NeRF for Outdoor Scene Relighting

Viktor Rudnev[1,2], Mohamed Elgharib[1], William Smith[3], Lingjie Liu[1],
Vladislav Golyanik[1], and Christian Theobalt[1]

[1]MPI for Informatics, SIC [2]Saarland University, SIC [3]University of York

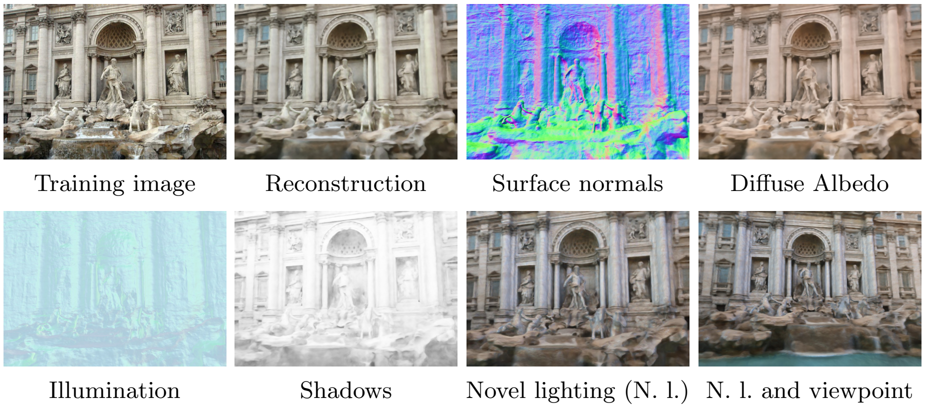


Fig. 1: **NeRF-OSR is the first neural radiance fields approach for outdoor scene relighting.** We learn a neural representation of the scene geometry, diffuse albedo and illumination-dependent shadows from a set of images capturing the same site from different viewpoints and at different times. The learnt intrinsics enable simultaneous editing of both the scene's lighting and viewpoint.

Abstract. Photorealistic editing of outdoor scenes from photographs requires a profound understanding of the image formation process and an accurate estimation of the scene geometry, reflectance and illumination. A delicate manipulation of the lighting can then be performed while keeping the scene albedo and geometry unaltered. We present NeRF-OSR, *i.e.,* the first approach for outdoor scene relighting based on neural radiance fields. In contrast to the prior art, our method allows simultaneous editing of illumination and camera viewpoint using only a collection of outdoor photos shot in uncontrolled settings. Moreover, it enables direct control over the scene illumination, as defined through a spherical harmonics model. For evaluation, we collect a new benchmark dataset of several outdoor sites photographed from multiple viewpoints and at different times[1]. For each time, a 360° environment map is captured together with a colour-calibration chequerboard to allow accurate numerical evaluations on real data against ground truth. Comparisons against SoTA show that NeRF-OSR enables controllable lighting and viewpoint editing at higher quality and with realistic self-shadowing reproduction.

[1] see the project web page `https://4dqv.mpi-inf.mpg.de/NeRF-OSR/`

1 Introduction

Controllable lighting editing of real scenes from photographs is a long-standing and challenging problem, with several applications in virtual and augmented reality [8,20,21,27,46]. It requires explicit modelling of the image formation process and an accurate estimation of the material properties and scene illumination. Such scene decomposition enables manipulating the lighting in isolation while maintaining the integrity of the remaining scene components (*e.g.,* albedo and geometry.) While several methods for controllable lighting editing exist, some solutions are dedicated to a specific class of objects such as human faces [18, 34] and human bodies [8, 21]. Other solutions are designed for processing either indoor [7, 20, 30, 33, 45, 49] or outdoor [1, 5, 27, 44, 46] scenes. Due to the very different nature of indoor and outdoor data, methods for relighting them were largely treated separately in the literature. In this work, we focus on outdoor scene relighting. Unlike existing methods [1,5,27,44,46], our approach is the first to simultaneously edit both scene illumination and camera viewpoint.

The recently proposed Neural Radiance Fields (NeRF) [23] is a powerful neural 3D scene representation capable of self-supervised training from 2D images recorded by a calibrated monocular camera [26, 39, 40, 48]. At test time, NeRF can produce photorealistic novel scene views. While there were a few attempts to extend NeRF for lighting editing [3, 19, 33, 35, 49], existing approaches are either designed for a specific object class [35], require known or single illumination condition for training [33, 49] or they do not model important outdoor illumination effects such as cast shadows [3]. Most existing NeRF-based relighting methods [3, 33, 35, 49] are not designed for outdoor scenes captured in uncontrolled settings. An exception to this is, at first sight, NeRF in the Wild (NeRF-W) [19] trained from uncontrolled images, factoring per-image appearance into an embedding space. However, NeRF-W and more recent follow-ups [4, 37] do not perform intrinsic image decomposition and thus semantically meaningful parametric control of lighting, shadows or even albedo is not possible.

This paper addresses the shortcomings of existing methods and presents NeRF-OSR, *i.e.,* the first approach based on neural radiance fields that can change both illumination and camera viewpoint of outdoor scenes photographed in uncontrolled settings, in a high-quality and semantically meaningful way; see Fig. 1. Our approach models the image formation process, disentangling the input image into its intrinsic components and scene illumination. It also contains a dedicated network for learning shadows, whose realistic reproduction is crucial for high-quality outdoor scene relighting. NeRF-OSR is trained in a self-supervised manner on multiple images of a site photographed from different viewpoints and under different illuminations. We evaluate our method qualitatively and quantitatively on a variety of outdoor scenes and show that it outperforms state of the art. Aspects of the novelty of our work include:

- NeRF-OSR, *i.e.*, the first method using neural radiance fields for outdoor scene relighting supporting simultaneous and semantically meaningful editing of scene illumination and camera viewpoint. Our model has explicit control over the scene intrinsics, including local shading, shadows and even albedo.

- Our method learns a neural scene representation that decomposes the scene into spatial occupancy, illumination, shadowing and diffuse albedo reflectance. It is trained in a self-supervised manner from outdoor data captured from various viewpoints and at different illuminations.
- A new and biggest in literature benchmark dataset for outdoor scene relighting. It includes eight buildings photographed from 3240 viewpoints and at 110 different times. In addition, it is the first one that includes colour-calibrated 360° environment maps, which allows accurate numerical evaluations.

2 Related Work

Scene Relighting. There are several methods for outdoor illumination editing [1, 5, 10–12, 27, 36, 44, 46, 47]. Some of them focus on integrating objects into images in an illumination-consistent manner [10, 44], while others process the full scene [1, 5, 11, 12, 27, 36, 46, 47]. Duchene *et al.* [5] estimate scene reflectance, shading and visibility from multiple views shot at fixed lighting. They produce novel relighting effects such as moving cast shadows. Barron *et al.* [1] formulate inverse rendering through statistical inference. Given a single RGB image of an object, their method estimates the most likely shape normals, reflectance, shading and illumination that can reproduce the examined image. They assume piecewise smooth and low-entropy reflectance images and isotropic (with frequent bends) surfaces. Philip *et al.* [27] guide relighting via a proxy geometry estimated from multi-view images. Their neural network translates image-space buffers of the examined scene into the desired relighting. The buffers include shadow masks (estimated from the extracted geometry), normal maps and illumination components. Philip *et al.*'s method is trained with high-quality synthetic data. However, it is primarily designed to edit only the illumination of the input and not the camera viewpoint. Furthermore, their illumination model is limited to sun lighting and can not handle other cases such as cloudy skies.

Yu and Smith [47] estimate the albedo, normals and lighting of an outdoor scene from just a single image; lighting is modelled through spherical harmonics (SH) with a statistical model as a prior. Relighting is then achieved by editing the reconstructed illumination (using a low-frequency model). Yu *et al.* [46] train a method for scene relighting given a single image in a self-supervised manner on a large corpus of uncontrolled outdoor images. A neural renderer takes the original albedo and geometry, the target shading and the target shadowing, and relights the scene; a dedicated network predicts the target shadows. Next, residuals of the inverse rendering are also supplied to the neural renderer as input to better capture scene details. Impressive results are shown visually and validated numerically on a new benchmark dataset. In contrast to our approach, neither Yu *et al.* [46] nor Yu and Smith [47] can edit the camera viewpoint.

Recently, there were efforts in developing relighting methods using NeRF backbone [3, 33, 35, 49]. Most of these methods operate in a setting different from ours, *i.e.,* they either require input images with a single illumination condition [49], assume a known illumination during training [33] or are designed for a specific class of objects such as faces [35]. The closest to our technique is NeRD by Boss *et al.* [3], in the sense it can operate on images of the same scene shot under different illuminations. Here, the spatially varying BRDF of the examined scene is estimated through the help of physically-based rendering. To allow fast rendering at arbitrary viewpoints and illumination, the learnt reflectance volume is converted into a relightable texture mesh. Unlike our NeRF-OSR, NeRD does not explicitly model shadows, which are crucial for high-quality outdoor scene relighting. Furthermore, it requires the examined object to be at a similar distance from all views—an assumption that can not be easily satisfied for outdoor photographs captured in an uncontrolled setup.

Style-based Editing. Scene relighting techniques are distantly related to style-based category of appearance editing methods [4,15,17,19,22,24,31,37]. Unlike relighting methods, the latter do not have a physical understanding of the scene illumination and seek to edit the overall appearance at once. Hence, they lack explicit parametric control over the local shading and shadows. In contrast, our NeRF-OSR performs scene intrinsic decomposition and seeks to edit illumination in isolation from albedo and geometry. It also directly models illumination-based shadows, which is crucial for high-quality outdoor relighting. Next, our intrisic decomposition allow editing applications that are not possible by style-based methods by any means (*e.g.,* inserting objects by editing the albedo channel separately and then relighting the entire composited new scene).

3 Method

NeRF-OSR takes as input multiple RGB images of a single scene, shot at different timings and from different viewpoints. It then renders the examined scene from an arbitrary viewpoint and under various illuminations. Our method estimates the scene intrinsics explicitly and has direct access to the scene illumination. It also includes a dedicated component for predicting shadows, *i.e.,* an essential feature of outdoor scene illumination.

An overview of NeRF-OSR is shown in Fig. 2. At its heart is a neural radiance fields (NeRF), *i.e.,* a neural implicit scene representation for volumetric rendering. Our method is trained in a self-supervised manner on outdoor data captured in uncontrolled settings and can render photorealistic views. Next, we describe in Sec. 3.1 the NeRF model [23] without view-dependent effects, which we build upon. We then discuss our illumination model and how it is adapted in a volumetric-based representation in Secs. 3.2–3.3. The objective function is presented in Sec. 3.4, followed by a discussion of the training details (Sec. 3.5).

3.1 Neural Radiance Fields (NeRF)

For each point $\mathbf{x}$ in 3D space, NeRF [23] defines its density $\sigma(\mathbf{x})$ and colour $\mathbf{c}(\mathbf{x})$. To render an image, a ray is cast from the camera origin $\mathbf{o}$, in a direction $\mathbf{d}$ corresponding to each of the output pixels. N_{depth} points $\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}$ are sampled along each ray, where $\mathbf{x}_i = \mathbf{o} + t_i\mathbf{d}$ and $\{t_i\}_{i=1}^{N_{\text{depth}}}$ are the corresponding ray

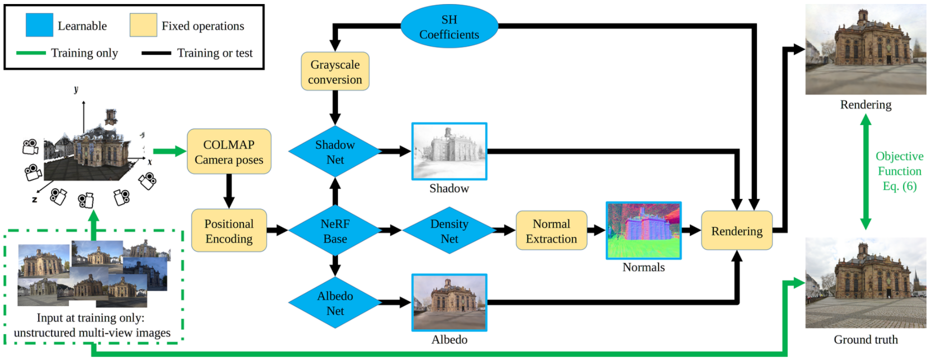


Fig. 2: **Our NeRF-OSR uses outdoor images of a site photographed in an uncontrolled setting (dashed green) to recover a relightable implicit scene model.** It learns the scene intrinsics and illumination as expressed by the SH coefficients. Here, a dedicated neural component learns shadows. During the test, NeRF-OSR can synthesise novel images at arbitrary camera viewpoints and scene illumination; the user directly supplies the desired camera pose and the scene illumination, either from an environment map or via SH coefficients.

depths. The final colour in the image space $\mathbf{C}(\mathbf{o}, \mathbf{d})$ is obtained by integrating the density and colour along the ray $(\mathbf{o}, \mathbf{d})$ as follows:

$$\mathbf{C}(\mathbf{o}, \mathbf{d}) = \mathbf{C}\left(\{\mathbf{x}_i\}_{i=1}^{N_\text{depth}}\right) = \sum_{i=1}^{N_\text{depth}} T(t_i)\alpha(\sigma(\mathbf{x}_i)\delta_i)\mathbf{c}(\mathbf{x}_i), \tag{1}$$

where $T(t_i) = \exp\left(-\sum_{j=1}^{N_\text{depth}-1} \sigma(\mathbf{x}_j)\delta_j\right)$, $\delta_i = t_{i+1} - t_i$, and $\alpha(y) = 1 - \exp(-y)$. The depths $\{t_i\}_{i=1}^{N_\text{depth}}$ are selected using stratified sampling from the uniform distribution, spanning the depths along $(\mathbf{o}, \mathbf{d})$ starting from the near and ending at the far camera plane. Both density $\sigma(\mathbf{x})$ and colour $\mathbf{c}(\mathbf{x})$ are modelled using MLPs, and the final rendering is trained in a self-supervised manner using the observed ground-truth per-pixel colours.

To better capture small details, NeRF uses *hierarchical volume sampling* for $\{t_i\}_{i=1}^{N_\text{depth}}$, *i.e.,* instead of performing a single rendering pass, points are first sampled in stratified manner. The densities at these points are then used for importance sampling in the final pass. The final model is thus learnt by supervising the rendered pixel colours of both passes with the ground-truth colours.

3.2 Spherical Harmonics NeRF

While (1) allows for high-quality free viewpoint synthesis, $\mathbf{c}(\mathbf{x})$ are defined only through an MLP that does not encode the lighting. In other words, such formulation learns a Lambertian model of the scene under a fixed lighting. The more generalised model with view direction dependencies [23] learns a slice of the apparent BRDF at a fixed illumination. Nonetheless, this learnt representation still does not have a semantic meaning of the underlying scene intrinsics and has no direct control over the lighting.

To allow relighting, we introduce an explicit 2nd-order Spherical Harmonics (SH) illumination model [2] and redefine the rendering equation (1) as follows:

$$\mathbf{C}\left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}, \mathbf{L}\right) = \mathbf{A}\left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}\right) \odot \mathbf{L}\mathbf{b}\left(\mathbf{N}\left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}\right)\right), \tag{2}$$

where $\odot$ denotes elementwise multiplication. $\mathbf{A}(\mathbf{x}) \in \mathbb{R}^3$ is the accumulated albedo colour, generated in the similar way as in (1), *i.e.,* by integrating the output of an albedo MLP. $\mathbf{L} \in \mathbb{R}^{9\times 3}$ is the per-image learnable SH coefficients, and $\mathbf{b}(\mathbf{n}) \in \mathbb{R}^9$ is the SH basis. $\mathbf{N}(\mathbf{x})$ is the surface normal computed from the accumulated ray density. It is defined as

$$\mathbf{N}\left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}\right) = \frac{\hat{\mathbf{N}}\left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}\right)}{\left\|\hat{\mathbf{N}}\left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}\right)\right\|^2}, \tag{3}$$

$$\text{where } \hat{\mathbf{N}}\left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}\right) = \sum_{i=1}^{N_{\text{depth}}} \left(\frac{\partial}{\partial \mathbf{x}_i}\sigma(\mathbf{x}_i)\right) \odot T(t_i)\alpha(\sigma(\mathbf{x}_i)\delta_i). \tag{4}$$

To extract $\mathbf{N}$, we first differentiate the density of points on the ray with respect to the original x-, y-, z-components of the ray samples, accumulate them over all N_{depth} samples on the ray with weights $T(t_i)\alpha(\sigma(\mathbf{x}_i)\delta_i)$, and normalise the resulting vector to a unit sphere. Note that in (2), we render in screen space using screen space albedo and normals accumulated from the neural volume. The accumulation makes the albedo and surface normal estimates less noisy and aids convergence. It also means we only make a single shading calculation rather than the alternative of one per sample point and accumulating shaded colours.

All terms of (2) are learnable except for the SH basis $\mathbf{b}(\cdot)$ and the normal extraction operator $\mathbf{N}(\cdot)$, which are based on fixed, explicit models. The proposed lighting model integration allows for explicit relighting by varying $\mathbf{L}$. While it accounts for Lambertian effects, it lacks direct shadow generation, which is crucial for modelling and subsequent relighting of outdoor scenes.

3.3 Shadow Generation Network

To allow for explicit shadow control during relighting, we introduce a dedicated shadow model $S\left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}, \mathbf{L}\right)$ and extend the rendering equation (2):

$$\mathbf{C}\left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}, \mathbf{L}\right) = S\left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}, \mathbf{L}\right) \mathbf{A}\left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}\right) \odot \mathbf{L}\mathbf{b}\left(\mathbf{N}\left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}\right)\right). \tag{5}$$

The shadow model is defined with a scalar computed by an MLP $s(\mathbf{x}, \mathbf{L}) \in [0, 1]$. The final shadow value is computed by accumulating along the ray into $S\left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}, \mathbf{L}\right) \in [0, 1]$, in the same way as in (1). Fig. 2 shows the high-level diagram of the proposed NeRF-OSR. Note that the shadow prediction network takes as input the SH coefficients in their grey-scale version, *i.e.,* $\mathbf{L} \in \mathbb{R}^{1\times 9}$ and not $\mathbb{R}^{3\times 9}$. This is motivated by the fact that shadows depend only on the spatial light distribution. Unlike traditional ray-tracing approaches as the one used in [27, 33], our shadow estimator operates much more efficiently, through just same single forward pass as albedo and geometry. We argue that it is a strength, as it makes the method much more computationally scalable, while still allowing us to relight using completely new illumination conditions.

3.4 Objective Function

We optimise the following loss function:

$$\mathcal{L}(\mathbf{C}, \mathbf{C}^{(\mathrm{GT})}, S) = \mathrm{MSE}(\mathbf{C}, \mathbf{C}^{(\mathrm{GT})}) + \lambda\, \mathrm{MSE}(S, 1), \tag{6}$$

where $\mathrm{MSE}(\cdot,\cdot)$ is the mean squared error. The first term is a reconstruction loss defined on the estimated colour $\mathbf{C}$ and the corresponding ground truth $\mathbf{C}^{(\mathrm{GT})}$. The second term regularises shadows. The shadow network S absorbs all greyscale lighting effects that cannot be explained by SH. To limit it to learning only shadows, we select the largest value of the regularisation strength λ that does not degrade the PSNR of the reconstructed images. Experimentation shows that removing the regulariser usually leads to S learning all the illumination components, except for the chromaticity—thus making the SH lighting useless.

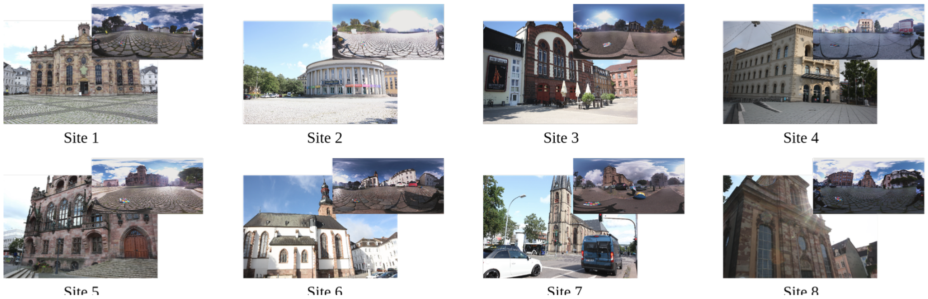
3.5 Training and Implementation Details

Our self-supervised model is trained on RGB images of an outdoor scene photographed from various viewpoints and under different illumination. We next describe several strategies for training our method and their importance.

Frequency Annealing. We noticed empirically that training the model as-is leads to noisy normal maps. Above some threshold on the number of the positional encoding (PE) frequencies, the initially generated noise (at the start of the training) becomes very hard to manipulate; it hardly converges to the correct geometry. Hence, we alleviate this by using the annealing scheme slightly modified from Deformable NeRF [26], *i.e.,* we add an annealing coefficient $\beta_k(n)$ to each of the PE components $\gamma_k(\mathbf{x})$: $\gamma'_k(\mathbf{x}) = \gamma_k(\mathbf{x})\beta_k(n)$, where $\beta_k(n) = \frac{1}{2}(1 - \cos(\pi \mathrm{clamp}(\alpha - i + N_{\mathrm{fmin}}, 0, 1)))$, $\alpha(n) = (N_{\mathrm{fmax}} - N_{\mathrm{fmin}})\frac{n}{N_{\mathrm{anneal}}}$, n is the current training iteration, N_{fmax} is the total number of used PE frequencies (the proposed model uses 12), N_{fmin} is the number of used PE frequences at the start (we use 8), N_{anneal} is tuned empirically to $3 \cdot 10^4$ for all sequences. This training strategy enables significantly improved geometry predictions.

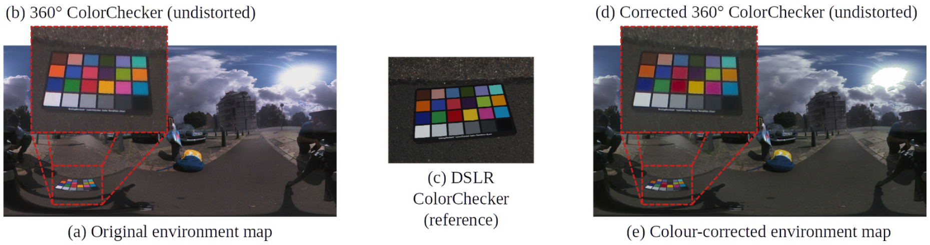
Ray Direction Jitter. To improve the generalisability of NeRF-OSR, we apply a sub-pixel jitter to the ray direction. Here, instead of shooting in the pixel centres, a jitter ψ is used as follows: $x_i = \mathbf{o} + t_i(\mathbf{d} + \psi)$. We sample ψ uniformly, such that the resulting ray still confines to the boundaries of its designated pixel.

Shadow Network Input Jitter. Since the shadows are generated in a learning-based fashion instead of using direct geometric approaches, there remains the possibility of overfitting to the training lightings. To mitigate this


Site 1 Site 2 Site 3 Site 4

Site 5 Site 6 Site 7 Site 8

Fig. 3: Sample views from the new benchmark dataset for outdoor scene relighting. The dataset has 3240 views captured in 110 different recording sessions.


(b) 360° ColorChecker (undistorted)

(d) Corrected 360° ColorChecker (undistorted)

(a) Original environment map

(c) DSLR ColorChecker (reference)

(e) Colour-corrected environment map

Fig. 4: For each recording session, we capture a colour chequerboard with the DSLR and 360° cameras. We colour-correct the 360° maps to match the DSLR.

effect, we add a slight normal noise ε to the environment coefficients as input of the shadow generation network:

$$S'\left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}, \mathbf{L}\right) = S\left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}, \mathbf{L} + \varepsilon\right), \tag{7}$$

where $\varepsilon \sim \mathcal{N}(0, 0.025I)$. (7) can be interpreted as a locality condition, *i.e.,* in similar lighting conditions, shadows should not be too different. This allows the model to learn smoother transitions between different lightings.

Implementation. We use NeRF++ [48] with the background network disabled as the code base and work within the unit sphere bounds of the foreground network. For training and evaluation, we use two Nvidia Quadro RTX 8000 GPUs. We train the model for $5 \cdot 10^5$ iterations using a batch size of 2^{10} rays, which takes ≈2 days.

4 A New Benchmark for Outdoor Scene Relighting

Several datasets for outdoor sites exist [9, 14, 15, 32, 46]. Most of them [9, 14, 15, 32] were collected with the task of 3D scene reconstruction in mind and not relighting. Hence, they mostly contain publicly available photos collected in uncontrolled settings. Furthermore, they do not provide environment maps, which are important for evaluating relighting techniques numerically on real data against ground truth. Examples of such datasets are the PhotoTourism [9, 32] and the MegaDepth [14]. The MegaDepth dataset consists of multi-view images of several sites that were initially a part of the Landmarks10k dataset [13]. Here, the depth signal is extracted using COLMAP [29] and the multi-view stereo (MVS) approach [28]. While MegaDepth was originally released as a benchmark for single-view depth extraction, it was used by Yu *et al.* [46] (one of the most recent relighting works). However, it can only evaluate methods qualitatively.

To allow for numerical evaluation on real data against ground truth, Yu *et al.* [46] recorded *one site* from different viewpoints and at different times of the day using a DSLR camera, along with the environment maps. Unfortunately, this benchmark is limited in two ways: First, it contains a single site. Second, the captured environment maps were not colour-corrected with respect to the DSLR camera of the main recordings. *Hence, numerical results obtained with this dataset would always differ from the ground truth by an unknown, possibly nonlinear, colour transformation.* Therefore, any error metric must first compute an optimal transformation (Yu *et al.* [46] used a per-colour channel linear scaling). This makes it hard to separate the behaviour of the examined relighting methods from the corrective behaviour of this normalisation.

Hence, we present a new benchmark for outdoor scene relighting. Our dataset is the first of its kind in terms of size and the ability to perform accurate numerical evaluations on real data against ground truth. It is much larger than Yu *et al.* [46], containing eight sites captured from various viewpoints using a DSLR camera (3240 viewpoints captured in 110 different recording sessions). Multiple recording sessions were performed for each site, at different times of the day; all sessions cover different weathers, including sunny and cloudy days. We also capture a 360° shot of the environment map for each session. Unlike Yu *et al.* [46], we explicitly account for the colour calibration between the environment maps and the DSLR camera of the main recordings. To this end—for every session in the test set—we also simultaneously capture the "GretagMacbeth ColorChecker" colour calibration chart with the DSLR and the 360° cameras. We then apply the second-order method of Finlayson *et al.* [6] to colour-correct the environment maps by calibrating their ColorChecker values to the ColorChecker values of the corresponding DSLR image. Finally, we manually align the environment maps to the world coordinates using COLMAP [29] reconstructions of each site.

Fig. 3 shows samples from the various sites from our dataset and the corresponding environment maps. See Fig. 4 for the colour-corrected environment maps. Note that we target scenes with minimal specular effects (such as brick or wood buildings). All data was captured in exposure brackets of five photos ranging from -3 to +3 EV for the DSLR photos and from -2 to +2 EV for the environment maps. We used the darkest capture for the 360° environment maps so that the sun is least overexposed. For the ColorChecker calibration with DSLR, we use images that are dark enough so that the white cells of the chequerboard are not overexposed. The DSLR image resolution is 5184×3456 pixel, while the resolution of the environment maps is 5660×2830 pixel.

5 Results

We evaluate the performance of NeRF-OSR on various real-world sites. We examine three sites from our newly proposed dataset and the Trevi Fountain from the PhotoTourism dataset [19]. These scenes include a variety of features. This includes large and small scale details as the sculptures (Site 1, Trevi), structural details such as trees and umbrellas (Site 3), a piecewise-smooth surfaces casting a lot of shadows on itself (Site 2), water (Trevi) and surrounding buildings (in all). Furthermore, Trevi Fountain shows performance on data collected completely from the internet through crowdsourcing. Note that only qualitative evaluation on Trevi Fountain is possible due to the absence of environment maps. We also evaluate the various design choices of our method in an ablative study.

Among existing scene relighting methods (see Sec. 2), we primarily compare against Yu *et al.* [46] and Philip *et al.* [27] as they handle a similar type of input data like ours; outdoor scenes photographed in uncontrolled settings and have a direct semantic understanding of the scene illumination. We note however, despite this, both Yu *et al.* [46] and Philip *et al.* [27] are designed to edit only the illumination of the input image, while our method can edit both the illumination and the viewpoint. This makes both these methods [27,46] not direct competitors to our method, but still the most related in literature. NeRV [33] can not be applied to our data as their setup is fundamentally different from ours. It requires a training scene to be illuminated by known lighting while our technique uses data shot in unknown lightings. We also do not compare quantitatively against NeRF-W [19] or other style-based based methods as they do not perform intrinsic decomposition, don't have a physical understanding of the scene illumination and can not edit lighting according to an environment map. In contrast, the intrinsic decomposition of NeRF-OSR enables applications that are inaccessible for style-based methods. For instance, we show how we can edit the albedo of an examined scene and relighting the entire resulting composited scene (Fig. 7-middle). We also show how our method can achieve real-time rendering with conventional computer graphics methods using the extracted mesh and albedo (Fig. 7-left).

We note that Boss *et al.* [3] (NeRD) requires the examined object to be at a similar distance from all views—an assumption that is fundamentally violated for outdoor data captured in uncontrolled setup and in our data. As confirmed by the authors of NeRD, this makes the reconstruction nearly impossible for our data. Furthermore, attempting to run NeRD on our data by the paper authors resulted in ray distance variation that is very large, requiring a large number of samples per ray. Thus it was computationally infeasible to process our data with NeRD. NeRF-OSR is the first method that can simultaneously edit the viewpoint and lighting of outdoor sites using neural radiance fields. It also extracts the underlying scene intrinsics and has a dedicated illumination-based shadow component It produces photorealistic results and significantly outperforms state of the art. It is also not limited by synthesising soft shadows only and can synthesise novel hard shadows as well (see Figs. 1 and 5).

Data Pre-Processing. Since NeRF-OSR does not aim to synthesise dynamic objects and discards them (*e.g.,* cars, people and bikes) from the training

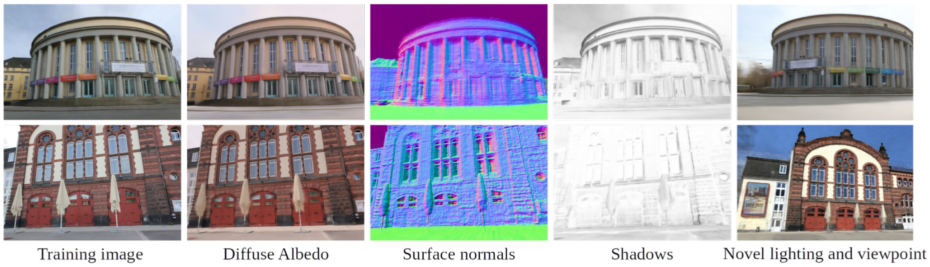


Fig. 5: Our NeRF-OSR renders photorealistic novel views and simultaneously edits lighting. It also estimates the underlying scene semantics including a dedicated shadows component. Moreover, it can also synthesise hard shadows.

stage. Although we attempted to reduce their presence during our recordings, the uncontrolled nature of the data makes eliminating them during capture impossible. We, therefore, use the segmentation method of Tao *et al.* [38] to obtain high-quality masks of such objects. Furthermore, even though NeRF-OSR can synthesise the sky and vegetation (*e.g.,* trees), it is not possible to evaluate their predictions due to their highly varying appearance, especially when recordings sessions span different weather seasons. Hence, we also estimate the masks of these regions and exclude them from our evaluation. For Sites 1–3, we keep five recording sessions for testing and use the rest for training. The resulting training/test splits are: 160/95 views for Site 1, 301/96 views for Site 2 and 258/96 views for Site 3.

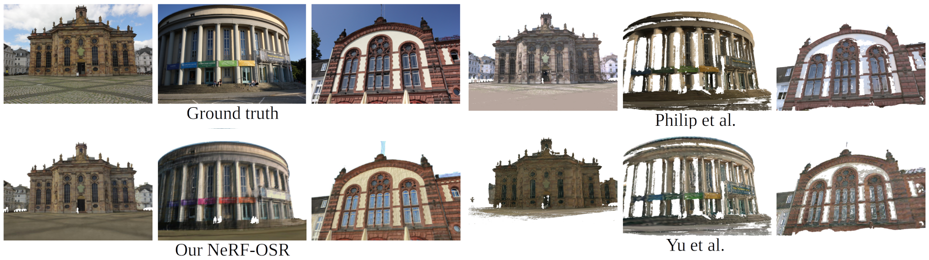
Relighting with Ground-Truth Environments. We quantitatively evaluate the parametric lighting control of our method and show that it can reproduce novel lighting using lighting coefficients extracted from environment maps. From each recording session of our dataset, we select one photo from the test set as the source. With Site 1, this gives five source images in total. We render all five images at the observed viewpoints and illumination directly. However, for Philip *et al.* [27] and Yu *et al.* [46], only the illumination of a given image can be edited. Hence, for each source image, we relight it using the illumination of the four other source images of the same site. We then cross-project the output to the camera viewpoint from which the target illumination was extracted. This is done by utilising the COLMAP reconstructions. We use segmentation masks to evaluate performance on regions where consistent predictions can be made and cross-projected for other methods. This is usually the main building. Here, we compute several metrics, including MSE, MAE and SSIM. For SSIM we report the average over the segmentation mask, using scikit-image [41] implementation. Here, we use an SSIM metric with a window size of 5 and the segmentation mask eroded by the same window size to remove the impact of the pixels outside of the mask on the metric value.

Tab. 1 reports the results of this experiment (the averages over all evaluated images). Fig. 6 shows several ground-truth images and views rendered by the compared methods. NeRF-OSR outperforms related techniques quantita-

Method	PSNR ↑	MSE ↓	MAE ↓	SSIM ↑
	Site 1			
Yu *et al.* [46]	18.71	0.014	0.088	0.4
Philip *et al.* [27] (d/s)	17.37	0.019	0.105	0.429
Ours (d/s)	**19.86**	**0.011**	**0.08**	**0.626**
Yu *et al.* [46] (u/s)	17.87	0.017	0.097	0.378
Philip *et al.* [27]	16.63	0.023	0.113	0.367
Ours	**18.72**	**0.014**	**0.09**	**0.468**
No shadows	17.82	0.017	0.101	0.418
No annealing	17.16	0.02	0.108	0.324
No ray jitter	18.43	0.015	0.093	0.433
No shadow jitter	18.28	0.016	0.095	0.413
No shadow regulariser	17.62	0.018	0.105	0.373

Method	PSNR ↑	MSE ↓	MAE ↓	SSIM ↑
	Site 2			
Yu *et al.* [46]	15.43	0.031	0.136	0.363
Philip *et al.* [27] (d/s)	11.85	0.07	0.21	0.184
Ours (d/s)	**15.83**	**0.026**	**0.128**	**0.556**
Yu *et al.* [46] (u/s)	15.28	0.032	0.138	0.385
Philip *et al.* [27]	12.34	0.065	0.2	0.272
Ours	**15.43**	**0.029**	**0.133**	**0.517**
	Site 3			
Yu *et al.* [46]	15.84	0.028	0.123	0.392
Philip *et al.* [27] (d/s)	12.85	0.054	0.169	0.164
Ours (d/s)	**17.38**	**0.021**	**0.106**	**0.576**
Yu *et al.* [46] (u/s)	15.17	0.033	0.133	0.376
Philip *et al.* [27]	12.28	0.062	0.179	0.319
Ours	**16.65**	**0.024**	**0.114**	**0.501**

Table 1: Quantitative evaluation of the relighting capabilities of different techniques. We report the metrics for Sites 1, 2 and 3 from our dataset. Our technique significantly outperforms related methods [27, 46]. "d/s" and "u/s" are shorthands for "downscaled" and "upscaled", respectively. Bottom left: ablation study of our various design choices. Our full model achieves the best result.


Ground truth

Philip et al.

Our NeRF-OSR

Yu et al.

Fig. 6: Relighting using ground-truth environment map. Since Philip *et al.* [27] and Yu *et al.* [46] can not edit the camera viewpoint—unlike NeRF-OSR—we cross-project their result on the ground-truth view. Our approach captures the illumination significantly better than related methods. See Tab. 1 for the corresponding numerical evaluations.

tively and qualitatively. While our method and Philip *et al.*generate results at 1280×844 pixel, Yu *et al.* can only generate results at 303×200 pixel. Hence in Tab. 1 we also compare against Yu *et al.* in a setting where we downscale the output of our method to Yu *et al.*'s default resolution (see d/s in Tab. 1). Despite this, our method still outperforms Yu *et al.*which shows our more superior performance is not due to differences in the output resolution. We note that comparing against style-based methods like NeRF-W here is not feasible as they do not have a semantic representation of light and thus can not edit the light according to an environment map.

Ablation Study. We evaluate the design choices of our method through an ablation study. We follow the same evaluation procedure as for the relighting comparison and report results as an average taken over all output images. For our approach, that are five images of Site 1. For Philip *et al.* [27], that are 20 images

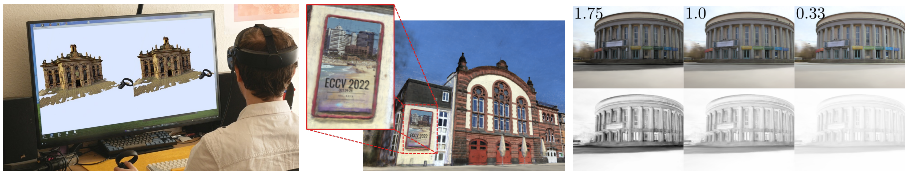


Fig. 7: (Left) Real-time interactive rendering of the extracted model in VR (screen capture is overlayed over the display for clearer image). Here, the sunlight illuminates the left side of the building and casts clear hard shadows on the right side. And an example of editing the scene albedo (middle) and shadows (right), independently of illumination and other intrinsics.

in total. Tab. 1-(bottom) reports the PSNR, MSE, MAE and SSIM of various tested settings. Results show that the best performance is obtained by using the full version of NeRF-OSR. We note that since all metrics in Tab. 1 are computed only over masked regions, they are expected to be of a higher performance if the entire image was evaluated, while filling the unmasked regions with black.

Real-time Interactive Rendering in VR. In contrast to style-based methods such as [19], our rendering is an explicit function of geometry, albedo, shadow, and the lighting conditions (see Eq. 5). Our model provides direct access to albedo and geometry. The lighting and shadows can be generated from the geometry using multiple, potentially non-differentiable, lighting models at render-time. Hence, if we can extract geometry and albedo at sufficient resolution, we can use them without the slow NeRF ray-marching at little to no loss of quality, compared to the original neural model.

We extract the geometry and albedo from the learned model of Site 1 as a mesh using Marching Cubes [16] at resolution 1000^3 voxels. Then we use them in our interactive VR renderer implemented with C++, OpenGL and SteamVR. The lighting model consists of the sun and a simple geometry-based shadow map [43]: $\mathbf{C}_{\text{interactive}} = \mathbf{C}_{\text{ambient}} + s \odot \mathbf{C}_{\text{sun}} \max\{0, \mathbf{D}_{\text{sun}}^T \mathbf{N}\}$, where $\mathbf{C}_{\text{ambient}}$ is the ambient colour, s is 0 when the rendered point is occluded and 1 if not, according to the shadow map, $\mathbf{C}_{\text{sun}}$ is the colour of the sun, $\mathbf{D}_{\text{sun}}$ is the direction of the sun and $\mathbf{N}$ is the normal of the mesh. The user can interactively move in the scene and control the sun direction with their controllers. The demo runs in real-time on a desktop computer with an Intel i7-4770 CPU, an Nvidia GeForce GTX 970 (4GB VRAM) GPU and an Oculus Rift S HMD. The system RAM usage of the application is below 3 GB. We provide an extensive demo in the supplementary video and show an extract in Fig. 7-(left).

Albedo and Shadow Editing. Another application of our intrinsic decomposition is to edit the scene albedo, without affecting the illumination or shadows. Such application is not possible by style-based methods by any means, *e.g.*, NeRF-W [19], as they do not perform image decomposition. In Fig. 7-(middle), we replace the announcement poster in Site 3 with an ECCV 2022 poster. Note how the replaced poster looks natural with the rest of the scene.

In Fig. 7-(right), we edit the shadow strength post-render. Please find extended video results of this experiment in our supplementary video, where we also show relighting results with the composited announcement poster.

6 Discussion and Conclusion

We have shown that the second-order SH lighting model is capable of producing plausible relightings. While sunlit environments can contain shadows not well represented by a second-order SH, we believe our learned shadow component compensates for this (see Figs. 7, right, for examples of novel hard shadows). Nevertheless, the SH illumination model can still be restricted in terms of high-frequency illumination, specularities and spatially varying illumination. Capturing such effects would enable reconstruction of view-dependent effects and more challenging scenes, including nighttime conditions.

Despite our method outperforms related approaches numerically and visually, some blur could exist. We believe this is due to some inaccuracies in geometry estimation. More specifically, we learn a disentangled representation of the image intrinsics, allowing many novel applications (Fig. 7). This, however, requires precise geometry, as even tiny bumps in the learned geometry can lead to significant change in normals and, hence, errors in the computed illumination. Hence the model can smooth some parts of the geometry in favour of having more accurate lighting. This leads to overall better relighting results, compared to other methods as shown in Tab. 1. Nevertheless, future work can further improve results by examining more sophisticated geometry models (*e.g.,* a hybrid volume density or implicit surface representation [25,42]). Our method needs only a set of in-the-wild photos taken from different times and views. To this end, we have evaluated our approach on our newly collected dataset and on the "Trevi" scene from [9,32]. This scene was collected completely from the Internet and is widely used in literature [15, 19, 22]. Recall that ground-truth environment maps are only used for evaluation (as in Sec. 5) and not required for our method to work. While the datasets we examined show practical use-cases of our method, future work could investigate using as few as a single illumination condition during test. Finally, incorporating more priors of the outdoor scenes could be an interesting future research direction.

Concluding Remarks. We presented the first method for simultaneous novel view and novel lighting generation of outdoor scenes captured from uncontrolled settings. We have shown that posed images with varying illumination are sufficient to train a neural representation of scene intrinsics and estimate per-image illumination. Our method outperforms related techniques subjectively and quantitatively on several sequences, including the newly collected benchmark dataset with ground-truth environment maps.

Acknowledgements. We thank Christen Millerdurai for the help with the dataset recording. This work was supported by the ERC Consolidator Grant 4DRepLy (770784).

References

1. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **37**(8), 1670–1687 (2015)
2. Basri, R., Jacobs, D.W.: Lambertian reflectance and linear subspaces. IEEE TPAMI **25**(2), 218–233 (2003)
3. Boss, M., Braun, R., Jampani, V., Barron, J.T., Liu, C., Lensch, H.P.: Nerd: Neural reflectance decomposition from image collections. In: International Conference on Computer Vision (ICCV) (2021)
4. Chen, X., Zhang, Q., Li, X., Chen, Y., Feng, Y., Wang, X., Wang, J.: Hallucinated neural radiance fields in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12943–12952 (2022)
5. Duchêne, S., Riant, C., Chaurasia, G., Moreno, J.L., Laffont, P.Y., Popov, S., Bousseau, A., Drettakis, G.: Multiview intrinsic images of outdoors scenes with an application to relighting. ACM Trans. Graph. **34**(5) (2015)
6. Finlayson, G.D., Mackiewicz, M., Hurlbert, A.: Color correction using root-polynomial regression. IEEE Transactions on Image Processing **24**(5), 1460–1470 (2015)
7. Garon, M., Sunkavalli, K., Hadap, S., Carr, N., Lalonde, J.F.: Fast spatially-varying indoor lighting estimation. In: Computer Vision and Pattern Recognition (CVPR) (2019)
8. Guo, K., Lincoln, P., Davidson, P., Busch, J., Yu, X., Whalen, M., Harvey, G., Orts-Escolano, S., Pandey, R., Dourgarian, J., Tang, D., Tkach, A., Kowdle, A., Cooper, E., Dou, M., Fanello, S., Fyffe, G., Rhemann, C., Taylor, J., Debevec, P., Izadi, S.: The relightables: Volumetric performance capture of humans with realistic relighting. ACM Trans. Graph. **38**(6) (2019)
9. Jin, Y., Mishkin, D., Mishchuk, A., Matas, J., Fua, P., Yi, K.M., Trulls, E.: Image Matching across Wide Baselines: From Paper to Practice. International Journal of Computer Vision (IJCV) (2020)
10. Karsch, K., Hedau, V., Forsyth, D., Hoiem, D.: Rendering synthetic objects into legacy photographs. ACM Trans. Graph. **30**(6) (2011)
11. Laffont, P.Y., Bousseau, A., Paris, S., Durand, F., Drettakis, G.: Coherent intrinsic images from photo collections. ACM Trans. Graph. **31**(6) (2012)
12. Lalonde, J.F., Efros, A.A., Narasimhan, S.G.: Webcam clip art: Appearance and illuminant transfer from time-lapse sequences. ACM Trans. Graph. **28**(5), 1–10 (Dec 2009)
13. Li, Y., Snavely, N., Huttenlocher, D., Fua, P.: Worldwide pose estimation using 3d point clouds. In: European Conference on Computer Vision (ECCV). pp. 15–29 (2012)
14. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Computer Vision and Pattern Recognition (CVPR) (2018)
15. Li, Z., Xian, W., Davis, A., Snavely, N.: Crowdsampling the plenoptic function. In: European Conference on Computer Vision (ECCV) (2020)
16. Lorensen, W.E., Cline, H.E.: Marching cubes: A high resolution 3d surface construction algorithm. ACM siggraph computer graphics **21**(4), 163–169 (1987)
17. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. In: Computer Vision and Pattern Recognition (CVPR). pp. 6997–7005 (2017)
18. Mallikarjun, B.R., Tewari, A., Dib, A., Weyrich, T., Bickel, B., Seidel, H.P., Pfister, H., Matusik, W., Chevallier, L., Elgharib, M., et al.: Photoapp: Photorealistic appearance editing of head portraits. ACM Transactions on Graphics **40**(4) (2021)

19. Martin-Brualla, R., Radwan, N., Sajjadi, M.S.M., Barron, J.T., Dosovitskiy, A., Duckworth, D.: NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In: Computer Vision and Pattern Recognition (CVPR) (2021)
20. Meka, A., Maximov, M., Zollhoefer, M., Chatterjee, A., Seidel, H.P., Richardt, C., Theobalt, C.: Lime: Live intrinsic material estimation. In: Computer Vision and Pattern Recognition (CVPR) (2018)
21. Meka, A., Pandey, R., Haene, C., Orts-Escolano, S., Barnum, P., Davidson, P., Erickson, D., Zhang, Y., Taylor, J., Bouaziz, S., Legendre, C., Ma, W.C., Overbeck, R., Beeler, T., Debevec, P., Izadi, S., Theobalt, C., Rhemann, C., Fanello, S.: Deep relightable textures - volumetric performance capture with neural rendering. In: ACM Transactions on Graphics (Proceedings SIGGRAPH Asia). vol. 39 (2020)
22. Meshry, M., Goldman, D.B., Khamis, S., Hoppe, H., Pandey, R., Snavely, N., Martin-Brualla, R.: Neural rerendering in the wild. In: Computer Vision and Pattern Recognition (CVPR). pp. 6871–6880 (2019)
23. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision (ECCV) (2020)
24. Nam, S., Ma, C., Chai, M., Brendel, W., Xu, N., Kim, S.: End-to-end time-lapse video synthesis from a single outdoor image. Computer Vision and Pattern Recognition (CVPR) pp. 1409–1418 (2019)
25. Oechsle, M., Peng, S., Geiger, A.: Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: International Conference on Computer Vision (ICCV) (2021)
26. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. International Conference on Computer Vision (ICCV) (2021)
27. Philip, J., Gharbi, M., Zhou, T., Efros, A.A., Drettakis, G.: Multi-view relighting using a geometry-aware network. ACM Trans. Graph. **38**(4) (2019)
28. Schönberger, J.L., Zheng, E., Frahm, J.M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: European Conference on Computer Vision (ECCV). pp. 501–518 (2016)
29. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Computer Vision and Pattern Recognition (CVPR). pp. 4104–4113 (2016)
30. Sengupta, S., Gu, J., Kim, K., Liu, G., Jacobs, D.W., Kautz, J.: Neural inverse rendering of an indoor scene from a single image. In: International Conference on Computer Vision (ICCV) (2019)
31. Shih, Y., Paris, S., Durand, F., Freeman, W.T.: Data-driven hallucination of different times of day from a single outdoor photo. ACM Trans. Graph. **32**(6) (2013)
32. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: Exploring photo collections in 3d. ACM Trans. Graph. **25**(3), 835–846 (2006)
33. Srinivasan, P.P., Deng, B., Zhang, X., Tancik, M., Mildenhall, B., Barron, J.T.: Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In: Computer Vision and Pattern Recognition (CVPR) (2021)
34. Sun, T., Barron, J.T., Tsai, Y.T., Xu, Z., Yu, X., Fyffe, G., Rhemann, C., Busch, J., Debevec, P., Ramamoorthi, R.: Single image portrait relighting. ACM Trans. Graph. **38**(4) (Jul 2019)
35. Sun, T., Lin, K.E., Bi, S., Xu, Z., Ramamoorthi, R.: Nelf: Neural light-transport field for portrait view synthesis and relighting. In: Eurographics Symposium on Rendering (2021)
36. Sunkavalli, K., Matusik, W., Pfister, H., Rusinkiewicz, S.: Factored time-lapse video. ACM Trans. Graph. **26**(3) (2007)

37. Tancik, M., Casser, V., Yan, X., Pradhan, S., Mildenhall, B., Srinivasan, P., Barron, J.T., Kretzschmar, H.: Block-NeRF: Scalable large scene neural view synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8248–8258 (2022)
38. Tao, A., Sapra, K., Catanzaro, B.: Hierarchical multi-scale attention for semantic segmentation. arXiv preprint arXiv:2005.10821 (2020)
39. Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Tretschk, E., Wang, Y., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S., Simon, T., Theobalt, C., Niessner, M., Barron, J.T., Wetzstein, G., Zollhoefer, M., Golyanik, V.: Advances in Neural Rendering. arXiv e-prints (2021)
40. Tretschk, E., Tewari, A., Golyanik, V., Zollhöfer, M., Lassner, C., Theobalt, C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In: International Conference on Computer Vision (ICCV) (2021)
41. Van der Walt, S., Schönberger, J.L., Nunez-Iglesias, J., Boulogne, F., Warner, J.D., Yager, N., Gouillart, E., Yu, T.: scikit-image: image processing in python. PeerJ **2**, e453 (2014)
42. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. Neural Information Processing Systems (NeurIPS) (2021)
43. Williams, L.: Casting curved shadows on curved surfaces. In: Proceedings of the 5th annual conference on Computer graphics and interactive techniques. pp. 270–274 (1978)
44. Xing, G., Zhou, X., Peng, Q., Liu, Y., Qin, X.: Lighting Simulation of Augmented Outdoor Scene Based on a Legacy Photograph. Computer Graphics Forum (2013)
45. Xu, Z., Sunkavalli, K., Hadap, S., Ramamoorthi, R.: Deep image-based relighting from optimal sparse samples. ACM Transactions on Graphics **37**(4), 126 (2018)
46. Yu, Y., Meka, A., Elgharib, M., Seidel, H.P., Theobalt, C., Smith, W.: Self-supervised outdoor scene relighting. In: European Conference on Computer Vision (ECCV) (2020)
47. Yu, Y., Smith, W.A.: Inverserendernet: Learning single image inverse rendering. In: Computer Vision and Pattern Recognition (CVPR) (2019)
48. Zhang, K., Riegler, G., Snavely, N., Koltun, V.: Nerf++: Analyzing and improving neural radiance fields. arXiv:2010.07492 (2020)
49. Zhang, X., Srinivasan, P.P., Deng, B., Debevec, P., Freeman, W.T., Barron, J.T.: Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. ACM Trans. Graph. (2021)