition. As described in Eq. 17, our inputs and synthesized images can be decomposed by their illumination-invariant color (albedo), illumination-variant shading, light color, and non-Lambertian residuals. Because our proposed ExtremeNeRF provides per-scene optimization, we can relight the synthesized image by replacing the shading image, as long as the synthesized image remains plausible. Fig. 11 shows the relighting results, achieved by replacing the shading image of the synthesized scene. Note that shiny, blurry effects in the scenes are caused by [34], which is used as our full-image intrinsic decomposition method.

A.4. Failure Cases

If the outdoor scenes have challenging illumination, such as the *tent* scene in our dataset, our proposed method may struggle to synthesize plausible results. Figure 12 shows the synthesized results for the *tent* scene produced by our method and the baseline method from [41]. The less-qualified results may be due to inadequate decomposition of the intrinsic components of the scene, which causes the cross-view regularization to fail. However, despite the difficulties in regularizing the intrinsic components, our proposed method still shows better synthesizing performance compared to the baseline, which produces a highly distorted color image.

A.5. Additional Qualitative Results

NeRF Extreme. Figure 13 shows additional qualitative comparisons for our NeRF Extreme benchmark, using 3, 6, and 9 view scenarios. Our method performs better than the others, especially in synthesizing depth. This demonstrates that our albedo regularization pipeline successfully removes the distortions that can arise in an unconstrained illumination environment.

Light-varying DTU. Figure 14 shows additional qualitative results for DTU [1] using 3 view scenarios. Although our proposed method sometimes produces competitive or inferior quantitative results compared to RegNeRF [41] (as shown in Table 3 in the paper), it often produces less dis-

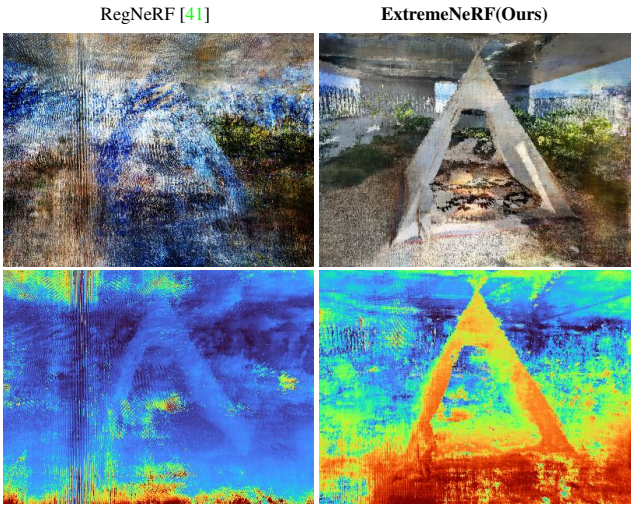


Figure 12: **Failure cases.**

torted results for both color and depth maps. We would like to emphasize that the DTU dataset is not an ideal input for our method, due to its insufficient global context, which is important for successful intrinsic decomposition.

B. Experimental Details

In this section, we provide details about the proposed methods and experiments.

B.1. Datasets

NeRF Extreme. As already mentioned in the paper, we would like to emphasize that our proposed NeRF Extreme dataset, is the first frontal-facing, in-the-wild multi-view dataset with varying illumination. A detailed explanation of the dataset statistics can be found in Sec. C. Fig. 15 shows the indoor and outdoor scenes included in our dataset. In the experiments, we use image IDs of (0, 14, 29), (0, 5, 12, 17, 28, 29), (0, 4, 7, 11, 14, 18, 22, 25, 29) for 3, 6, and 9 views scenarios, respectively. All images are used in a resolution of 300×400.

Light-varying DTU. DTU [1] consists of images taken under structured cameras and light sources. There exist 7 number of lighting variations per scene. Previous works which deal with view synthesis under consistent illumination [38, 26, 63, 10, 4, 41, 56, 30, 9, 61] have used the dataset with fixed mild lighting condition.In this paper, we randomly choose lighting conditions for each scene. Note that view synthesis performance may vary a lot depending on which viewing directions and lighting conditions are selected. Following the evaluation protocol used by the previous works [63, 41], we use scan IDs (8, 21, 30, 31, 34, 38, 40, 41, 45, 55, 63, 82, 103, 110, 114) as the dataset, while using image IDs (25, 22, 28) as inputs. In the cases of the lighting condition, (4, 1, 5) are used for (25, 22, 28) images, respectively. All images are used in a resolution of 300×400, following the evaluation protocols of the previous work [41]. Note that all metrics are calculated without masks.

B.2. Baselines

In the cases of RegNeRF [41], the official configurations for DTU [1] and LLFF [37] are used for evaluations of light-varying DTU and NeRF Extreme, respectively, owing to their identical or similar data characteristics. For light-varying DTU, 44K train epochs are used for 3 view cases, for both RegNeRF and ours. For NeRF Extreme, 64K, 140K, and 21K train epochs are used for 3, 6, and 9 view cases, respectively. Our ExtremeNeRF shares the configurations with RegNeRF. Note that the official implementation of RegNeRF lacks color regularization done by the normalizing-flow model. In the case of NeROIC [31], we use the publicly available instructions and code for the experiments. Since the authors have notified that the proposed method is not able to handle the frontal-facing scenes, we adjust the implementation code to be able to run on the frontal-facing scene. Instead of the foreground masks which should be provided to run the code, we enable all the image pixels to be regarded as foreground ones. It may result in degraded performance of the baseline, however, is the only way to make a comparison. We follow the provided configurations for the training options.

B.3. Metrics

Cross-Color-Ratio (CCR). Cross-color-ratio (CCR) [19] refers to the illumination-invariant image gradients that only depend on the albedo transitions of an image. For two adjacent RGB pixels x_1, x_2, CCRs are defined as:

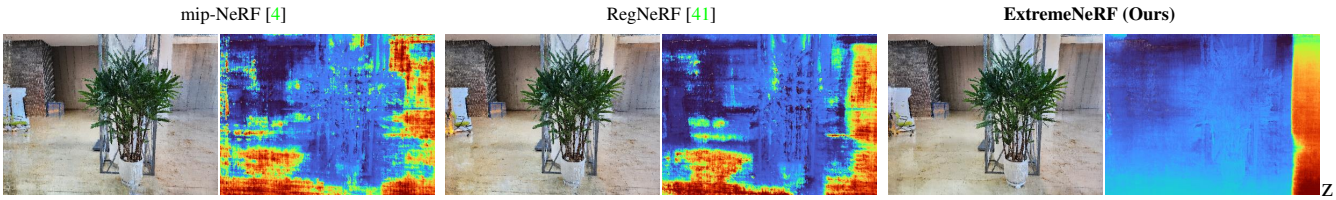
$$M_{RG} = \frac{R_{x_1} G_{x_2}}{R_{x_2} G_{x_1}}, M_{RB} = \frac{R_{x_1} B_{x_2}}{R_{x_2} B_{x_1}}, M_{GB} = \frac{G_{x_1} B_{x_2}}{G_{x_2} B_{x_1}}, \tag{10}$$

Cross-Color-Ratio-Difference (CCRD). According to Das et al. [12], CCR reflects albedo properties sufficiently to improve the performance of intrinsic decomposition. By taking the logarithm of both side of Eq. 10 as suggested by [12], our cross-color-ratio-difference (CCRD) metric can be formulated as:

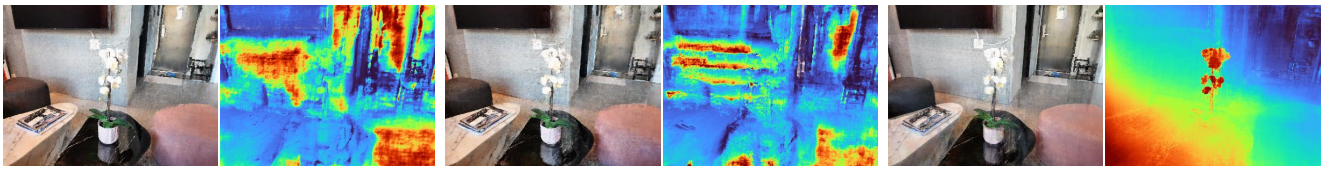
$$\text{CCRD} = \sum_{x \in \mathcal{P}} |\text{ccr}(x) - \text{ccr}_{\text{GT}}(x)| \tag{11}$$

where $\text{ccr}(x)$ and $\text{ccr}_{\text{GT}}(x)$ indicate log-scale CCR values at a pixel x of the synthesized image and the ground-truth image, respectively, and $\mathcal{P}$ denotes all pixels of the image.

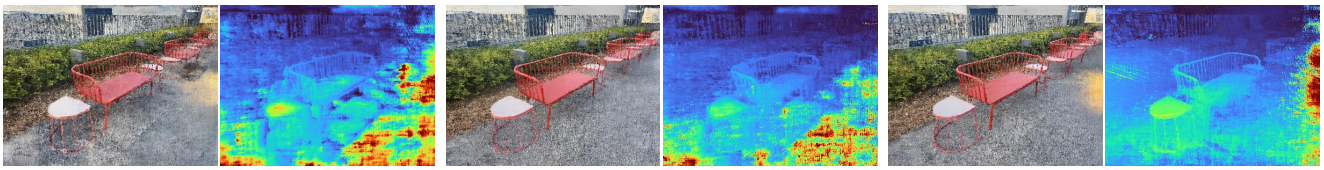
Absolute Relative error (Abs Rel). Absolute Relative error (Abs Rel) is one of the most commonly used metrics



(a) 3 Input Views



(b) 6 Input Views



(c) 9 Input Views

Figure 13: **Additional qualitative results on NeRF Extreme.**

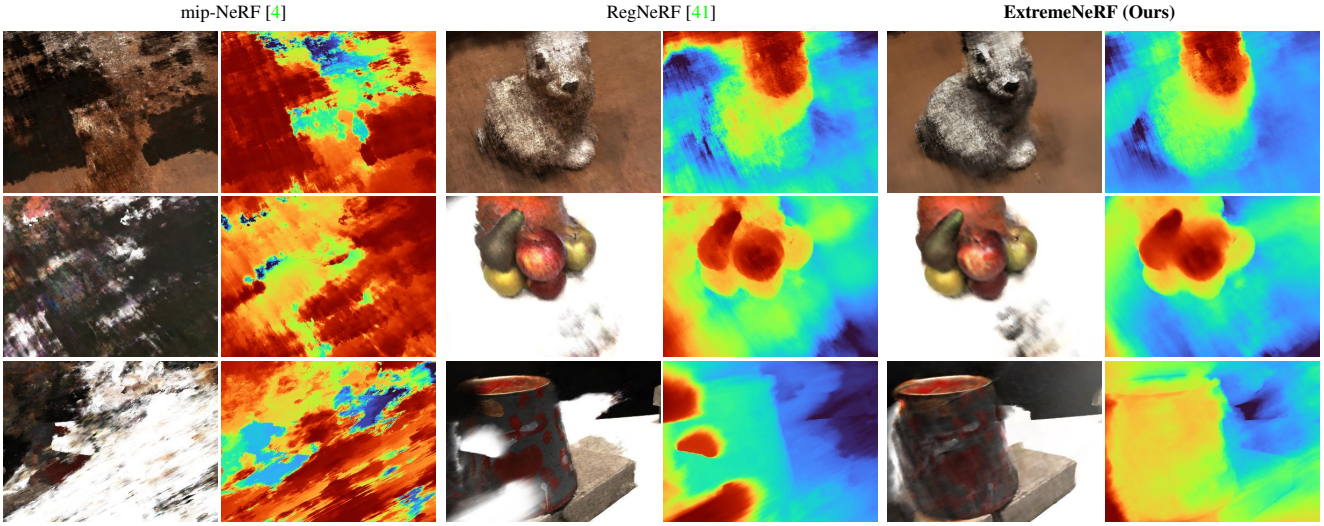


Figure 14: **Additional qualitative results on light-varying DTU [1] for 3 input views.**

for depth estimation tasks. In order to measure the scale-invariant plausibility of the synthesized depth, we adopted Abs Rel as:

$$\text{Abs Rel} = \sum_{x \in \mathcal{P}} |\bar{d}(x) - \bar{d}_{\text{GT}}(x)|, \tag{12}$$

where $\bar{d}(x)$ and $\bar{d}_{\text{GT}}(x)$ indicate the synthesized depth and the ground truth depth, which are normalized to scale 0 to 1, respectively. Since most of the view-synthesis datasets except DTU [1] are not providing depth information, we estimated the depth maps by [20] and used them as a pseudo-depth ground truth.

B.4. Network Architectures

FIDNet architectures. IIDWW[34] uses a variant of UNet[25, 47] architectures with a shared encoder and 2 decoders. A 3-dimensional light color c is also predicted as a by-product of the network. The publicly available model takes 256×384 sized full-resolution images as inputs. Images are resized before and after intrinsic decomposition.

PIDNet architectures. As a downsized model of FIDNet, PIDNet consists of 4 numbers of 4×4 sized convolution/deconvolution layers that are connected to each other using

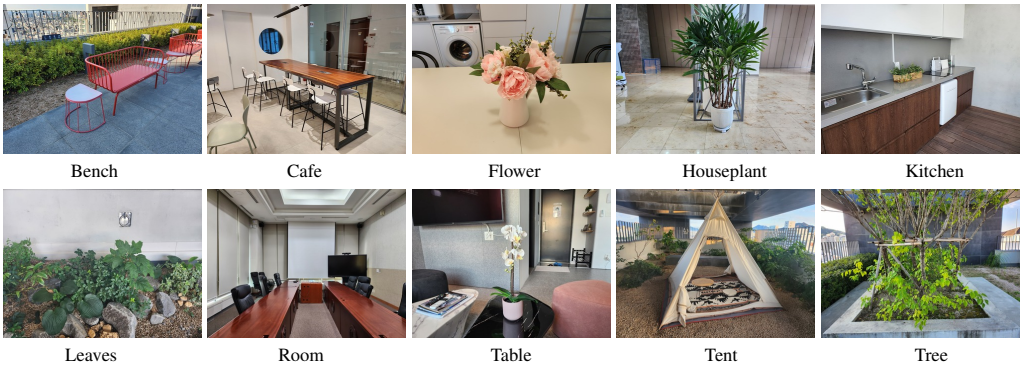


Figure 15: **NeRF Extreme dataset.** This dataset is newly built to provide multi-view images under varying illumination, which can be used to train and evaluate the robust NeRF. Exemplified images are selected from their test sets with mild-lighting conditions.

	Name	Filter	Input	Output	Connectivity
Encoder	Conv1	kernel=(4×4), strides=(2×2)	$32 \times 32 \times 3$	$16 \times 16 \times 32$	-
	(LeakyRelu-Conv2-BN)	kernel=(4×4), strides=(2×2)	$16 \times 16 \times 32$	$8 \times 8 \times 64$	-
	(LeakyRelu-Conv3-BN)	kernel=(4×4), strides=(2×2)	$8 \times 8 \times 64$	$4 \times 4 \times 128$	-
	(LeakyRelu-Conv4)	kernel=(4×4), strides=(2×2)	$4 \times 4 \times 128$	$2 \times 2 \times 256$	-
Decoder	(LeakyRelu-Deconv1-BN)	kernel=(4×4), strides=(2×2)	$2 \times 2 \times 256$	$4 \times 4 \times 128$	$4 \times 4 \times 128$
	(LeakyRelu-Deconv2-BN)	kernel=(4×4), strides=(2×2)	$4 \times 4 \times 256$	$8 \times 8 \times 64$	$8 \times 8 \times 64$
	(LeakyRelu-Deconv3-BN)	kernel=(4×4), strides=(2×2)	$8 \times 8 \times 128$	$16 \times 16 \times 32$	$16 \times 16 \times 32$
	(LeakyRelu-Deconv4)	kernel=(4×4), strides=(2×2)	$16 \times 16 \times 64$	$32 \times 32 \times 3$ (albedo) $32 \times 32 \times 1$ (shading)	-
Color Prediction	AvgPool	kernel=(1×1)	$2 \times 2 \times 256$	$2 \times 2 \times 256$	-
	(LeakyRelu-Conv5)	kernel=(3×3), strides=(1×1)	$2 \times 2 \times 256$	$2 \times 2 \times 128$	-
	(LeakyRelu-Flatten-FC)	-	$2 \times 2 \times 128$	3	-

Table 7: **Patch-wise intrinsic decomposition network(PIDNet) architecture.**

skip-connections, with additional FC layers for light color prediction. Tab. 7 shows the details of PIDNet. Note that PIDNet takes 32×32 sized patch as an input. Decoders for albedo and shading are identical except for the last channel dimension (3 for the albedo and 1 for the shading image).

B.5. Loss Functions

Edge-preserving loss. Motivated by [21], we use the gradient-based edge-preserving loss, to enforce the input and the novel view patches to preserve geometric properties. Using an occlusion-aware weight term, $\omega_{occ}(x)$, which already has been discussed in the paper, our edge-preserving loss on the predicted albedo can be formulated as:

$$\mathcal{L}_{\text{edge}} = \sum_{x' \in \mathcal{P}'} \omega_{occ}(x) \|\partial(\hat{a}(x) - \hat{a}(x'))\|^2, \tag{13}$$

where ∂ denotes the partial derivatives of the vertical and the horizontal directions, and $\mathcal{P}'$ are all the pixels in the target image.

Patch-wise intrinsic smoothness loss. Similar to the previous work [34] which uses various kinds of smoothness terms to give constraints to the network, we give smoothness constraints to our patch-wise intrinsic decomposition network (PIDNet). Our patch-wise intrinsic smoothness loss is formulated as:

$$\mathcal{L}_{\text{pid}} = \sum_{x' \in \mathcal{P}'} \sum_{y \in \mathcal{N}(x')} \|\hat{a}(y) - \hat{a}(x')\|^2, \tag{14}$$

where y is one of the 4-neighbor adjacent pixels $\mathcal{N}(x')$ for x', and $\mathcal{P}'$ denotes all the pixels of the target image.

Chromaticity consistency loss. Similar to [62], we adopted the chromaticity consistency loss to enforce the consistency between the input and novel view patches. Our chromaticity consistency loss is formulated as:

$$\mathcal{L}_{\text{chrom}} = \sum_{x' \in \mathcal{P}'} \|\hat{ch}(x) - \hat{ch}(x')\|_2^2, \tag{15}$$

where $\hat{ch}(x)$ and $\hat{ch}(x')$ indicate the extracted albedo at x and x' from the novel view and the input view, respectively. $\mathcal{P}$ denotes all the pixels in the novel view.

Total loss functions. For each x and x' in the batch-wisely sampled input and the novel view patch, the total loss of our proposed framework is

$$\begin{aligned}\mathcal{L}_{\text{total}} =& \lambda_{\text{c}}\mathcal{L}_{\text{color}} + \lambda_{\text{a}}\mathcal{L}_{\text{ac}} + \lambda_{\text{dc}}\mathcal{L}_{\text{dc}} + \lambda_{\text{ds}}\mathcal{L}_{\text{ds}} \\ &+ \lambda_{\text{e}}\mathcal{L}_{\text{edge}} + \lambda_{\text{pid}}\mathcal{L}_{\text{pid}} + \lambda_{\text{chrom}}\mathcal{L}_{\text{chrom}},\end{aligned} \tag{16}$$

Algorithm 1 Depth Consistency, Pytorch-like

```
def get_corresponding_coords(K, coords, c2w_1, c2w_2
    , depth):
    im_coord = coords*depth
    cam_coord = np.linalg.inv(K) @ im_coord
    w2c = np.linalg.inv(c2w_2)
    cam_mat = w2c @ c2w_1
    return K @ cam_mat @ cam_coord

def init_mask_tg_rays(tg_coords, full_rays):
    lrc_mask = np.ones_like(tg_coords)
    lrc_mask = (0 if outside of image region else 1)
    tg_rays = full_rays[tg_coords]
    return lrc_mask, tg_rays

# src renderings from MLP
src_renderings = model.apply(src_rays)
depth = src_renderings['depth_pred']

# calc target coord & rays
uvd_target = get_corresponding_coords
        (K, src_coords, src_pose, tg_pose, depth)
tg_coords = uvd_target[:2,:]//uvd_target[2,:]

depth_tilde = uvd_target[2,:]
lrc_mask, tg_rays = init_mask_tg_rays
        (tg_coords, tg_full_rays)

# tg renderings from MLP
tg_renderings = model.apply(tg_rays)
depth_hat = tg_renderings['depth_pred']

error_proj = (depth_tilde - depth_hat)**2

loss_dc = np.mean(tv_norm(error_proj))
```

Algorithm 2 Albedo Consistency, Pytorch-like

```
src_patch = src_renderings['rgb']
src_shading, src_albedo, src_rgb = pidnet.apply(
    src_patch)

# intrinsic smoothness loss
pid_loss = np.mean(tv_norm(src_albedo))

# get target albedo from FIDNet result
tg_albedo = tg_full_albedo[tg_coords]

tg_chrome = rgb_to_chromaticity(tg_albedo)
src_chrome = rgb_to_chromaticity(src_albedo)
src_patch_chrome = rgb_to_chromaticity(src_patch)

# chromaticity consistency loss
loss_chrom = np.mean((src_chrome - tg_chrome)**2
        + (src_chrome - src_patch_chrome)**2)

# scale value using least square
tg_albedo = ls_scale_val(src_albedo, tg_albedo)

occ_weight = r_e * (1- error_proj/max(error_proj))

# albedo consistency loss
loss_ac = np.mean(occ_weight*(src_albedo - tg_albedo
    )**2)

src_grad = np.gradient(np.exp(lrc_mask*src_albedo))
tg_grad = np.gradient(np.exp(lrc_mask*tg_albedo))

# edge preserving loss
loss_edge = np.mean(occ_weight*(src_grad - tg_grad)
    **2)
```

where $\lambda_{\mathrm{c}}, \lambda_{\mathrm{a}}, \lambda_{\mathrm{dc}}, \lambda_{\mathrm{ds}}, \lambda_{\mathrm{edge}}, \lambda_{\mathrm{pid}}, \lambda_{\mathrm{chrom}}$ are weights parameters for each loss, respectively. In our experiment, losses are weighted as: $\lambda_{\mathrm{c}} = 1.0$, $\lambda_{\mathrm{a}} = 1.0$, $\lambda_{\mathrm{dc}} = 1.0$, $\lambda_{\mathrm{ds}} = 0.1$, $\lambda_{\mathrm{edge}} = 0.01$, $\lambda_{\mathrm{pid}} = 1.0$, $\lambda_{\mathrm{chrom}} = 0.01$. We provide detailed algorithms in Alg. 1 and 2.

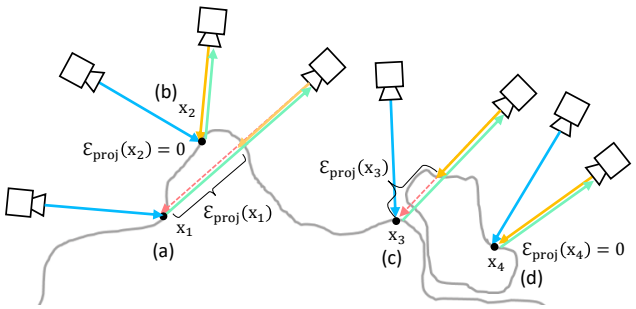


Figure 16: **Projection situation examples.** Blue and yellow arrows are the synthesized depths $\hat{d}(x)$ and $\hat{d}(x')$, respectively, while green arrow is $\tilde{d}(x')$ obtained by projective transformation.

B.6. Occlusion Handling

Problem situations. In Fig. 16, we illustrate possible situations that may occur in the geometry alignment stage.

- (a): Projection to x_1 shows a situation where projection error $\mathcal{E}_{\mathrm{proj}}(x_1)$ is caused by the self-occlusion. It refers to the expected occlusion error case, which does not relate to the ill-synthesized geometry. Using occlusion mask $m(x)$ can exclude the projection case since it should be maintained even after the end of the optimization.
- (b): Projection to x_2 shows an ideal projection case. The synthesized depth $\hat{d}(x_2)$ is identical to the depth $\tilde{d}(x_2)$ obtained by projective-transformation - i.e. $\mathcal{E}_{\mathrm{proj}}(x_2) = 0$.
- (c): Projection to x_3 shows a projection error caused by the false-positive density measure of NeRF. By enforcing $\mathcal{E}_{\mathrm{proj}}(x_3)$ to become zero, the floating artifacts can be expected to be removed.
- (d): Projection to x_4 shows a zero projection error which cannot be optimized by enforcing $\mathcal{E}_{\mathrm{proj}} = 0$. It is because the projection error $\mathcal{E}_{\mathrm{proj}}(x_4)$ is already 0. By assuming that the projection errors caused by self-occlusions would have smooth projection errors while ill-synthesized artifacts are not, we can optimize the depth consistency loss $\mathcal{L}_{\mathrm{dc}}$ described in Eq. 8 in the paper.

Error rate coefficient r_e. As described earlier, our projection error $\mathcal{E}_{\mathrm{proj}}$ includes errors that result from both occlusions and ill-synthesized depth. To prevent over-smoothing of synthesized results on occluded areas, errors resulting from occlusion should be disregarded during training. However, our occlusion-aware weight term is formulated as Eq. 8 in the paper. If the network can effectively

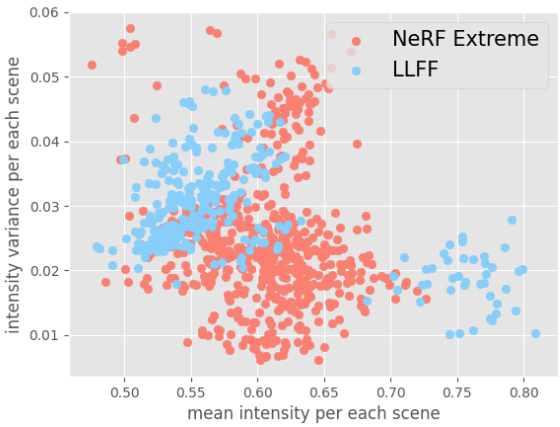


Figure 17: **Intensity distribution of our NeRF Extreme and LLFF [37].** Our dataset shows a larger variance in per-image lighting intensity distribution than LLFF [37]. Lighting intensities are obtained from the shading images extracted by [12].

regularize the ill-synthesized depth, a portion of the occlusion pairs included as a regularization target will increase during training. To prevent the regularization of occlusion pairs, the error rate coefficient r_e will decrease the contribution of the consistency regularization as training progresses.

B.7. Image Formation Model

Basic intrinsic decomposition methods are based on the Lambertian assumption. However, most real-world objects have surface characteristics whose reflectances vary upon viewing directions. Thus, we use the image formation model suggested by [34] as follows:

$$\log I = \log A + \log S + c + N, \tag{17}$$

which takes light color vector c and non-Lambertian residuals N into account.

C. NeRF Extreme

C.1. Pose Estimation

Similar to LLFF [37], all the camera poses are obtained by COLMAP [48] structure from motion framework. All the images and corresponding camera poses will be publicly available.

C.2. Dataset Statistics

In order to verify the illumination diversity of our dataset, we extract per-image intensity distribution from shading images, which are obtained by the state-of-the-art intrinsic decomposition framework [12]. A shading image is suitable for evaluating the illumination diversity of a dataset since it provides environment-dependent information about the image. Fig. 17 shows an intensity distribution of ours and existing LLFF [37], which has similar dataset characteristics. Each dot in the distribution indicates per-image intensity characteristics. Our dataset shows more scattered distribution compared to LLFF, indicating a larger illumination diversity.

C.3. Lighting Variations

All of the images in our dataset were captured using off-the-shelf cameras. Galaxy z-flip 4 was used to reproduce the casual image-capturing setup. Fig. 20 and 21 show all the training and the test images(2 rows from the last) belonging to our *tent* and *bench* scene, respectively. Outdoor scenes are captured at different times and in different sunlight, to be taken under varying illumination conditions. Fig. 18 and 19 show all the training and test images belonging to our *table* and *room* scenes, respectively. Indoor scenes are captured with turned-on/off lights, and closed/open curtains, to get illumination variance.

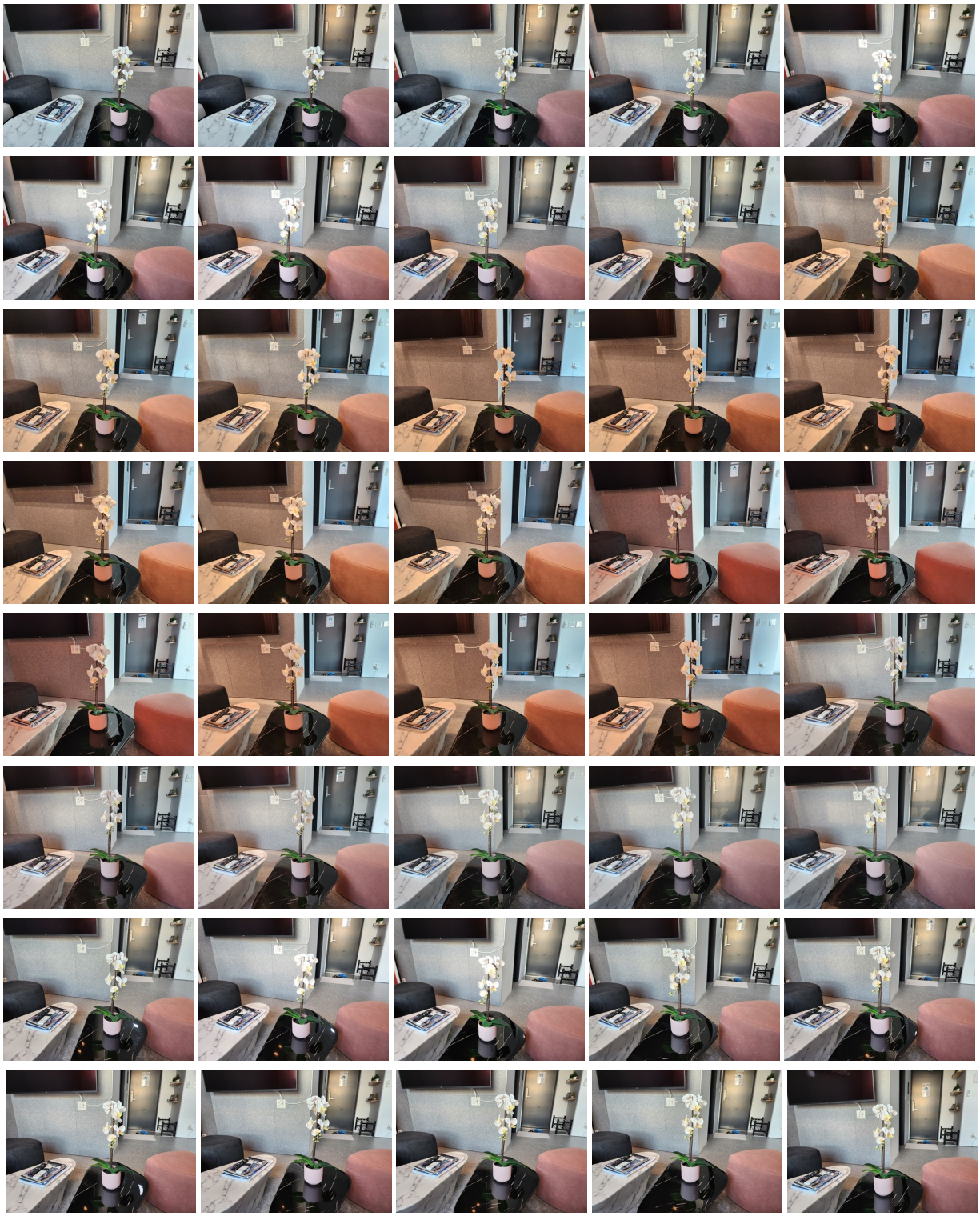


Figure 18: **Illumination variation samples of the *table* scene in NeRF Extreme.**

Figure 19: **Illumination variation samples of the *room* scene in NeRF Extreme.**

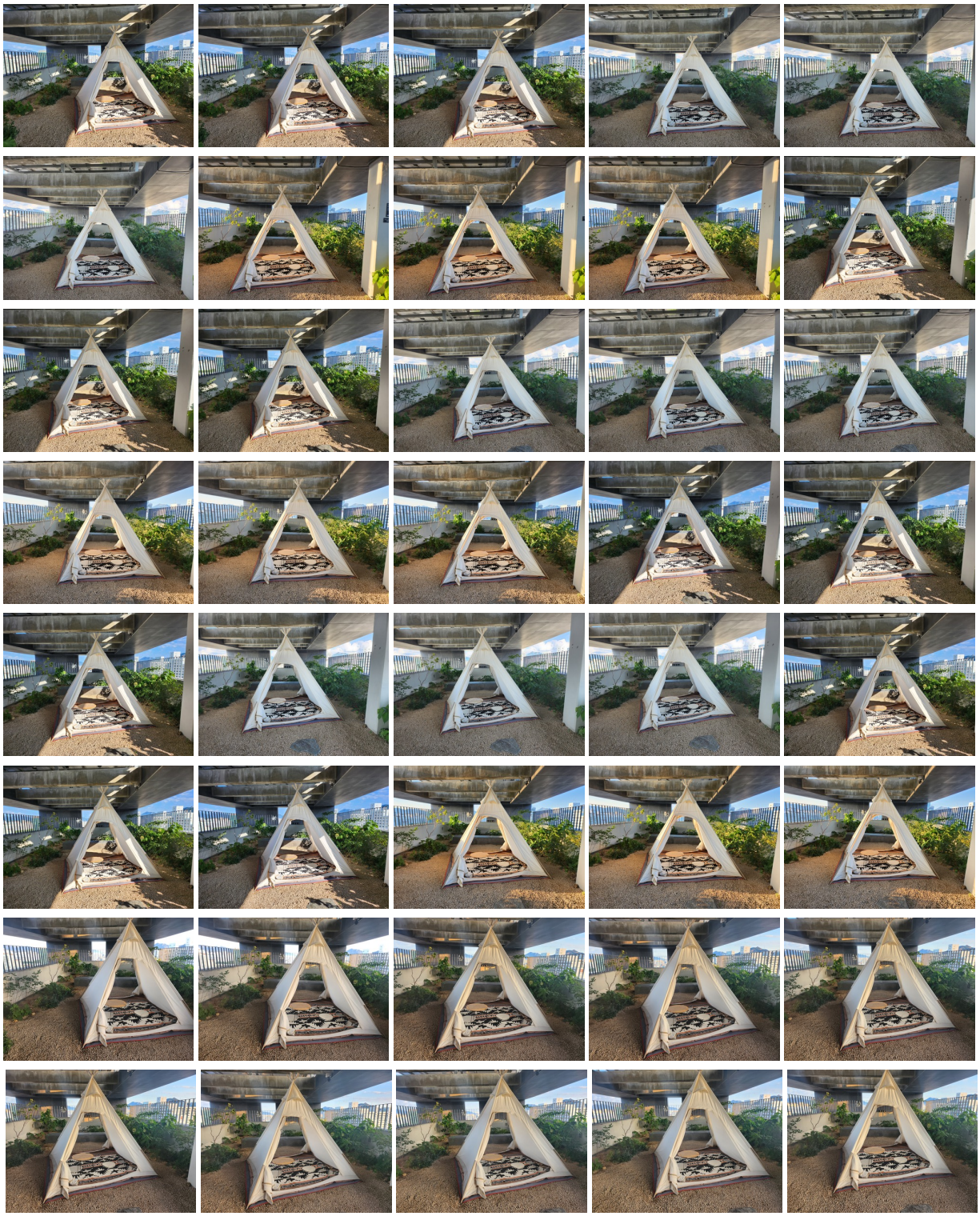


Figure 20: **Illumination variation samples of the *tent* scene in NeRF Extreme.**

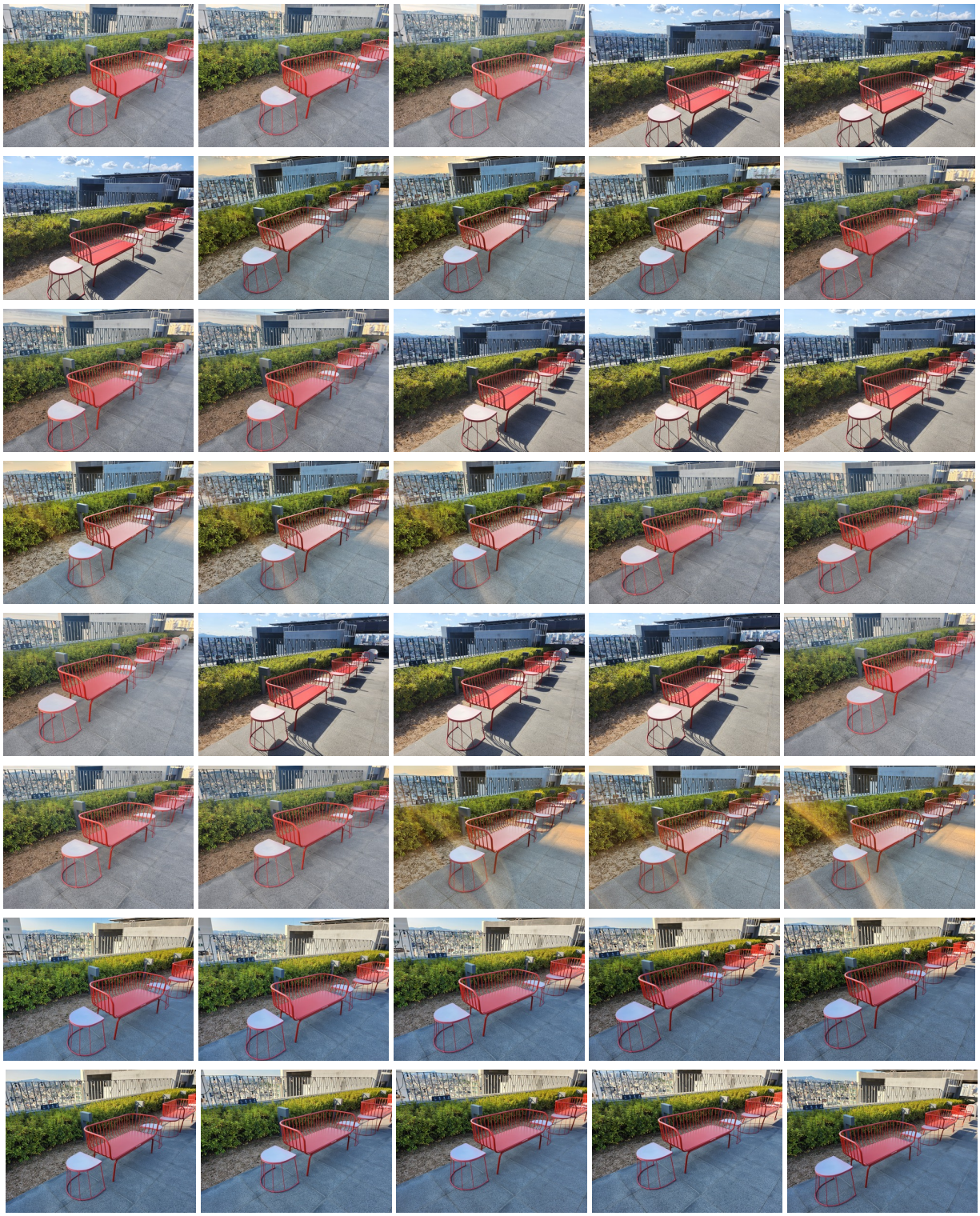


Figure 21: **Illumination variation samples of the *bench* scene in NeRF Extreme.**

References

[1] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, pages 1–16, 2016. 3, 6, 7, 8, 10, 11, 12

[2] Relja Arandjelović and Andrew Zisserman. Nerf in detail: Learning to sample for view synthesis. *arXiv preprint arXiv:2106.05264*, 2021. 1

[3] 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*, 34:26289–26301, 2021. 2

[4] 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*, pages 5855–5864, 2021. 1, 6, 7, 8, 9, 11, 12

[5] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T Barron, Ce Liu, and Hendrik Lensch. Nerd: Neural reflectance decomposition from image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12684–12694, 2021. 2, 3, 5, 6, 9

[6] Mark Boss, Andreas Engelhardt, Abhishek Kar, Yuanzhen Li, Deqing Sun, Jonathan T. Barron, Hendrik P.A. Lensch, and Varun Jampani. SAMURAI: Shape And Material from Unconstrained Real-world Arbitrary Image collections. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2, 3, 5, 6, 9

[7] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. *Advances in Neural Information Processing Systems*, 34:10691–10704, 2021. 2, 5, 6, 9

[8] James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. 6

[9] 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*, pages 14124–14133, 2021. 1, 2, 11

[10] 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*, pages 7911–7920, 2021. 1, 2, 11

[11] François Darmon, Bénédicte Bascle, Jean-Clément Devaux, Pascal Monasse, and Mathieu Aubry. Improving neural implicit surfaces geometry with patch warping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6260–6269, 2022. 2

[12] Partha Das, Sezer Karaoglu, and Theo Gevers. Pie-net: Photometric invariant edge guided network for intrinsic image decomposition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19790–19799, 2022. 3, 6, 11, 15

[13] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. *arXiv preprint arXiv:2212.03267*, 2022. 2

[14] 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*, pages 12882–12891, 2022. 2

[15] Arnab Dey, Yassine Ahmine, and Andrew I Comport. Mip-nerf rgb-d: Depth assisted fast neural radiance fields. *arXiv preprint arXiv:2205.09351*, 2022. 2

[16] Thibaud Ehret, Roger Marí, and Gabriele Facciolo. Nerf, meet differential geometry! *arXiv preprint arXiv:2206.14938*, 2022. 2

[17] Majed El Helou, Ruofan Zhou, Johan Barthas, and Sabine Süsstrunk. VIDIT: Virtual image dataset for illumination transfer. *arXiv preprint arXiv:2005.05460*, 2020. 3

[18] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14346–14355, 2021. 1

[19] Theo Gevers and Arnold WM Smeulders. Color-based object recognition. *Pattern recognition*, 32(3):453–464, 1999. 6, 11

[20] Khang Truong Giang, Soohwan Song, and Sungho Jo. CURVATURE-GUIDED DYNAMIC SCALE NETWORKS FOR MULTI-VIEW STEREO. In *International Conference on Learning Representations*, 2022. 6, 12

[21] Clément Godard, Oisin Mac Aodha, and Gabriel J Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 270–279, 2017. 5, 13

[22] 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*, pages 5875–5884, October 2021. 1

[23] Tao Hu, Shu Liu, Yilun Chen, Tiancheng Shen, and Jiaya Jia. Efficientnerf efficient neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12902–12911, 2022. 1

[24] Xin Huang, Qi Zhang, Ying Feng, Hongdong Li, Xuan Wang, and Qing Wang. Hdr-nerf: High dynamic range neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18398–18408, June 2022. 1

[25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 12

[26] 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*, pages 5885–5894, 2021. 1, 2, 6, 11

[27] Yifan Jiang, Peter Hedman, Ben Mildenhall, Dejia Xu, Jonathan T Barron, Zhangyang Wang, and Tianfan Xue. Alignerf: High-fidelity neural radiance fields via alignment-aware training. *arXiv preprint arXiv:2211.09682*, 2022. 1

[28] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image matching across wide baselines: From paper to practice. *International Journal of Computer Vision*, 129(2):517–547, 2021. 3

[29] Mohammad Mahdi Johari, Yann Lepoittevin, and François Fleuret. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18365–18375, June 2022. 1, 2

[30] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12912–12921, June 2022. 2, 11

[31] Zhengfei Kuang, Kyle Olszewski, Menglei Chai, Zeng Huang, Panos Achlioptas, and Sergey Tulyakov. Neroic: neural rendering of objects from online image collections. *ACM Transactions on Graphics*, 41(4):1–12, 2022. 2, 6, 9, 11

[32] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. Mine: Towards continuous depth mpi with nerf for novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12578–12588, October 2021. 1

[33] Zuoyue Li, Tianxing Fan, Zhenqiang Li, Zhaopeng Cui, Yoichi Sato, Marc Pollefeys, and Martin R Oswald. Compnvs: Novel view synthesis with scene completion. *arXiv preprint arXiv:2207.11467*, 2022. 2, 5

[34] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9039–9048, 2018. 3, 5, 6, 10, 12, 13, 15

[35] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5741–5751, 2021. 1

[36] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 2, 3, 5, 6, 9

[37] 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*, 38(4):1–14, 2019. 3, 5, 6, 8, 11, 15

[38] 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 *European Conference on Computer Vision*, pages 405–421. Springer, 2020. 1, 3, 11

[39] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics*, 41(4):102:1–102:15, July 2022. 1

[40] Lukas Murmann, Michael Gharbi, Miika Aittala, and Fredo Durand. A multi-illumination dataset of indoor object appearance. In *2019 IEEE International Conference on Computer Vision*, Oct 2019. 3

[41] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 1, 2, 3, 6, 7, 8, 9, 10, 11, 12

[42] 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*, pages 5865–5874, 2021. 1

[43] 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*, pages 14335–14345, 2021. 1

[44] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10901–10911, 2021. 1

[45] Konstantinos Rematas, Ricardo Martin-Brualla, and Vittorio Ferrari. Sharf: Shape-conditioned radiance fields from a single view. *arXiv preprint arXiv:2102.08860*, 2021. 2

[46] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12892–12901, 2022. 2

[47] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 12

[48] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, 2016. 5, 15

[49] Jianxiong Shen, Antonio Agudo, Francesc Moreno-Noguer, and Adria Ruiz. Conditional-flow nerf: Accurate 3d modelling with reliable uncertainty quantification. *arXiv preprint arXiv:2203.10192*, 2022. 1

[50] Zifan Shi, Yujun Shen, Jiapeng Zhu, Dit-Yan Yeung, and Qifeng Chen. 3d-aware indoor scene synthesis with depth priors. *arXiv preprint arXiv:2202.08553*, 2022. 2

[51] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 2

[52] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P. Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8248–8258, June 2022. 1, 2

[53] Jonathan Tremblay, Moustafa Meshry, Alex Evans, Jan Kautz, Alexander Keller, Sameh Khamis, Charles Loop, Nathan Morrical, Koki Nagano, Towaki Takikawa, et al. Rtmv: A ray-traced multi-view synthetic dataset for novel view synthesis. *arXiv preprint arXiv:2205.07058*, 2022. 3

[54] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 1, 2

[55] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. Sparf: Neural radiance fields from sparse and noisy poses. *arXiv preprint arXiv:2211.11738*, 2022. 2

[56] Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. Clip-nerf: Text-and-image driven manipulation of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3835–3844, June 2022. 2, 11

[57] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 1, 2

[58] Zirui Wang, Shangzhe Wu, Weidi Xie, Min Chen, and Victor Adrian Prisacariu. Nerf–: Neural radiance fields without known camera parameters. *arXiv preprint arXiv:2102.07064*, 2021. 1

[59] Daniel Watson, William Chan, Ricardo Martin-Brualla, Jonathan Ho, Andrea Tagliasacchi, and Mohammad Norouzi. Novel view synthesis with diffusion models. *arXiv preprint arXiv:2210.04628*, 2022. 2

[60] Yair Weiss. Deriving intrinsic images from image sequences. In *Proceedings Eighth IEEE International Conference on Computer Vision*, volume 2, pages 68–75. IEEE, 2001. 3

[61] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image. *arXiv preprint arXiv:2204.00928*, 2022. 1, 2, 4, 11

[62] Weicai Ye, Shuo Chen, Chong Bao, Hujun Bao, Marc Pollefeys, Zhaopeng Cui, and Guofeng Zhang. Intrinsicnerf: Learning intrinsic neural radiance fields for editable novel view synthesis. *arXiv preprint arXiv:2210.00647*, 2022. 2, 3, 5, 13

[63] 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*, pages 4578–4587, 2021. 1, 2, 6, 11

[64] Yu-Jie Yuan, Yu-Kun Lai, Yi-Hua Huang, Leif Kobbelt, and Lin Gao. Neural radiance fields from sparse rgb-d images for high-quality view synthesis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2022. 2

[65] Jason Zhang, Gengshan Yang, Shubham Tulsiani, and Deva Ramanan. Ners: neural reflectance surfaces for sparse-view 3d reconstruction in the wild. *Advances in Neural Information Processing Systems*, 34:29835–29847, 2021. 1

[66] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 1

[67] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics*, 40(6):1–18, 2021. 2

[68] Xuanmeng Zhang, Zhedong Zheng, Daiheng Gao, Bang Zhang, Pan Pan, and Yi Yang. Multi-view consistent generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18450–18459, 2022. 2, 4