ReLight My NeRF: A Dataset for Novel View Synthesis and Relighting of Real World Objects

ReLight My NeRF: A Dataset for Novel View Synthesis and Relighting of Real World Objects

Marco Toschi*, Riccardo De Matteo*,◇, Riccardo Spezialetti*, Daniele De Gregorio
Eyecan.ai
{marco.toschi, riccardo.spezialetti, daniele.degregorio}@eyecan.ai
Luigi Di Stefano, Samuele Salti
University of Bologna

Figure 1. **A random subset of scenes from our dataset**. Our dataset enables the study of joint novel view synthesis and relighting from real images, as it provides scenes framed from the same viewpoint under varying light positions (shown along rows) as well as captured from different viewpoints under the same light position (shown along columns).

Abstract

*In this paper, we focus on the problem of rendering novel views from a Neural Radiance Field (NeRF) under unobserved light conditions. To this end, we introduce a novel dataset, dubbed ReNe (**Re**lighting **NeRF**), framing real world objects under one-light-at-time (OLAT) conditions, annotated with accurate ground-truth camera and light poses. Our acquisition pipeline leverages two robotic arms holding, respectively, a camera and an omni-directional point-wise light source. We release a total of 20 scenes depicting a variety of objects with complex geometry and challenging materials. Each scene includes* 2000 *images, acquired from 50 different points of views under 40 different OLAT conditions. By leveraging the dataset, we perform an ablation study on the relighting capability of variants of the vanilla NeRF architecture and identify a lightweight architecture that can render novel views of an object under novel light conditions, which we use to establish a non-trivial baseline for the dataset. Dataset and benchmark are available at* https://eyecan-ai.github.io/rene.

1. Introduction

Inverse rendering [29, 47, 52, 73] addresses the problem of estimating the physical attributes of an object, such as its geometry, material properties and lighting conditions, from a set of images or even just a single one. This task is a longstanding problem for the vision and graphics communities, since it unlocks the creation of novel renderings of an object from arbitrary viewpoints and under unobserved lighting conditions. An effective and robust solution to this problem would have significant value for a wide range of applications in gaming, robotics and augmented reality.

Recently, Neural Radiance Fields (NeRF) [36] has contributed tremendously to the novel view synthesis sub-task of inverse rendering pipelines. By mapping an input 5D vector (3D position and 2D viewing direction) to a 4D continuous field of volume density and color by means of a neural network, NeRF learns the geometry and appearance of a single scene from a set of posed images. The appealing results in novel view synthesis have attracted a lot of attention from the research community and triggered many follow-up

* Joint first authorship. ◇ Work done while at Eyecan.ai

Dataset	Multiple categories	Real-World	Background Shadows	Public	Light Supervision
Gross *et al*. [14]	✗	✓	✗	✓	✓
Sun *et al*. [57]	✗	✓	✗	✓	✓
Wang *et al*. [68]	✗	✓	✗	✓	✓
Zhang *et al*. [74]	✗	✓	✗	✓	✓
Srinivasan *et al*. [55]	✓	✗	✗	✓	✓
Zhang *et al*. [75]	✓	✓	✓	✓	✗
Zhang *et al*. [75]	✓	✗	✗	✓	✓
Bi *et al*. [2]	✓	✓	✗	✗	✓
ReNe	✓	✓	✓	✓	✓

Table 1. Overview of relighting datasets. Our dataset is the first one featuring a variety of objects and materials captured with real-world sensors that provides ground-truth light positions and also presents challenging cast shadows.

works aimed at overcoming the main limitations of NeRF, *e.g*. reduce inference runtime [23, 24, 27, 48, 49, 71], enable modeling of deformable objects [9, 25, 42, 45, 46, 62], and generalization to novel scenes [5,15,19,39,50,53,61,63,72]. However, less attention has been paid to the relighting ability of NeRFs. Although NeRF and its variants represent nowadays the most compelling strategy for view synthesis, the learned scene representation entangles material and lighting and, thus, cannot be directly used to generate views under novel, unseen lighting conditions. A few existing works [3, 4, 55, 75] try to overcome this structural NeRF limitation by learning to model the scene appearance as a function of *reflectance*, which accounts for both scene geometry and lighting modeling. However, these methods incur a great computation cost, mainly due to the need to explicitly model light visibility and/or the components of a microfacet Bidirectional Reflectance Distribution Function (BRDF) [65] like diffuse albedo, specular roughness, and point normals: for instance NeRV [55] uses 128 TPU cores for 1 day while [3] is trained on 4 GPUs for 2 days.

As highlighted in Tab. 1, one of the main challenges to foster research in this direction is the absence of real-world datasets featuring generic and varied objects with ground-truth light direction, both at training and test time. The latter is key to create a realistic quantitative benchmark for relighting methods, whose availability is one of the main driving forces behind fast-paced development of a machine learning topic. Indeed, the above mentioned NeRF-like methods for relighting mainly consider a handful of synthetic images to provide quantitative results (3 and 4 scenes in NeRV [55] and NeRFactor [75], respectively) while no quantitative results are provided in [3]. The only dataset with scenes acquired by a real sensor is proposed in [2], which, however, assumes images captured under collocated view and lighting setup, *i.e*. a smartphone with flash light on, which limits the amount of cast shadows and simplifies the task. Moreover, the dataset is not publicly available. Some available real-world datasets with ground-truth light positions feature human faces or portraits [14, 57, 68, 74]. In these datasets, the subject is usually seated in the center of a light-stage with cameras arranged over a dome array positioned in front of the subject. Although these dataset provide real scenes with ground-truth annotations for light positions, the background is masked-out and shadows cast on the background, which are hard to model because they require a precise knowledge about the geometry of the overall scene, are ignored.

Therefore, in this paper we try to answer the research questions: can we design a data acquisition methodology suitable to collect a set of images of an object under *one-light-at-time* (OLAT) [74] illumination with high-quality camera and light pose annotations which requires minimal human supervision? We then leverage it to investigate a second question: can we design a novel Neural Radiance Field architecture to learn to perform relighting with reasonable computational requirements? To answer the first question, we design a capture system relying on two robotic arms. While one arm holds the camera and shots pictures from viewpoints uniformly distributed on a spherical wedge, the other moves the light source across points uniformly distributed on a dome. As a result, we collect the ReNe dataset, made of 20 scenes framing daily objects with challenging geometry, varied materials and visible cast shadows, composed of 50 camera view-points under 40 OLAT light conditions, i.e. 2000 frames per scene. Examples of images from the dataset are shown in Fig. 1. With a subset of images from each scene, we create a novel hold-out dataset for joint relighting and novel view synthesis evaluation that will be used as an online benchmark to foster research on this important topic. As regards the second question, thanks to the new dataset we conduct a study on the relighting capability of NeRF. In particular, we investigate on how the standard NeRF architecture can be modified to take into account the position of the light when generating the appearance of a scene. Our study shows that by estimating color with two separate sub-networks, one in charge of soft-shadow prediction and one responsible for neurally approximating the BRDF, we can perform an effective relighting, *e.g*. cast complex shadows. We provide results of our novel architecture as a reference baseline for the new benchmark.

In summary, our contributions include:

- a novel dataset made out of sets of OLAT images of real-world objects, with accurate camera and light pose annotations;
- a study comparing different approaches to enable NeRF to perform relighting alongside novel view synthesis;
- a new architecture, where the stage responsible for radiance estimation is split into two separate networks,

that can render novel views under novel unobserved lighting conditions;

- a public benchmark for novel view synthesis and relighting of real world objects, that will be maintained on an online evaluation server.

2. Related Work

In this paper, we focus on relighting static objects by Neural Radiance Fields. We briefly discuss the related works and their datasets below.

Image Relighting without NeRF. There exist several works in the field of image based relighting [8, 32, 44, 51]. The proposed methods differ for the adopted technique to model the light transport function in form of discrete light transport matrix. With recent advances in deep learning, new techniques have been introduced to address relighting [57], with many of them designed to relight human portraits [38, 58, 59, 70, 76]. Another bunch of works pursues joint relighting and novel view synthesis [11, 34, 74].

Relighting Datasets. Most of relighting datasets contain captures of human bodies or face portraits, they leverage complex capture setups such as calibrated multi-view lightstages with cameras and LED lights, where the cameras are synchronized with the lights in order to flash one LED per capture [38,58,59,76]. Alternative works on generic objects do exist [51, 70], but they consider a mix of real and synthetic data, with the latter dominating the former in terms of number of images. The dataset most similar to ours is the one adopted by [2], which involves a robotic arm setup holding a Samsung Galaxy Note 8. However, this acquisition framework assumes a collocated view and lighting setup, *e.g*. built-in flash of a smart-phone camera lens, that inherently limits the shadows cast. Furthermore, the dataset has not been released.

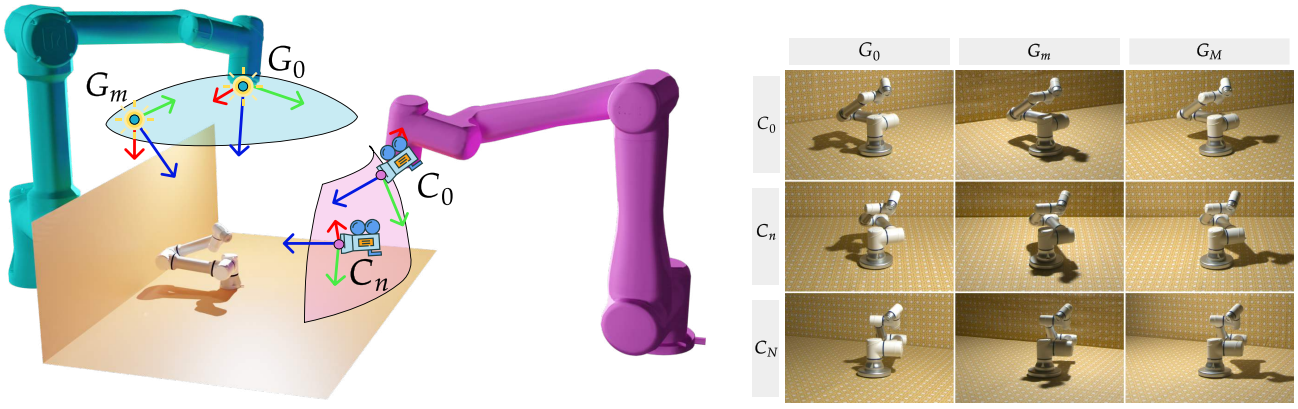
Neural Radiance Fields. Novel view synthesis has been a longstanding problem within the computer vision and computer graphics fields [7, 13, 20]. The advent of deep learning gave rise to explicit methods that train CNNs for this very purpose [22, 26, 35, 54, 56, 64, 77]. In the past few years, NeRF has advanced greatly in becoming the main scene representation for view synthesis [36]. NeRF learns a continuous volumetric function parameterized by means of a fully connected neural network optimized over a set of observed images with known camera poses using gradient descent. Due to its effectiveness, NeRF has inspired many subsequent works that extend its continuous neural volumetric representation for generative modeling [5, 19, 53], dynamic scenes [10, 21, 21, 30, 41, 46, 69], non-rigidly deforming objects [9, 9, 40, 42, 43, 62], multi-resolution images [1, 60], phototourism images with changing illumination settings [30, 61] and relighting [2, 4, 30, 55].

Neural Radiance Fields for Relighting. One of main limitations of NeRF is that is not suitable for relighting. Indeed, NeRF treats the particles within scene representation as elements that *emit* light instead of being modelled as particles that *reflect* the incoming light sent out from external light sources. To straddle this divide, Neural Reflectance Field [2] models both scene geometry and *reflectance* regressing volume density, normal and material properties, *e.g*. Bidirectional Reflectance Distribution Function (BRDF) [65], for any 3D location within the volume of a scene. The shading at each point is then estimated using the light, view direction, the normal at that point and the BRDF. For instance, [28] leverages a differentiable path tracer to optimize a spatially-varying BSDF tied to an the environment map, that can be used to render novel views under a novel OLAT condition. Unfortunately, this approach requires marching rays from all points sampled along the camera ray to any source of lighting in the scene, which restricts it use only in a collocated camera-light setup. NeRV ameliorates the computational cost of [2] by replacing the visibility estimation between scene points and light sources with a neural approximation of the true visibility field, which acts as a lookup table during rendering. Thanks to this insight, NeRV simulates direct illumination from environment lighting as well as one-bounce indirect illumination. NeRD [4] does not model visibility or shadows, but uses an analytic BRDF model [6] to learn a volumetric representation which stores SVBRDF [16] parameters at each 3D point instead of a radiance field. A relightable textured mesh is then extracted from that volume to allow fast rendering and relighting. Finally, NeRFactor [75] starts from two pre-trained networks: a NeRF of the scene and a BRDF network trained on the MERL dataset [31]. This knowledge is then distilled into four MLP-based networks to predict for each surface location its normal vector, light visibility, albedo and a BRDF latent code. The MLP outputs are injected into classical volumetric rendering and the networks parameters are optimized minimizing the re-rendering loss.

We take inspiration from this line of works to design the study reported in Sec. 5, where we split the MLP-stack of NeRF in two separate networks: one responsible for the visibility and one that resembles the BRDF.

3. Dataset Acquisition Framework

While providing both lights and cameras poses in a controlled manner with ground-truths annotations is straightforward in synthetic environments [55, 75], it is cumbersome and complex in the real world, especially if humans are involved, as done in [14, 57, 68, 74]. Our solution is to deploy extremely repeatable and safe machines such as industrial *cobots* (*collaborative robots*). In particular, we use a pair of robots, dubbed *LightBot* and *CameraBot*, to position the light and camera independently. Thanks to their high repeatability (*+/- 0.03 mm*), we can calibrate their tra-



(a) (b)

Figure 2. **Overview of our dataset acquisition framework.** (a) The *LightBot* and *CameraBot* moving light and camera respectively; the trajectories of the two robots are two non-intersecting sections of a hemisphere around the object of interest. (b) Grid of images in which each row depicts the same viewpoint as the light conditions change; each column, on the other hand, represents the same light condition as the viewpoint changes.

jectories with respect to a common world reference frame only once before acquisition starts, and then obtain accurate 6-DoF pose knowledge of both light and camera for all scenes by repeating the same trajectories. The proposed framework is sketched in Fig. 2a, while some acquisitions for the Robotoy scene are shown in Fig. 2b.

Camera/Light Calibration. In order to calibrate both robots with respect to a common reference frame we use a ChArUco board [12]. We let the *CameraBot* see the board at each waypoint of its trajectory in order to calibrate extrinsics parameters for each view. Thanks to the high repeatability of the robot, it is possible to perform this procedure only once before we start to record scenes: the robot will be able to place the camera at the previously calibrated positions with negligible error. This allows us to register in a common reference frame all the camera poses for each scene without instrumenting the scene with a pattern, which alters the realism of the scene and makes the setup fragile, as the pattern may be inadvertently moved across scans. As for the *LightBot*, we assume that the center of the point light corresponds to the geometric center of the LED. We are then able to register it with the calibration pattern by letting the end effector physically touch the central reference point of the pattern with the center of the LED at calibration time. We note we cannot use the same procedure for the *CameraBot* as the camera optical center is behind the lens. That closes the calibration loop since both robots, now, will have the pattern itself as their world reference frame, thus a common coordinate system. Even with the *LightBot*, high repeatability assures that the light pose registered at calibration time will be identical across all scenes of the dataset.

Trajectories. As depicted in Fig. 2a, we generate two trajectories, roughly belonging to the same hemisphere, whose upper part is used for the lights while the side part for the camera. This made it possible to capture several front-facing scenes with the light moving over the object of interest. With this setup, then, for each scene we collected 50 viewpoints each under 40 different light locations. To have waypoints uniformly distributed on the selected spherical region for each end effector, we create trajectories as sequences of centers of equal-area subregions, following the methodology proposed by [17]. For the sake of illustration, Fig. 2b shows a grid of images in which each row depicts frames with the same camera pose but different lighting conditions and vice-versa for columns.

Background texture. In preliminary experiments with a uniform background as scenario, we found out that NeRF struggles to correctly estimate density in such a case and training does not converge. Hence, in our dataset textured walls are shown in the background.

Hardware. We equipped the *LightBot* and the *CameraBot*, respectively an Universal Robots UR5e and an Elite Robotics EC66, with a consumer headlight and an industrial camera. The headlight is the Velamp Metros IH523, equipped with 5 COB LEDs that provide a diffuse 100° beam up to 150 Lm with 6000K light temperature. We used it at half power (70 Lm) to maximize its runtime. The camera is a Basler acA1440-73gc with its optical axis mounted at 45° from the robot flange normal vector, this to simplify the computation of inverse kinematics of the *CameraBot*. Images are captured in 1.6MP resolution (1440px × 1080px) using a Basler lens with 8mm focal length.

4. The ReNe Dataset

In this section, we summarize the data we collected using our framework to establish the ReNe dataset. Our dataset $\mathcal{S} = \{\mathcal{I}_s\}_{s=1}^{S}$ contains S scenes, where each scene is a collection of RGB images $\mathcal{I}_s = \{I_{s,n,m} \mid n = 1, \dots, N, m = 1, \dots, M\}$, $I_{n,m} \in [0,1]^{H \times W \times 3}$ acquired using camera poses $\mathcal{C}_s = \{C_{s,n}\}_{n=1}^{N}$ and lit with a point light source placed in $\mathcal{G}_s = \{G_{s,m}\}_{m=1}^{M}$. Each C_n and G_m represent the 6D pose $\{\mathbf{R}, \mathbf{t}\} \in SE(3)$ for camera and light source, respectively. $\mathbf{R}$ denotes rotation, $\mathbf{R} \in SO(3)$, and $\mathbf{t}$ denotes translation, $\mathbf{t} \in \mathbb{R}^3$. The dataset features 40000 images divided in $S = 20$ scenes taken by $N = 50$ frontal point of views in $M = 40$ different lighting conditions, *i.e.* 2000 frames per scene.

Train, Validation and Test splits. We split the datasets into train, val and test subsets to make it easy to compare in a fair and consistent way different approaches to NeRF relighting. In order to avoid data leakage from validation and test, we randomly create the set of held out viewpoints for validation, $\mathcal{C}^{val}$, and for testing, $\mathcal{C}^{test}$, as well as the set of held out light positions for testing, $\mathcal{G}^{test}$. Each of them consists of 3 indexes. To build the validation set, for each viewpoint $n_{val} \in \mathcal{C}^{val}$, we randomly pick 3 light poses among the 37 available to form $\mathcal{G}_{n_{val}}$. In this way, in the validation set we have an unseen pair (viewpoint, light) composed of a viewpoint not present in the training set and a light used for other training viewpoints, $\mathcal{I}^{val} = \{I_{s,n_{val},m_{val}} | n_{val} \in \mathcal{C}_{val} \wedge m_{val} \in \mathcal{G}_{n_{val}} \wedge s = 1, \dots, S\}$. For test set, we build two sub-splits namely *easy* and *hard*. The former considers all the images $\mathcal{I}^{easy} = \{I_{s,n,m} | n \in \mathcal{C}^{test}, m = 1, \dots, M, m \notin \mathcal{G}^{test}\}$, while the latter is composed of images $\mathcal{I}^{hard} = \{I_{s,n,m} | (n,m) \in \mathcal{C}^{test} \times \mathcal{G}^{test}\}$. With this partition, the test set is comprised of novel viewpoints lit by light seen at training time in the easy split, and of novel viewpoints under a never seen light in the hard split. We sketch the splits in Fig. 3.

5. Benchmarking the Relighting Capability of NeRF

In the section, we explore our second research question: equipped with the ReNe dataset, can we add relighting capabilities to NeRF in a simple and lightweight way? We start with a brief overview of NeRF (Sec. 5.1), we then discuss alternative ways to make NeRF able to perform novel view synthesis and relighting simultaneously (Sec. 5.2). We compare our choices through an ablation study using the validation set described in Sec. 4, whose results are discussed in Sec. 5.3. Finally, in Sec. 5.4, we present the results for the best method on both test sets of the ReNe dataset.

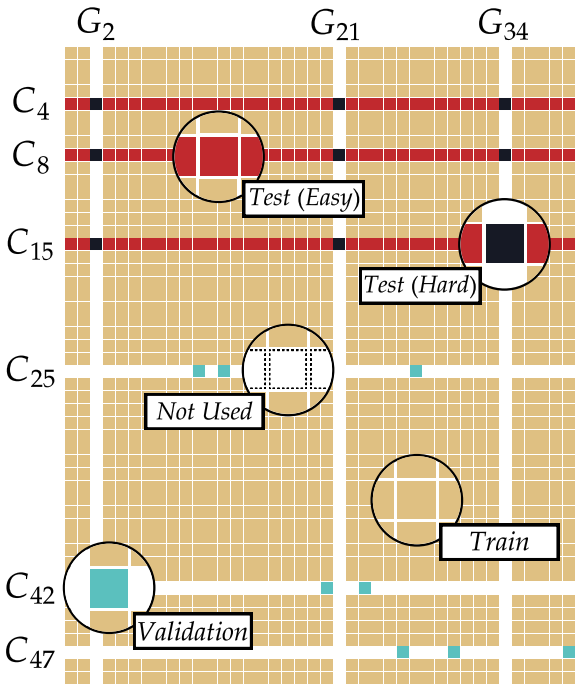


Figure 3. **Dataset splits.** Each scene $\mathcal{I}_s$ is divided in 4 different splits: ■ Samples used for training ■ Samples used for validation ■ Samples used for Easy Test ■ Samples used for Hard Test □ Samples never used.

5.1. NeRF Overview

A Neural Radiance Field (NeRF) [36] maps a 5D input representing camera pose, *i.e.* 3D coordinates $\mathbf{x} = (x, y, z)$ along with the 2D viewing directions $\mathbf{d} = (\theta, \phi)$, into a 4D color-density output $(\mathbf{c}, \sigma)$ by means of a function $F_\Theta(\mathbf{x}, \mathbf{d}) \rightarrow (\mathbf{c}, \sigma)$ approximated by the weights of an MLP. Specifically, here we consider a vanilla NeRF architecture [37] which estimates color and density using two MLPs as $(\sigma, \mathbf{e}) = \Psi_{geo}(\mathbf{x})$ and $\mathbf{c} = \Psi_{rgb}(\mathbf{e}, \mathbf{d})$, with σ being interpreted as the probability of a ray terminating at (x, y, z) and $\mathbf{e}$ being a feature embedding. Following [33], the color $C(\mathbf{r})$ rendered from a camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ can be obtained solving the integral:

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))c(\mathbf{r}(t), \mathbf{d})dt \qquad (1)$$

where $T(t) = exp\left(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds\right)$ represents the accumulated transmittance from t_n to t along the ray r, and t goes from near plane t_n to far plane t_f. The image formation procedure is performed by simply aggregating the result of the integral for all pixels, *i.e.* samples along the rays, of the target image. Although NeRF works well for synthesizing novel views, as shown by the previous equations, it has no way of modeling the incoming light since

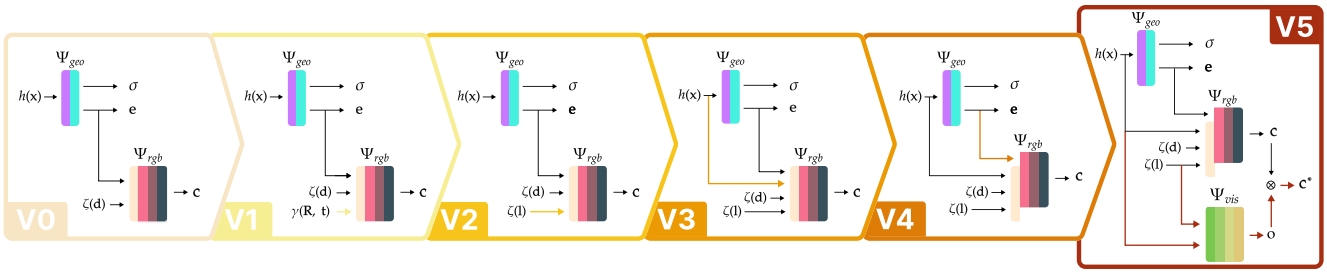


Figure 4. **Overview of architectures.** We start from the architecture of [37] and propose variants to inject information about the light position in the scene. Highlighted arrows in each variant underline an architectural change from the one that precedes.

it only takes into account position and orientation for the camera. In the next section, we will introduce variations of the classical approach that also take into account the light position.

Positional Encoding. As showed in NeRF [35] enconding $\mathbf{x}$ and $\mathbf{d}$ into a higher dimensional space allows the neural network to overcome its spectral bias and makes it able to reproduce the high-frequency content of the input signal. We follow [37] and encode $\mathbf{x}$ using a multiresolution hash encoding, $h(\mathbf{x})$, whilst $\mathbf{d}$ is projected onto the first 16 coefficients of the spherical harmonics, $\zeta(\mathbf{d})$.

5.2. On The Relighting Capability of NeRF

As illustrated in Fig. 4, we take as a starting point the architecture presented in [37] and delineate 5 proposals to enable NeRF to perform relighting while adding minimum complexity to the architecture. Recalling Sec. 5.1, the naive architecture, namely V0, has two sub-networks, Ψ_{geo}, which outputs σ a measure of the particle density at $\mathbf{x}$, and Ψ_{rgb} which is responsible for the color, $\mathbf{c}$. Previous works [66, 66] demonstrated that Ψ_{geo} learns the scene geometry, while Ψ_{rgb} models the view-dependent color appearance. Hence, our key idea is to modify the input to the Ψ_{rgb} injecting light position and let the model learn the interaction between the geometry and light. We train one network on each scene minimizing the error between the ground-truth images and the rendered ones. As input data we consider a set of multi-view images of an object, $\mathcal{I}_s$, illuminated under known lighting conditions, $\mathcal{G}_s$, and the camera poses of these images, $\mathcal{C}_s$. We will consider the subscript s implicit in what follows.

V1. A straightforward way to enable relighting with NeRF is to condition Ψ_{rgb} with a latent representation that stores the world position and orientation of the light, G_m for a given observation $I_{n,m}$. To do so, we flatten the $\mathbf{R}$ and $\mathbf{t}$ parts of G_m and encode this 12-dimensional vector into a 156-dimensional latent code using the positional encoding with Fourier Features proposed in [36], $\gamma(\mathbf{R}, \mathbf{t})$. Our procedure resembles the approach adopted by NeRF in the Wild [30], but the condition embedding is used to adapt NeRF to variable lighting conditions instead of transient parts of the images.

V2. As the next iteration, we change the parameterization for G_m. Instead of directly encoding $\mathbf{R}$ and $\mathbf{t}$ in the same way for each sample $\mathbf{x}$ used to query Ψ_{rgb}, we construct a 3D unit vector $\mathbf{l} = \frac{\mathbf{t}-\mathbf{x}}{||\mathbf{t}-\mathbf{x}||}$ which let the network be aware of the relative position between the light source and the query location. Since, differently from the previous approach, $\mathbf{l}$ changes for each query point, we conjecture this may help NeRF in modulating the output color $\mathbf{c}$ based on the incoming light information. As done for $\mathbf{d}$, $\mathbf{l}$ is encoded using the spherical harmonics, $\zeta(\mathbf{l})$.

V3. The previous version resembles somehow the BRDF where $\mathbf{l}$ acts as the incoming light direction at $\mathbf{x}$, and $\mathbf{d}$ is the unit vector pointing from $\mathbf{x}$ toward the camera. Following this analogy, we may consider $\mathbf{e}$ to represent an N-dimensional embedding of the normal vector at $\mathbf{x}$. Since all these information are relative to $\mathbf{x}$, we explore whether having $h(\mathbf{x})$ as input can be beneficial for Ψ_{rgb}. Without such skip connection the knowledge about the location of the query point may vanishes as the depth of the network increases.

V4. The previous version may be negatively influenced by the diverse scales of its inputs, which are on one side spherical and hash-grid encoded vectors like $\zeta(\mathbf{d})$, $\zeta(\mathbf{l})$, and $h(\mathbf{x})$, while on the other there is the ReLU output $\mathbf{e}$. To test this hypothesis, we propose a variant where we concatenate $\mathbf{e}$ with the output of the first layer of Ψ_{rgb} and use this vector as input for its second layer, rather than feeding Ψ_{rgb} directly with $\mathbf{e}$.

V5. The original NeRF learns a continuous 3D field of particles that absorb and *emit* light. As a result only the amount of outgoing light from a location is modeled, without taking into account the fact that the outgoing light is the result of interactions between incoming light and the material properties of an underlying surface. This issue can be alleviated by replacing Eq. (1) with a physically-based volume rendering for non-emissive and non-absorptive volumes, which replaces the emitted color of each 3D point along the ray, $c(\mathbf{r}(t), \mathbf{d})$, with $L_r(\mathbf{x}, \omega_o)$ that represents the scattered light at $\mathbf{x}$ along ω_o. As done in [2], assuming a single point light

source of fixed intensity, L_r can be approximated as:

$$L_r(\mathbf{x}, \omega_o) = \int_S f_p(\mathbf{x}, \omega_o, \omega_i) L_i(\mathbf{x}, \omega_i) d\omega_i \qquad (2)$$

where S is a unit sphere, f_p represents a differentiable reflectance model (like the BRDF), and L_i represents the incident radiance at x from direction ω_i. Please note that in this equation ω_o correspond to $\mathbf{d}$ and ω_i is analogous to $\mathbf{l}$. While we can assume that f_p is approximated by Ψ_{rgb}, computing L_i requires marching a large number of light rays for all shading points on all camera rays to determine the transmittance between the light and each 3D point. We avoid this cumbersome procedure by introducing a neural approximation of L_i in the form of an MLP aimed at predicting a scalar value $o = \Psi_{vis}(h(\mathbf{x}), \zeta(\mathbf{l}))$, which allows us to efficiently query the point-to-light visibility. The final pixel color is then obtained with $\mathbf{c}^\star = o \cdot \mathbf{c}$.

5.3. Ablation study

Dataset. To assess the relative merits of the incremental modifications to NeRF proposed in the previous section and identify the best architecture, we consider a subset of scenes presenting a variety of challenges: Cube, that has thin wireframe structures and shadows; Savannah, that pictures a jagged leaves tree with a complex shadow; Reflective, that is full of specular reflections, and FlipFlop, where light passes through semi-transparent plastic strings.

Implementation Details. We used the torch-ngp pytorch implementation[1] of Nvidia Instant-NGP [37] for fast training of Neural Radiance Fields models. During training, we randomly sample 4096 pixel rays as a batch to train our networks. We use Adam optimizer [18] with an initial learning rate of 0.01 (other Adam hyperparameters are left at default values of $\beta_1 = 0.9$, $\beta_2 = 0.999$). For volumetric integration, we use 128 samples in coarse volume and 128 additional adaptive samples in fine volume to compute the final radiance. The optimization for a single scene typically takes around 100–200k iterations to converge on a single NVIDIA RTX 2080Ti GPU (about 5 hours). We use early-stopping if the model doesn't improve for 10 consecutive epochs. We supervise the regressed RGB values with the ground truth values from the captured images using the L2 loss. For Ψ_{geo} we use 2 fully-connected ReLU layers with 64 channels, while Ψ_{rgb} and Ψ_{vis} uses 4 fully-connected ReLU layers with 64 channels.

Results. For a quantitative evaluation of our methods we consider two widely adopted metrics of image quality assessment: Structural Similarity (SSIM) [67] and Peak Signal-to-Noise Ratio (PSNR) [35]. We report the results of the ablation study in Tab. 2. Results show how all the proposed modifications to NeRF improve performance and contribute to the overall good results of the most effective version, which is V5. These results provide experimental support to the intuitions presented above about the higher effectiveness of feeding to the network the relative light position with respect to the sample $\mathbf{x}$ instead of the absolute light pose (row 1 versus row 2), as well as the importance of providing $\mathbf{x}$ as input to Ψ_{rgb} (row 2 versus row 3). Separating inputs to Ψ_{rgb} is also beneficial and, together with Ψ_{vis} contributes to the good performance of V5.

Some qualitative examples comparing and contrasting V1 against V5 are reported in Fig. 5. Zooming in on cast shadows we can appreciate how careful handling of the light pose as well as the proposed modifications to the vanilla NeRF architecture enable rendering of sharper and more coherent shadows than the straightforward extension that conditions NeRF also on light pose.

[1] https://github.com/ashawkey/torch-ngp

Method	$\gamma(\mathbf{R},\mathbf{t})$	$\mathbf{l}$	*skip*	*inputs*	Ψ_{vis}	Cube		Savannah		Reflective		FlipFlop	
						PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
V1	✓	✗	✗	✗	✗	24.37	0.52	22.53	0.44	23.57	0.51	24.12	0.51
V2	✗	✓	✗	✗	✗	24.73	0.54	23.70	0.52	23.68	0.52	24.42	0.56
V3	✗	✓	✓	✗	✗	25.38	0.56	24.39	0.55	24.65	0.58	25.06	0.57
V4	✗	✓	✓	✓	✗	25.41	0.57	24.79	0.58	24.24	0.56	25.27	0.58
V5	✗	✓	✓	✓	✓	**26.11**	**0.61**	**25.23**	**0.61**	**25.00**	**0.59**	**25.46**	**0.60**

Table 2. Quantitative relighting and view synthesis result of the proposed modifications to NeRF. For each scene, we report on the left the PSNR and on the right the SSIM. Legend: $\gamma(\mathbf{R},\mathbf{t})$ indicates the use of absolute light position, $\mathbf{l}$ of relative position wrt $\mathbf{x}$, *skip* indicates that $\mathbf{x}$ is provided as input to Ψ_{rgb}, *inputs* that the embedding $\mathbf{e}$ is inserted in Ψ_{rgb} at the second layer, while Ψ_{vis} that the MLP approximating $L_i(x, \omega_i)$ is used.

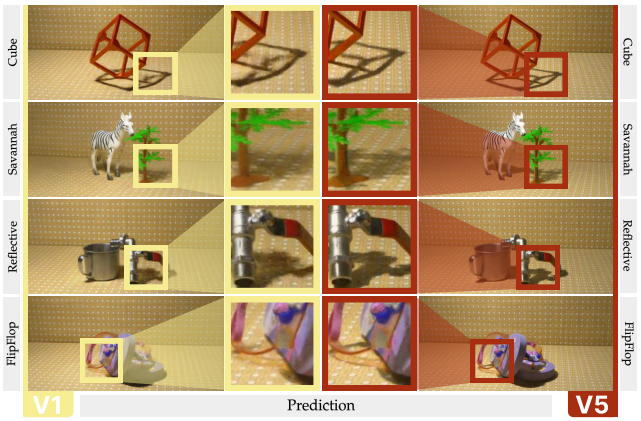


Figure 5. **Qualitative results for ablation.** The same image as rendered by V1 and V5 networks. The side-by-side comparison clearly shows how V5 reproduces much sharper details in shadows (best seen in Reflective) and is able to better handle complex reflections (best seen in FlipFlop).

5.4. Results on benchmark

Finally, we report the performance for the best architecture on the overall set of held out test scenes in Tab. 3.

	Ours				*Instant-NGP*			
Name	*Easy*		*Hard*		*Easy*		*Hard*	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Apple	26.44	0.62	26.25	0.62	20.89	0.45	20.95	0.45
Cheetah	25.66	0.61	24.64	0.60	19.37	0.44	19.60	0.44
Cube	24.90	0.54	23.98	0.53	20.14	0.42	20.31	0.42
Dinosaurs	25.75	0.65	24.98	0.64	19.58	0.42	19.66	0.41
FlipFlop	25.85	0.61	25.42	0.61	20.38	0.45	20.36	0.45
Fruits	25.93	0.62	25.72	0.62	20.16	0.45	20.21	0.44
Garden	25.74	0.66	25.08	0.66	19.76	0.45	19.70	0.45
Helicopters	25.12	0.61	24.73	0.61	19.34	0.37	19.37	0.37
Kittens	25.90	0.64	24.96	0.63	18.52	0.37	18.65	0.37
Lego	26.07	0.61	25.77	0.61	20.75	0.46	20.76	0.46
Lunch	25.84	0.60	24.71	0.59	19.32	0.46	19.38	0.45
Plant	26.55	0.67	25.93	0.67	20.62	0.44	20.66	0.44
Reflective	25.79	0.61	25.28	0.61	20.09	0.43	20.11	0.42
Robotoy	26.24	0.65	25.55	0.65	20.77	0.50	20.78	0.50
Savannah	25.15	0.62	24.31	0.61	19.08	0.40	19.18	0.40
Shark	25.59	0.57	25.32	0.56	20.54	0.42	20.53	0.41
Stegosaurus	25.87	0.63	25.65	0.63	20.84	0.43	20.91	0.42
Tapes	25.84	0.58	25.41	0.57	19.34	0.41	19.55	0.41
Trucks	25.80	0.67	25.16	0.66	19.81	0.44	19.87	0.44
Wooden toys	25.69	0.61	25.24	0.60	20.19	0.48	20.22	0.48
Average	**25.79**	**0.62**	**25.20**	**0.61**	19.97	0.43	20.04	0.43

Table 3. Quantitative results across all 20 scenes. We report both Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity (SSIM) by comparing the novel view synthetized image with its corresponding groundtruth.

We compare our method against the pytorch implementation of Nvidia Instant-NGP [37]. Overall, the proposed architecture achieves satisfying results on this challenging task and sets a non trivial baseline for the online benchmark, clearly outperforming the naive baseline represented by the standard NeRF. We show some qualitative results for the easy test split in Fig. 6, and in Fig. 7 for the hard one. These qualitative results confirm that the model is capable of convincingly rendering and lighting the scene from unseen viewpoints and light positions while using a simple and lightweight architecture. Interestingly, the output of Ψ_{vis}, reported in the rightmost columns of the figures, well approximate point-to-light visibility, especially in the easy split, even if it has not received direct supervision to emulate it, and it even reproduces some indirect illumination effects, confirming the feasibility of a neural approximation of the incident radiance $L_i(\mathbf{x}, \omega_i)$. Comparing with the ground-truth frame, while rendering of light on the object is indeed quite realistic, cast shadows are sharp and coherent with the scene geometry on the easy set, while more artifacts are present in the hard split.

6. Conclusion and Limitations

We have introduced the ReNe dataset, the first dual robot dataset framing real world objects under challenging one-light-at-time (OLAT) conditions and annotated with accurate camera and light poses. The main limitations of our dataset concern the absence of 360 degrees scans and the use of a challenging but unrealistic OLAT setup. By leveraging the training and validation splits of the dataset, we were able to perform an ablation study on lightweight modifications to a NeRF architecture that extend it to successfully perform novel view synthesis under unseen lighting. The best model emerged from the study has been tested on the held out test set to establish a non-trivial baseline for the benchmark. We hope the availability of a new dataset and baseline for the problem of NeRF relighting will attract new research around this challenging inverse rendering problem.

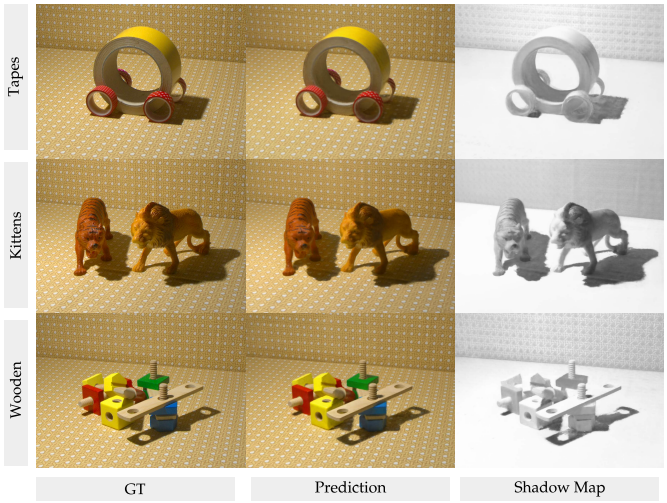


Figure 6. We show one image from the easy test split $\mathcal{I}^{easy}$ on the left. In the middle we can see the same image as rendered by our V5 model, while on the right the intermediate visibility output of Ψ_{vis} is shown, which is then multiplied by the predicted BRDF to determine the outgoing radiance at each point.

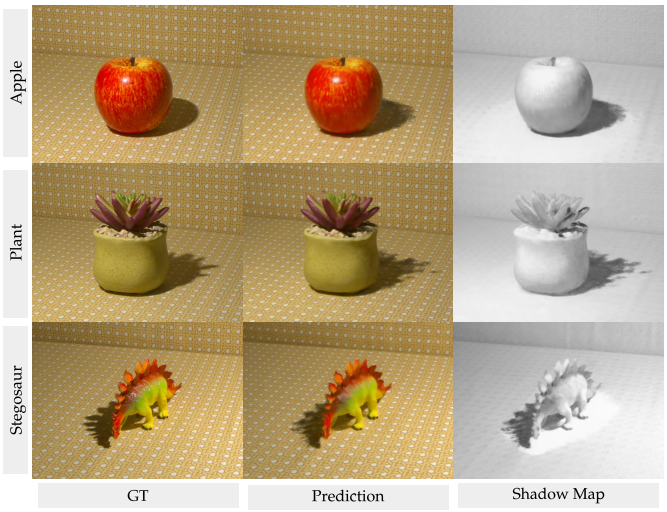


Figure 7. On the left the ground-truth image drawn from the hard test split $\mathcal{I}^{hard}$. In the middle the corresponding render produced by our V5 model. On the right the intermediate visibility output of our model.

References

[1] 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. 3

[2] Sai Bi, Zexiang Xu, Pratul Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824*, 2020. 2, 3, 6

[3] Sai Bi, Zexiang Xu, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Deep reflectance volumes: Relightable reconstructions from multi-view photometric images. In *European Conference on Computer Vision*, pages 294–311. Springer, 2020. 2

[4] 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] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 2, 3

[6] Robert L Cook and Kenneth E Torrance. A reflectance model for computer graphics. *ACM Siggraph Computer Graphics*, 15(3):307–316, 1981. 3

[7] Abe Davis, Marc Levoy, and Fredo Durand. Unstructured light fields. In *Computer Graphics Forum*, volume 31, pages 305–314. Wiley Online Library, 2012. 3

[8] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 145–156, 2000. 3

[9] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8649–8658, 2021. 2, 3

[10] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021. 3

[11] Duan Gao, Guojun Chen, Yue Dong, Pieter Peers, Kun Xu, and Xin Tong. Deferred neural lighting: free-viewpoint relighting from unstructured photographs. *ACM Transactions on Graphics (TOG)*, 39(6):1–15, 2020. 3

[12] Sergio Garrido-Jurado, Rafael Muñoz-Salinas, Francisco José Madrid-Cuevas, and Manuel Jesús Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014. 4

[13] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. siggraph'96: Proceedings of the 23rd annual conference on computer graphics and interactive techniques, 1996. 3

[14] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010. 2, 3

[15] Wonbong Jang and Lourdes Agapito. Codenerf: Disentangled neural radiance fields for object categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12949–12958, 2021. 2

[16] James T Kajiya. The rendering equation. In *Proceedings of the 13th annual conference on Computer graphics and interactive techniques*, pages 143–150, 1986. 3

[17] J A Kegerreis, V R Eke, P Gonnet, D G Korycansky, R J Massey, M Schaller, and L F A Teodoro. Planetary giant impacts: convergence of high-resolution simulations using efficient spherical initial conditions and swift. *Monthly Notices of the Royal Astronomical Society*, 487(4):5029–5040, 06 2019. 4

[18] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. 7

[19] Adam R Kosiorek, Heiko Strathmann, Daniel Zoran, Pol Moreno, Rosalia Schneider, Sona Mokrá, and Danilo Jimenez Rezende. Nerf-vae: A geometry aware 3d scene generative model. In *International Conference on Machine Learning*, pages 5742–5752. PMLR, 2021. 2, 3

[20] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996. 3

[21] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 3

[22] Zhengqi Li, Wenqi Xian, Abe Davis, and Noah Snavely. Crowdsampling the plenoptic function. In *European Conference on Computer Vision*, pages 178–196. Springer, 2020. 3

[23] David B Lindell, Julien NP Martel, and Gordon Wetzstein. Autoint: Automatic integration for fast neural volume rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14556–14565, 2021. 2

[24] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. 2

[25] Lingjie Liu, Marc Habermann, Viktor Rudnev, Kripasindhu Sarkar, Jiatao Gu, and Christian Theobalt. Neural actor: Neural free-view synthesis of human actors with pose control. *ACM Transactions on Graphics (TOG)*, 40(6):1–16, 2021. 2

[26] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 3

[27] Stephen Lombardi, Tomas Simon, Gabriel Schwartz, Michael Zollhoefer, Yaser Sheikh, and Jason Saragih. Mixture of volumetric primitives for efficient neural rendering. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 2

[28] Linjie Lyu, Ayush Tewari, Thomas Leimkühler, Marc Habermann, and Christian Theobalt. Neural radiance transfer fields for relightable novel-view synthesis with global illumination. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 153–169. Springer, 2022. 3

[29] Stephen Robert Marschner. *Inverse rendering for computer graphics*. Cornell University, 1998. 1

[30] 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. 3, 6

[31] Wojciech Matusik. *A data-driven reflectance model*. PhD thesis, Massachusetts Institute of Technology, 2003. 3

[32] Wojciech Matusik, Matthew Loper, and Hanspeter Pfister. Progressively-refined reflectance functions from natural illumination. *Rendering Techniques*, 1(2), 2004. 3

[33] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 5

[34] Abhimitra Meka, Rohit Pandey, Christian Haene, Sergio Orts-Escolano, Peter Barnum, Philip David-Son, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, et al. Deep relightable textures: volumetric performance capture with neural rendering. *ACM Transactions on Graphics (TOG)*, 39(6):1–21, 2020. 3

[35] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 3, 6, 7

[36] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1, 3, 5, 6

[37] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, July 2022. 5, 6, 7, 8

[38] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5124–5133, 2020. 3

[39] Michael Niemeyer and Andreas Geiger. Campari: Camera-aware decomposed generative neural radiance fields. In *2021 International Conference on 3D Vision (3DV)*, pages 951–961. IEEE, 2021. 2

[40] Atsuhiro Noguchi, Xiao Sun, Stephen Lin, and Tatsuya Harada. Neural articulated radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5762–5772, 2021. 3

[41] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2856–2865, 2021. 3

[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. 2, 3

[43] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 3

[44] Pieter Peers, Dhruv K Mahajan, Bruce Lamond, Abhijeet Ghosh, Wojciech Matusik, Ravi Ramamoorthi, and Paul Debevec. Compressive light transport sensing. *ACM Transactions on Graphics (TOG)*, 28(1):1–18, 2009. 3

[45] Sida Peng, Junting Dong, Qianqian Wang, Shangzhan Zhang, Qing Shuai, Hujun Bao, and Xiaowei Zhou. Animatable neural radiance fields for human body modeling. 2021. 2

[46] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2, 3

[47] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 117–128, 2001. 1

[48] Daniel Rebain, Wei Jiang, Soroosh Yazdani, Ke Li, Kwang Moo Yi, and Andrea Tagliasacchi. Derf: Decomposed radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14153–14161, 2021. 2

[49] 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. 2

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

[51] Peiran Ren, Yue Dong, Stephen Lin, Xin Tong, and Baining Guo. Image based relighting using neural networks. *ACM Transactions on Graphics (ToG)*, 34(4):1–12, 2015. 3

[52] Yoichi Sato, Mark D Wheeler, and Katsushi Ikeuchi. Object shape and reflectance modeling from observation. In *Proceedings of the 24th annual conference on Computer graphics and interactive techniques*, pages 379–387, 1997. 1

[53] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware im-

age synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 2, 3

[54] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 3

[55] 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, 3

[56] Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–184, 2019. 3

[57] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul E Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38(4):79–1, 2019. 2, 3

[58] Tiancheng Sun, Kai-En Lin, Sai Bi, Zexiang Xu, and Ravi Ramamoorthi. Nelf: Neural light-transport field for portrait view synthesis and relighting. *arXiv preprint arXiv:2107.12351*, 2021. 3

[59] Tiancheng Sun, Zexiang Xu, Xiuming Zhang, Sean Fanello, Christoph Rhemann, Paul Debevec, Yun-Ta Tsai, Jonathan T Barron, and Ravi Ramamoorthi. Light stage super-resolution: continuous high-frequency relighting. *ACM Transactions on Graphics (TOG)*, 39(6):1–12, 2020. 3

[60] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11367, 2021. 3

[61] Matthew Tancik, Ben Mildenhall, Terrance Wang, Divi Schmidt, Pratul P Srinivasan, Jonathan T Barron, and Ren Ng. Learned initializations for optimizing coordinate-based neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2846–2855, 2021. 2, 3

[62] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 2, 3

[63] 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. 2

[64] Richard Tucker and Noah Snavely. Single-view view synthesis with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 551–560, 2020. 3

[65] Bruce Walter, Stephen R Marschner, Hongsong Li, and Kenneth E Torrance. Microfacet models for refraction through rough surfaces. In *Proceedings of the 18th Eurographics conference on Rendering Techniques*, pages 195–206, 2007. 2, 3

[66] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 6

[67] Zhou Wang, Alan Bovik, Hamid Sheikh, and Eero Simoncelli. Image quality assessment: From error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13:600 – 612, 05 2004. 7

[68] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics (TOG)*, 39(6):1–13, 2020. 2, 3

[69] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021. 3

[70] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics (ToG)*, 37(4):1–13, 2018. 3

[71] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 2

[72] 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. 2

[73] Yizhou Yu, Paul Debevec, Jitendra Malik, and Tim Hawkins. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 215–224, 1999. 1

[74] Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, et al. Neural light transport for relighting and view synthesis. *ACM Transactions on Graphics (TOG)*, 40(1):1–17, 2021. 2, 3

[75] 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 (TOG)*, 40(6):1–18, 2021. 2, 3

[76] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7194–7202, 2019. 3

[77] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 3