

Neural Radiance Fields for Outdoor Scene Relighting

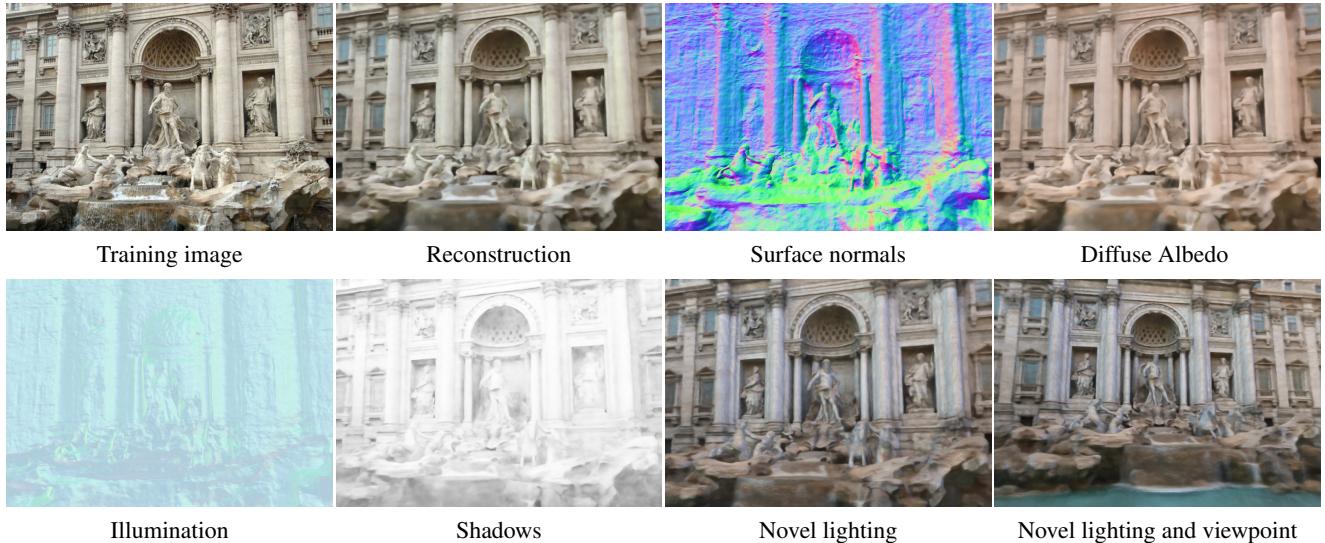
Viktor Rudnev¹Mohamed Elgharib¹William Smith²Lingjie Liu¹Vladislav Golyanik¹Christian Theobalt¹¹MPI for Informatics, SIC²University of York

Figure 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 sequence of uncontrolled outdoor images. The learnt models 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 technique allows simultaneous editing of both scene 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. It also includes a dedicated network for shadow reproduction, which is crucial for high-quality outdoor scene relighting. To evaluate the proposed method, we collect a new bench-

mark dataset of several outdoor sites, where each site is photographed from multiple viewpoints and at different timings. For each timing, 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 state of the art show that NeRF-OSR enables controllable lighting and viewpoint editing at higher quality and with realistic self-shadowing reproduction. Our method and the dataset will be made publicly available at <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 [18, 6, 40, 24, 17]. 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 [15, 31] and human bodies [18, 6]. Other solutions are designed for processing either indoor [43, 30, 39, 27, 5, 17] or outdoor [38, 3, 1, 24, 40] 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.

The recently proposed Neural Radiance Fields (NeRF) [20] is a powerful neural 3D scene representation capable of self-supervised training from 2D images recorded by a calibrated monocular camera [42, 23, 36, 35]. At test time, NeRF can produce photorealistic novel scene views. While there were a few attempts to extend NeRF for lighting editing [16, 30, 2, 32, 43], existing approaches are either designed for a specific object class [32], require known or single illumination condition for training [30, 43] or they do not model important outdoor illumination effects such as cast shadows [2]. Most existing NeRF-based relighting methods [30, 2, 32, 43] are not designed for outdoor scenes captured in uncontrolled settings. An exception to this is, at first sight, NeRF in the Wild (NeRF-W) [16]. NeRF-W is trained from uncontrolled images, factoring per-image appearance into an embedding space. However, this space has no direct physical interpretation of illumination, hence semantically meaningful parametric control of lighting or shadows is not possible.

This paper addresses the shortcomings of existing methods and presents NeRF-OSR, *i.e.*, a new 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 cast 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.*, a new method for outdoor scene relighting supporting simultaneous and semantically meaningful editing of scene illumination and camera viewpoint. Our model has explicit control over the physical illumination components, including local shading and shadows.

- 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 timings. In addition, it is the first one that includes colour-calibrated 360° environment maps, which allows accurate numerical evaluations on real data against ground truth.

2. Related Work

Scene Relighting. Several methods exist for outdoor illumination editing [33, 10, 8, 9, 38, 3, 1, 41, 40, 24]. Some of them focus on integrating objects into images in an illumination-consistent manner [8, 38], while others process the full scene [33, 10, 9, 3, 1, 24, 41, 40]. Duchene *et al.* [3] estimate scene reflectance, shading and visibility from multiple views shot at fixed lighting. Results show intrinsic decomposition with the ability to 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 searches for the most likely estimates of the shape normal, reflectance, shading and illumination that can reproduce the examined image. The work imposes several assumptions on the underlying components such as reflectance images should be piecewise smooth with low-entropy, and surfaces should be isotropic with frequent bends. Philip *et al.* [24] guide relighting through a proxy geometry estimated from multi-view images. A neural network is trained to translate image-space buffers of the examined scene into the desired relighting. The buffers include shadow masks, normal maps and illumination components. Here, the shadow masks are estimated from the extracted geometry and refined through a dedicated network. The approach of Philip *et al.* [24] is trained using high-quality synthetic data and produces impressive results on real data. However, their illumination model is limited to sun-lighting and can not handle other cases such as cloudy skies.

Yu and Smith [41] propose a monocular image reconstruction method that estimates 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 can then be achieved by editing the reconstructed illumination. While their approach shows interesting relighting, it is limited by a low-frequency illumination model that can lead to non-photorealistic results. The follow-up work addresses scene relighting given a single input image [40]. A neural renderer takes the original albedo and geometry, the target shading

and the target shadowing, and produces the desired relighting; a dedicated network predicts the target shadows. Furthermore, residuals of the inverse rendering process are also supplied to the neural renderer as input to better capture scene details during relighting. Yu *et al.* [40] train their solution in a self-supervised manner on a large corpus of uncontrolled outdoor images. Impressive results are shown visually and validated numerically on a new benchmark dataset. In contrast to our approach, neither Yu *et al.* [40] nor Yu and Smith [41] can edit the camera viewpoint.

Recently, there have been some efforts in developing relighting methods using NeRF backbone [30, 2, 32, 43]. Most of these methods operate in a setting different from ours. They either require input images with a single illumination condition [43], assume a known illumination during training [30], or are designed for a specific class of objects such as faces [32]. The closest to our technique is NeRD by Boss *et al.* [2], 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 very 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 another category of appearance editing methods that are style-based [28, 14, 19, 13, 21, 16]. Unlike relighting methods, the latter do not have a physical understanding of the scene illumination. Instead, they seek to edit the overall appearance at once. Hence, they lack explicit parametric control over the local shading and shadows. Style-based methods have largely evolved during the past years, with methods transferring the appearance of a target image [28, 14] and others modelling the appearance as a function of style [19, 13, 16, 21]. On the other hand, our NeRF-OSR directly understands the intrinsic scene 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.

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 com-

ponent 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 [20] 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 [20] 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 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:

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

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 $\{t_i\}_{i=1}^{N_{\text{depth}}}$. Here, instead of performing a single rendering pass, points are first sampled with stratified sampling. 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 [20] learns a slice of the apparent BRDF at a fixed illumination. Nonetheless, this learnt representation still does

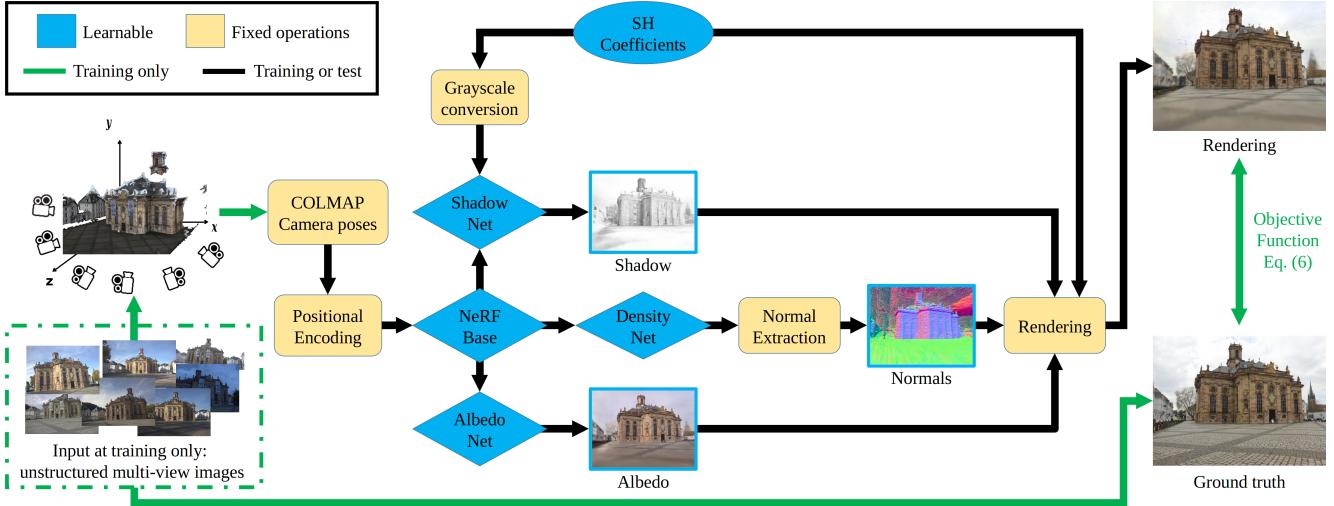


Figure 2: Our NeRF-OSR uses outdoor images of a site photographed in an uncontrolled setting (dashed green) to recover a relightable implicit scene model. It learns the scene intrinsics and illumination as expressed by the spherical harmonics (SH) coefficients. Here, a dedicated neural component learns shadows. During the test, our technique 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 directly via SH coefficients.

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 illumination model and redefine the rendering equation (1) as follows:

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

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 spherical harmonics (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

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

where

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

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 the environment \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) as follows:

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

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 Philip *et al.* [24], our shadow estimation component operates much more efficiently, through just one forward pass.

3.4. Objective Function

We optimise the following loss function:

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

where $\text{MSE}(\cdot, \cdot)$ is the mean squared error. The first term is a reconstruction loss defined on the estimated colour \mathbf{C} and the corresponding ground truth $\mathbf{C}^{(\text{GT})}$. The second term is a shadow regulariser. Here, λ controls the regularisation strength and is selected empirically as the largest value 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 [23], *i.e.*, we add an annealing coefficient $\beta_k(n)$ to each of the PE components $\gamma_k(\mathbf{x})$:

$$\gamma'_k(\mathbf{x}) = \gamma_k(\mathbf{x})\beta_k(n), \quad (7)$$

where $\beta_k(n) = \frac{1}{2}(1 - \cos(\pi \text{clamp}(\alpha - i + N_{\text{fmin}}, 0, 1)))$, $\alpha(n) = (N_{\text{fmax}} - N_{\text{fmin}})\frac{n}{N_{\text{anneal}}}$, 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 frequencies 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:

$$x_i = \mathbf{o} + t_i(\mathbf{d} + \psi), \quad (8)$$

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

$$S' \left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}, \mathbf{L} \right) = S \left(\{\mathbf{x}_i\}_{i=1}^{N_{\text{depth}}}, \mathbf{L} + \varepsilon \right), \quad (9)$$

where $\varepsilon \sim \mathcal{N}(0, 0.025I)$. (9) 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++ [42] code with the background network disabled as the base, thus only working 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 Benchmark for Outdoor Scene Relighting

Several datasets for outdoor sites exist [12, 13, 29, 7, 40]. However, most of them [12, 29, 7, 13] were collected with the task of 3D scene reconstruction in mind and not relighting. Hence, they were collected from highly uncontrolled settings, using publicly available photos from the internet. Furthermore, they do not contain environment maps, which are important for evaluating relighting techniques numerically on real data against ground truth. Examples of such datasets are the PhotoTourism [29, 7] and the MegaDepth [12]. The MegaDepth dataset consists of multi-view images of several sites that were initially a part of the Landmarks10k dataset [11]. Here, the depth signal is extracted using COLMAP [26] and the multi-view stereo (MVS) approach [25]. While MegaDepth was originally released as a benchmark for single-view depth extraction methods, it was used by Yu *et al.* [40], which is 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.* [40] presented a new dataset for outdoor scene relighting. Here, they recorded one site from different viewpoints and at different times of the day using a DSLR camera. The environment map for each time of the day was captured too. While this data is a new benchmark for outdoor scene relighting, it remains limited in two main ways. First, it only contains footage for one 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



Figure 3: Sample views from the new benchmark dataset for outdoor scene relighting. Our dataset contains eight sites shot from various viewpoints and at different timings using a DSLR camera and a 360° camera to capture the environment map. A colour chequerboard is also captured from both the DSLR and 360° cameras to account for colour calibration (see Fig. 4). The dataset has 3240 views captured in 110 different recording sessions.



Figure 4: For each recording session in our new benchmark dataset, we capture a colour chequerboard by the DSLR (a) and 360° (b) cameras. The chequerboards are used to colour-correct the environment maps with respect to the DSLR (see (e) and (f) for variants before and after the correction). This allows, for the first time, accurate numerical evaluations of outdoor scene relighting methods on real data against ground truth. Note that (c) shows the original colour chequerboard of (b) being reprojected from the 360° view into the regular view; its resulting colour-corrected version is shown in (d).

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.* [40] 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.

We present a new benchmark dataset 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. Our dataset is much larger than Yu *et al.* [40], containing eight sites captured from various viewpoints using a DSLR camera. Multiple recording sessions were performed for each site, at different times of the day. We also capture a 360° shot of the environment map for each session. However, unlike Yu *et al.* [40], we explicitly account for the colour calibration between the environment maps and the DSLR camera of the main record-

ings. To this end, for every session in the test set, we also capture the “GretagMacbeth ColorChecker” colour calibration chart with the DSLR and the 360° cameras simultaneously. We then apply the second-order method of Finlayson *et al.* [4] 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 [26] reconstructions of each site.

Fig. 3 shows samples from the different sites captured in our dataset, with their corresponding environment maps. Furthermore, Fig. 4 shows an example of the colour correction performed on the environment maps. Our dataset contains a total of 3240 viewpoints captured in 110 different recording sessions. The sessions cover different weathers, including sunny and cloudy days. All data were 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 resolution of the DSLR images 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 Photo-Tourism dataset [16]. Note that only qualitative evaluation on Trevi Fountain is possible. We also evaluate the various design choices of our method in an ablative study.

While several algorithms exist for scene relighting (see Sec. 2), we compare against Yu *et al.* [40] and Philip *et al.* [24], *i.e.*, approaches that 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 do not compare against Boss *et al.* [2] as it requires the examined object to be at a similar distance from all views—an assumption that is violated for outdoor data captured in uncontrolled setup. We also do not compare quantitatively against NeRF-W [16] or other style-based based methods as they do not have a physical understanding of the scene illumination.

NeRF-OSR is the first method that can simultaneously edit the viewpoint and lighting of outdoor sites. It also extracts the underlying scene intrinsics and has a dedicated illumination-based shadow component (see Figs. 1 and 5). It produces photorealistic results and significantly outperforms state of the art.

5.1. Data Pre-Processing

Since NeRF-OSR does not aim to synthesise dynamic objects, it is important to discard such objects (*e.g.*, cars, people and bikes) from the training 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.* [34] 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.

5.2. 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.* [24] and Yu *et al.* [40], 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. 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 quantitatively and qualitatively. Note that while ours and Philip *et al.* generate results at 1280×844 pixel, Yu *et al.* can only generate results at 303×200 pixel. Hence, we report the results for Yu *et al.* at the resolutions 303×200 pixel and 1280×844 pixel.

5.3. Ablation Study

We evaluate the design choices of our method through an ablation study. We follow the same evaluation procedure of Sec. 5.2 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.* [24], that are 20 images in total. Tab. 1-(bottom) reports the PSNR, MSE and MAE of various tested settings. Results show that the best performance is obtained by using the full version of NeRF-OSR.

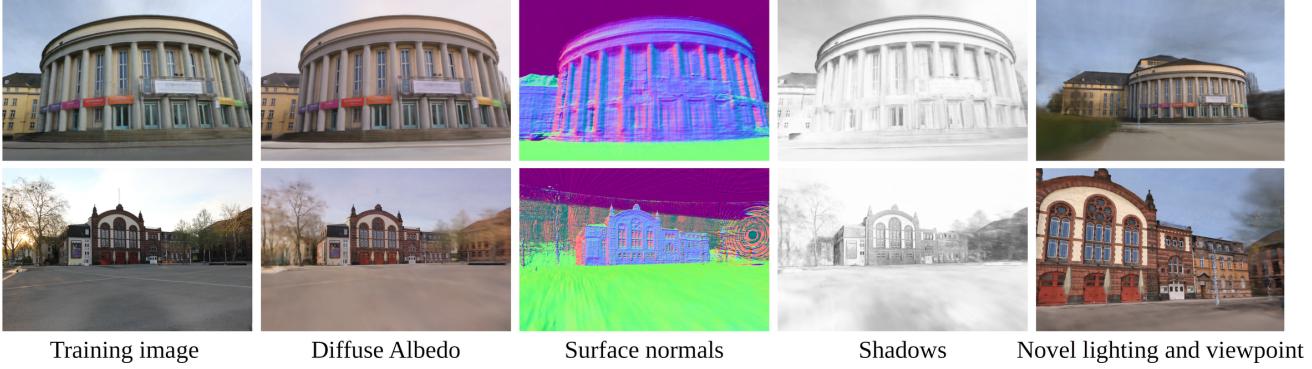


Figure 5: Our NeRF-OSR renders photorealistic novel views and simultaneously edits lighting. It also estimates the underlying scene semantics including a dedicated shadows component.

Method	PSNR \uparrow	MSE \downarrow	MAE \downarrow
Yu <i>et al.</i> [40]	18.71	0.0138	0.0881
Philip <i>et al.</i> [24] (d/s)	17.37	0.0194	0.1046
Ours (d/s)	19.86	0.0114	0.0802
Yu <i>et al.</i> [40] (u/s)	17.87	0.0167	0.0967
Philip <i>et al.</i> [24]	16.63	0.0229	0.1131
Ours	18.72	0.0143	0.0893
No shadows	17.82	0.0172	0.1012
No annealing	17.16	0.0195	0.1082
No ray jitter	18.43	0.0150	0.0931
No shadow jitter	18.28	0.0155	0.0954
No shadow regulariser	17.62	0.0181	0.1046

Table 1: Quantitative evaluation of the relighting capabilities of different techniques. We report the mean reconstruction error for Site 1 from our dataset. Our technique significantly outperforms related methods [40, 24]. “d/s” and “u/s” are shorthands for “downscaled” and “upscaled”, respectively. Bottom: ablation study of our various design choices. Our full model achieves the best result.

6. Limitations and Conclusions

We have 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 intrinsic scene properties and estimate per-image illumination.

There are many exciting extensions to consider for future work. Our SH illumination model is restricted in terms of the Lambertian reflectance assumption and restriction to low-frequency lighting effects. Capturing high-frequency illumination, specularities and spatially varying illumination would enable reconstruction of view-dependent effects and more challenging scenes, including nighttime conditions. Ideally, models of illumination and reflectance would

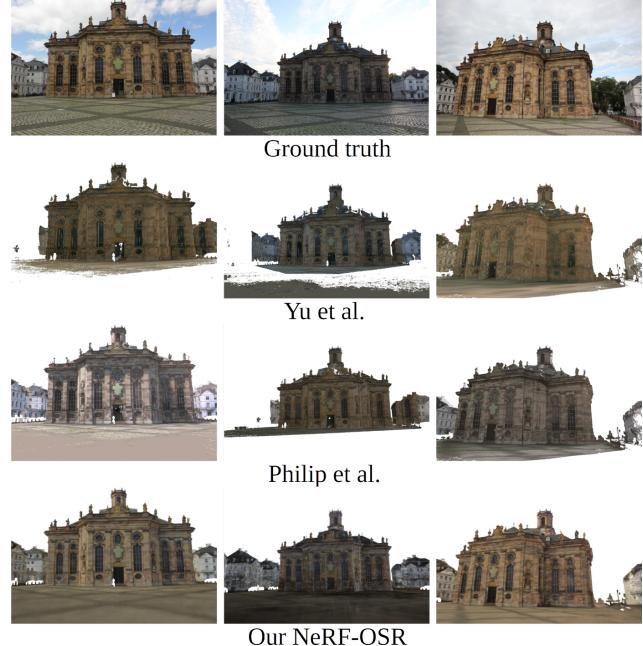


Figure 6: Relighting using ground-truth environment map. Since Philip *et al.* [24] and Yu *et al.* [40] can not edit the camera viewpoint—unlike NeRF-OSR—we cross-project their result on the ground-truth view. Our approach captures the illumination significantly better than related methods. See Tab. 1 for the corresponding numerical evaluations.

themselves be learnt from data in a self-supervised manner. Realism could be improved by including a discriminator, which could be scene-agnostic, allowing data for many scenes to improve the quality of a single scene model. An alternative geometry model (*e.g.*, a hybrid volume density or implicit surface representation [22, 37]) may yield higher quality surface normals. Including explicit geometric reasoning into our shadow calculation will likely improve the appearance of hard cast shadows.

References

- [1] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 37(8):1670–1687, 2015. 2
- [2] Mark Boss, Raphael Braun, Varun Jampani, Jonathan T. Barron, Ce Liu, and Hendrik P.A. Lensch. Nerd: Neural reflectance decomposition from image collections. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 7
- [3] Sylvain Duchêne, Clement Riant, Gaurav Chaurasia, Jorge Lopez Moreno, Pierre-Yves Laffont, Stefan Popov, Adrien Bousseau, and George Drettakis. Multiview intrinsic images of outdoors scenes with an application to relighting. *ACM Trans. Graph.*, 34(5), 2015. 2
- [4] Graham D Finlayson, Michal Mackiewicz, and Anya Hurlbert. Color correction using root-polynomial regression. *IEEE Transactions on Image Processing*, 24(5):1460–1470, 2015. 7
- [5] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-Francois Lalonde. Fast spatially-varying indoor lighting estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. 2
- [6] Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escalano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izadi. The re-lightables: Volumetric performance capture of humans with realistic relighting. *ACM Trans. Graph.*, 38(6), 2019. 1, 2
- [7] Yuhe Jin, Dmytro Mishkin, Anastasiia Mishchuk, Jiri Matas, Pascal Fua, Kwang Moo Yi, and Eduard Trulls. Image Matching across Wide Baselines: From Paper to Practice. *International Journal of Computer Vision (IJCV)*, 2020. 5
- [8] Kevin Karsch, Varsha Hedau, David Forsyth, and Derek Hoiem. Rendering synthetic objects into legacy photographs. *ACM Trans. Graph.*, 30(6), 2011. 2
- [9] Pierre-Yves Laffont, Adrien Bousseau, Sylvain Paris, Frédéric Durand, and George Drettakis. Coherent intrinsic images from photo collections. *ACM Trans. Graph.*, 31(6), 2012. 2
- [10] Jean-François Lalonde, Alexei A. Efros, and Srinivasa G. Narasimhan. Webcam clip art: Appearance and illuminant transfer from time-lapse sequences. *ACM Trans. Graph.*, 28(5):1–10, Dec. 2009. 2
- [11] Yunpeng Li, Noah Snavely, Dan Huttenlocher, and Pascal Fua. Worldwide pose estimation using 3d point clouds. In *European Conference on Computer Vision (ECCV)*, pages 15–29, 2012. 5
- [12] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 5
- [13] Zhengqi Li, Wenqi Xian, Abe Davis, and Noah Snavely. Crowdsampling the plenoptic function. In *European Conference on Computer Vision (ECCV)*, 2020. 3, 5
- [14] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6997–7005, 2017. 3
- [15] B R Mallikarjun, Ayush Tewari, Abdallah Dib, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Louis Chevallier, Mohamed Elgharib, et al. Photoapp: Photorealistic appearance editing of head portraits. *ACM Transactions on Graphics*, 40(4), 2021. 2
- [16] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. In *Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 7, 11, 13
- [17] Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. Lime: Live intrinsic material estimation. In *Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2
- [18] Abhimitra Meka, Rohit Pandey, Christian Haene, Sergio Orts-Escalano, Peter Barnum, Philip Davidson, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, Chloe Legendre, Wan-Chun Ma, Ryan Overbeck, Thabo Beeler, Paul Debevec, Shahram Izadi, Christian Theobalt, Christoph Rhemann, and Sean Fanello. Deep relightable textures - volumetric performance capture with neural rendering. In *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia)*, volume 39, 2020. 1, 2
- [19] Moustafa Meshry, Dan B. Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. In *Computer Vision and Pattern Recognition (CVPR)*, pages 6871–6880, 2019. 3
- [20] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3
- [21] Seonghyeon Nam, Chongyang Ma, M. Chai, William Brendel, N. Xu, and S. Kim. End-to-end time-lapse video synthesis from a single outdoor image. *Computer Vision and Pattern Recognition (CVPR)*, pages 1409–1418, 2019. 3
- [22] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *International Conference on Computer Vision (ICCV)*, 2021. 8
- [23] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *International Conference on Computer Vision (ICCV)*, 2021. 2, 5
- [24] Julien Philip, Michaël Gharbi, Tinghui Zhou, Alexei A. Efros, and George Drettakis. Multi-view relighting using a geometry-aware network. *ACM Trans. Graph.*, 38(4), 2019. 1, 2, 5, 7, 8
- [25] Johannes L. Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, pages 501–518, 2016. 5

- [26] Johannes L. Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Computer Vision and Pattern Recognition (CVPR)*, pages 4104–4113, 2016. [5](#) [7](#)
- [27] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W. Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *International Conference on Computer Vision (ICCV)*, 2019. [2](#)
- [28] Yichang Shih, Sylvain Paris, Frédo Durand, and William T. Freeman. Data-driven hallucination of different times of day from a single outdoor photo. *ACM Trans. Graph.*, 32(6), 2013. [3](#)
- [29] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, 2006. [5](#)
- [30] 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 *Computer Vision and Pattern Recognition (CVPR)*, 2021. [2](#) [3](#)
- [31] Tiancheng Sun, Jonathan T. Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38(4), July 2019. [2](#)
- [32] Tiancheng Sun, Kai-En Lin, Sai Bi, Zexiang Xu, and Ravi Ramamoorthi. Nelf: Neural light-transport field for portrait view synthesis and relighting. In *Eurographics Symposium on Rendering*, 2021. [2](#) [3](#)
- [33] Kalyan Sunkavalli, Wojciech Matusik, Hanspeter Pfister, and Szymon Rusinkiewicz. Factored time-lapse video. *ACM Trans. Graph.*, 26(3), 2007. [2](#)
- [34] Andrew Tao, Karan Sapra, and Bryan Catanzaro. Hierarchical multi-scale attention for semantic segmentation. *arXiv preprint arXiv:2005.10821*, 2020. [7](#)
- [35] Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, Yifan Wang, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, Tomas Simon, Christian Theobalt, Matthias Niessner, Jonathan T. Barron, Gordon Wetzstein, Michael Zollhoefer, and Vladislav Golyanik. Advances in Neural Rendering. *arXiv e-prints*, 2021. [2](#)
- [36] 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 *International Conference on Computer Vision (ICCV)*, 2021. [2](#)
- [37] 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. *Neural Information Processing Systems (NeurIPS)*, 2021. [8](#)
- [38] Guanyu Xing, Xuehong Zhou, Qunsheng Peng, Yanli Liu, and Xueying Qin. Lighting Simulation of Augmented Outdoor Scene Based on a Legacy Photograph. *Computer Graphics Forum*, 2013. [2](#)
- [39] Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. Deep image-based relighting from optimal sparse samples. *ACM Transactions on Graphics*, 37(4):126, 2018. [2](#)
- [40] Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and Will Smith. Self-supervised outdoor scene relighting. In *European Conference on Computer Vision (ECCV)*, 2020. [1](#) [2](#) [3](#) [5](#) [6](#) [7](#) [8](#)
- [41] Ye Yu and William AP Smith. Inverserendernet: Learning single image inverse rendering. In *Computer Vision and Pattern Recognition (CVPR)*, 2019. [2](#) [3](#)
- [42] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv:2010.07492*, 2020. [2](#) [5](#)
- [43] 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. *arXiv*, 2021. [2](#) [3](#)

Appendix

This appendix provides more details on the new dataset for outdoor scene relighting and the experiments. For video visualisations, see our project web page <http://4dqv.mpi-inf.mpg.de/NeRF-OSR/>.

A. Statistics of the Dataset

	Sessions	Views
Site 1	18	373
Site 2	17	423
Site 3	16	372
Site 4	11	401
Site 5	13	493
Site 6	12	379
Site 7	11	468
Site 8	12	331
Total	110	3240

Table 2: Statistics of our new benchmark dataset. The dataset currently contains eight sites recorded in 110 different sessions, each with a 360° environment map captured by LG R105. The total number of views captured by a DSLR camera Canon EOS 550D is 3240.

Tab. 2 lists the statistics of our new dataset. It consists of eight sites photographed from 3240 views in 110 different sessions. Our dataset is the first to allow numerical evaluation of relighting methods on real data against ground truth, thanks to the environment maps and the captured colour chequerboards. Kindly note that the dataset is in continuous expansion. We believe it will be valuable for the community, and we plan to release it.

B. Ablative Study

Fig. 7 demonstrates the impact of the design choices in NeRF-OSR on the final novel view renderings with relighting. Not using frequency annealing leads to an evident degradation in the output (the fourth row). This includes circular-shaped artefacts on the ground (the second and the third columns) and clear artefacts on the building (the first three columns, from the left). Removing the shadow regulariser often causes the shadow layer to learn all the illumination components, except the chromaticity, leading to significant artefacts (the first and the last columns). Removing the ray jitter leads to clear artefacts, as shown in the first column. Finally, removing shadow learning and shadow jitter produces less accurate reconstruction (the third column). The strength of shadow learning is more evident during timelapse relighting (see the videos on the project web page). The full model produces the best results, which is also reflected numerically in Tab. 1 of the main manuscript.

C. Video Results

We demonstrate the ability of NeRF-OSR to edit the camera viewpoint and illumination in videos that can be found on the project web page. Thus, we show timelapse relighting, where the camera viewpoint is fixed and the illumination changes by rotating the lighting 360° around the building. Our approach handles known lighting conditions and can generalise well to new ones. Even though some synthesised lightings can not occur in real life (*e.g.*, due to the sun trajectory covering only 180° of the sky at most), NeRF-OSR still produces a highly photorealistic output. We also change the viewpoint while keeping the scene illumination fixed. Moreover, we show results when both scene illumination and viewpoint were not seen during the training. Finally, we visualise the scene intrinsics, *i.e.*, normals, albedo, shadow and shading, as recovered by our technique.

Fig. 8 provides several screenshots from the video results, including simultaneous relighting and novel view synthesis of Site 3 (Fig. 8-(a)). Next, NeRF-OSR can be used for relighting using unrealistic and synthetic lighting (Fig. 8-(b)). Finally, while NeRF-W [16] can interpolate between the learnt appearances, our method decomposes the scene in its intrinsic components; it enables manipulation of the lighting, which results in interpretable editing of the novel views (Fig. 8-(c))).

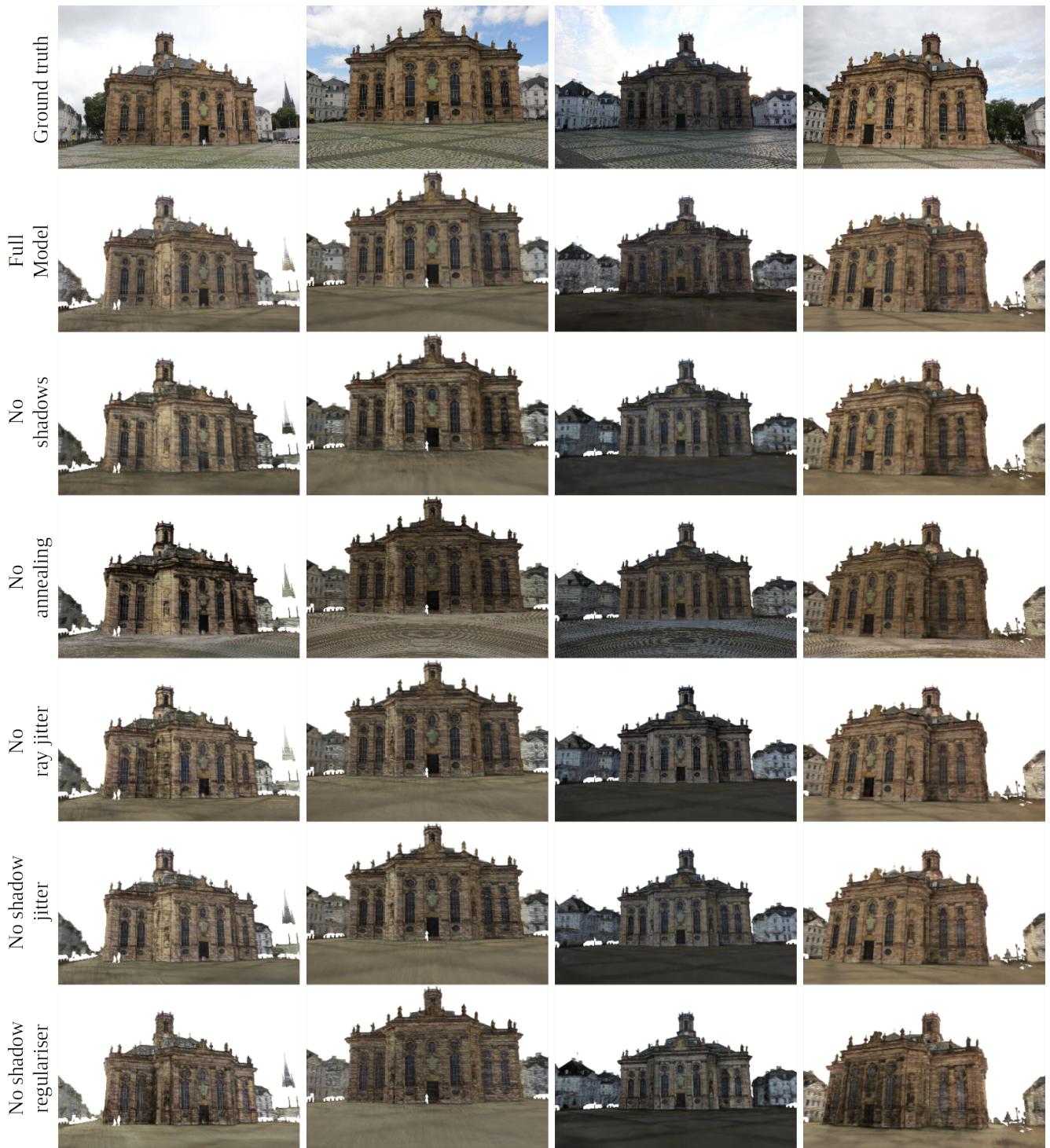


Figure 7: The impact of the various design choices in NeRF-OSR (Site 1). The four columns show the view-lighting combinations used in the quantitative evaluation against the ground truth (Tab. 1 of the main manuscript). The best result is obtained using the full model (the second row from the top). Best viewed with zoom.



Figure 8: Additional visualisations for various experiments. (a): Relighting and novel view synthesis of Site 3; (b): Relighting of Site 2 using natural and unrealistic light sources (illuminations); (c): Qualitative comparisons to NeRF-W [16]. For the corresponding full videos, see our project web page.