Self-supervised Outdoor Scene Relighting

Ye Yu[1], Abhimitra Meka[2], Mohamed Elgharib[2], Hans-Peter Seidel[2], Christian Theobalt[2], and William A. P. Smith[1]

[1] University of York, United Kingdom
{yy1571,william.smith}@york.ac.uk
[2] Max Planck Institute for Informatics, Saarland Informatics Campus, Germany

Abstract. Outdoor scene relighting is a challenging problem that requires good understanding of the scene geometry, illumination and albedo. Current techniques are completely supervised, requiring high quality synthetic renderings to train a solution. Such renderings are synthesized using priors learned from limited data. In contrast, we propose a self-supervised approach for relighting. Our approach is trained only on corpora of images collected from the internet without any user-supervision. This virtually endless source of training data allows training a general relighting solution. Our approach first decomposes an image into its albedo, geometry and illumination. A novel relighting is then produced by modifying the illumination parameters. Our solution capture shadow using a dedicated shadow prediction map, and does not rely on accurate geometry estimation. We evaluate our technique subjectively and objectively using a new dataset with ground-truth relighting. Results show the ability of our technique to produce photo-realistic and physically plausible results, that generalizes to unseen scenes.
Keywords: neural rendering, image relighting, inverse rendering

1 Introduction

Virtual relighting of real world outdoor scenes is an important problem that has wide applications. Performing such a relighting task involves correctly estimating and editing the various scene components – geometry, reflectance and the direct and indirect lighting effects. Measuring these high-dimensional parameters traditionally required the use of instruments such as LIDAR scanners and gonio-reflectometers and extensive manual effort [42,44]. This problem has been simplified by using only a small number of 2D images of a scene in a process known as image based rendering (IBR), but this leads to far fewer constraints on the unknown variables and runs into the problem of ill-posedness.

Multi-view and multi-illumination constraints have proved to be effective in solving this problem [12,4,31,48]. 2D images of a scene from different viewpoints and under different lighting conditions provide the necessary constraints to reconstruct the geometry of the scene and disambiguate the lighting from the reflectance. For example, the method of Laffont *et al.* [12], along with multi-view 3D reconstruction, also uses manual interactions to perform an intrinsic decomposition of the scene images into reflectance and shading layers. By re-projecting the reflectance layer from one viewpoint to another and recombining

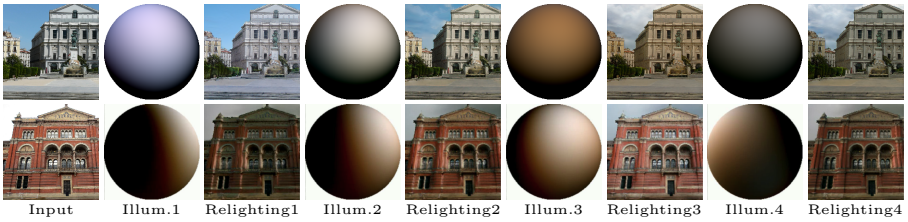

Fig. 1: We present a novel self-supervised technique to photorealistically relight an outdoor scene from a single image to any given target illumination condition. Our method is able to generate plausible shading, shadows, color-cast and sky region in the output image, while preserving the high-frequency details of the scene reflectance.

with the original shading image, lighting conditions of one image of a scene can be transferred to another. While this technique is effective, it is limited in its relighting capability because it cannot relight the scene under an arbitrary lighting condition of choice. The method of Duchene *et al.* [4] also performs a similar intrinsic decomposition of multi-view images, and additionally estimate the shadows and the parameters of a sun-lighting model for the scene. These parameters are then modified in a geometrically accurate way to achieve scene relighting. Philip *et al.* [31] similarly estimate shadows and sun-light model parameters, but skip the inverse rendering process and instead use a deep neural network to directly generate relighting results. Their network takes as input several 'illumination buffers' that are rendered using the reconstructed geometry and estimated sun-light model parameters. This method relies on high-quality ground-truth renderings of synthetic 3D models of outdoor scenes, requiring the availability of high-end computational hardware. While these techniques have been shown to generate high-quality relighting results on real scenes, they are limited by the availability of a multi-view images of the scene. They also rely on a sun-lighting model that only works for bright sunlight conditions and does not generalize to cloudy overcast skies, night-time lighting, or other desired target illumination conditions.

Another class of methods circumvent the problem of estimating the scene parameters by achieving relighting directly through lighting style transfer. These methods [37,13] change the lighting in a scene by learning the colour characteristics of images at different times of the day. Another set of methods [11,18] learn a more general class of style-transfer in which characteristics of a reference image are transferred to a target image, including the scene lighting. Such methods are not physically based and are limited in relighting a scene either based on a reference image or a particular time of the day.

In contrast, the method of Yu and Smith [48] proposes a novel formulation for the problem that allows for fully controlled relighting based on a single image of the scene. They demonstrate a learning method that at training time uses the constraints available from multi-view casual images of outdoor scenes sourced from the internet, to learn to estimate the scene appearance parameters. The network can then at test time estimate these parameters from a single image. By

modifying the lighting to a desired lighting environment, the image can be relit. While this method enabled relighting of a scene from a single 2D image to any arbitrary lighting, it was also limited by the low-frequency lighting model used in the decomposition that lead to non-photorealistic relighting results.

Recently, the advent of adversarial learning technique [6] has enabled neural networks to generate photorealistic images. 'Neural rendering' techniques based on this principle have shown promising results in various allied tasks such as novel-view synthesis [20], view-dependent effects rendering [43] and appearance modification [25].

Motivated by these two advances, we propose the first fully self-supervised neural rendering framework for performing photorealistic relighting of an outdoor scene from a single image with full lighting controllability (see Figure 1). Similar to the method of Yu and Smith [48], our method learns to estimate scene appearance parameters based on multi-view constraints at training time, without using any ground-truth synthetic 3D renderings. At test time, it takes as input a single 2D image and estimates the underlying appearance parameters such albedo, shading, shadows, lighting and normals. These physical parameters are then fed to a novel neural rendering framework, along with target lighting conditions, to generate photorealistic relighting of the scene and the sky region. By training our system in a completely self-supervised manner, it generalizes to unseen novel scenes and any target lighting condition of choice as provided by the user in the form of an environmental light map. We introduce a new high-resolution HDR multi-view & multi-illuminant evaluation dataset for outdoor relighting, and our extensive test results on the dataset show the efficacy of our method.

In summary, our main technical contributions are:

- The first fully-automatic single-image based relighting technique for outdoor scenes with full controllability of target lighting
- A novel self-supervised neural rendering framework that uses physical intrinsic decomposition layers of the scene to generate photorealistic relighting results without using any ground-truth data or synthetic 3D rendering
- A sky generation network that generates plausible sky region for the scene under a given target lighting environment
- A high-quality evaluation dataset for outdoor relighting with ground-truth HDR environment maps.

2 Related Work

Relighting a scene is a complex task. In order to perform physically accurate relighting, all components of light-transport in the scene need to be measured and modified, in a process known as inverse rendering [30]. Traditionally, this involved using special optical equipment to measure the geometry [50,17,22], surface reflectance [3,45,21,44,23] and environmental illumination [2,9,14,40], while also inverting the global illumination within the scene [49]. Image-based relighting techniques have attempted to simplify the problem by using only 2D images for the task. But using only 2D images makes the problem highly under-constrained and ambiguous.

Due to the ambiguous nature of the problem, recently there has been a lot of interest in applying learning based methods to solving it [52,38,5,27]. We restrict our discussion to methods that perform scene level relighting. Due to the very different nature of geometry and illumination in indoor and outdoor scenes, the two have often been treated as separate class of inverse rendering problems. Inverse Rendering in outdoor scenes has usually dealt with specific illumination models for natural illumination. [46] propose a single-image approach that accounts for environment lighting in outdoor scenes. Collections of photographs of a scene have been used to provide better constraints for relighting [7,36]. While we also use a dataset of casual photography of particular scenes to learn to perform inverse rendering and relighting, but at test time, we only rely on a single image of a scene to perform photorealistic relighting. The method of [37] performs lighting transfer by matching a single image to a large database of timelapses, but cannot treat cast shadows. Alternatively, online digital terrain and urban models registered to images can be used for approximate relighting [11].

Several methods on multi-view image relighting have been developed, both for the case of multiple images sharing single lighting conditions [4], and for images of the same location with multiple lighting conditions (typically from internet photo collections) [12]. For the single lighting condition, [4], first perform shadow classification and intrinsic decomposition using separate optimization steps. Despite impressive results, artifacts remain especially around shadow boundaries and the relighting method fails beyond limited shadow motion. More recently, several learning based methods have been suggested to perform relighting in outdoor scenarios [48,47,31,34]. A simpler version of the relighting problem, is of integrating virtual objects into real scenes in an illumination-consistent manner, have been solved by using proxy geometry and user interaction [10,46,28,24]. But these methods do not solve the problem of general relighting of scenes. Webcam sequences have also been used for relighting [13,41], although cast shadows often require manual layering.

3 Overview

Neural inverse rendering has been recently shown to enable convincing decomposition of both indoor [35] and outdoor [48] uncontrolled scenes into geometry (normal map), illumination and reflectance. These methods are self-supervised via a physics-based model of image formation. Such models are typically based on simple assumptions such as perfect Lambertian reflectance and ignore global illumination effects and shadowing. For this reason, re-illumination of the geometry and reflectance with novel lighting does not lead to photorealistic images. In addition, sky regions do not adhere to reflectance models, and so they are either missing from relighting results or the original sky is pasted back, making it inconsistent with the new lighting.

Our goal in this paper, motivated by the recent advances in neural inverse rendering [48], is to learn in a fully self-supervised fashion to perform photorealistic relighting of outdoor scenes from a single image. Photorealism is achieved by replacing classical model-based renderers with a learnt neural renderer that can

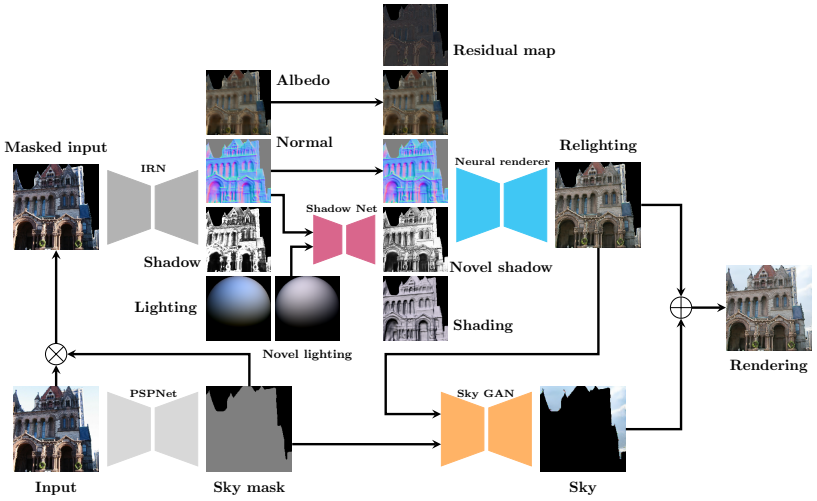

Fig. 2: For a given *Input* image of an outdoor scene, our method first performs a physical decomposition of the scene into various components. Using a pre-trained segmentation network (**PSPNet** [51]), the scene is separated from the sky. The scene is then decomposed by the **InverseRenderNet (IRN) [48]** into intrinsic image layers of *Albedo*, *Normal*, *Shadow* and *Lighting* . Given a target *Novel lighting* condition, **ShadowNet** uses the regressed scene normals to generate a target *Novel shadow* map for the scene. The scene albedo and normals, along with target lighting, shadow map, target shading and residual input map (see Section 5) are then fed to the **Neural renderer** to generate plausible *Relighting* of the scene. Given the output of the neural renderer, **SkyGAN** generates a convincing *Sky* region, and by compositing these together, a complete photorealistically relit *Rendering* is achieved.

take as input the various scene parameters along with a target lighting condition and generate plausible relighting results. The neural renderer particularly learns to synthesize global illumination effects such as plausible shadows, inter-reflections and view-dependent effects that are required for photorealism, which are much more difficult to simulate with model-based renderers. The neural renderer is trained using an adversarial loss to ensure that the generated images lie within the distribution of real images. A novel cycle consistency loss and direct supervision loss via cross projection of multi-view images is also used to ensure that the generated images exhibit the desired target lighting. We also present a sky generation network that learns to synthesize plausible skies that are consistent with the lighting within the rest of the image. An overview of our approach is shown in Figure 2.

4 Inverse rendering

We take as our starting point the inverse rendering network of Yu and Smith [48]. InverseRenderNet comprises an image-to-image network that estimates colour diffuse albedo, $\boldsymbol{\alpha}(p) = [\alpha_r(p), \alpha_g(p), \alpha_b(p)]^T$, and surface normal direction,

$\mathbf{n}(p) \in \mathbb{R}^3$, $\|\mathbf{n}(p)\| = 1$, for each pixel p. Illumination is represented using the parameters, $\mathbf{L} \in \mathbb{R}^{3\times 9}$, of an order 2 spherical harmonics model [32] leading to the following image formation model:

$$\mathbf{i}(p) = \boldsymbol{\alpha}(p) \odot \mathbf{L}\mathbf{b}(\mathbf{n}(p)), \tag{1}$$

where $\mathbf{b}(\mathbf{n}(p)) \in \mathbb{R}^9$ contains the spherical harmonic basis for normal direction $\mathbf{n}(p)$, $\odot$ is the elementwise product and $\mathbf{i}(p)$ the RGB colour at pixel p $\mathbf{L}$ is computed by solving a least squares system over all foreground pixels, $p \in \mathcal{F}$, i.e. those not labelled as sky by a PSPNet segmentation network [51]. $\mathbf{L}$ is further restricted to a statistical subspace learnt from real, outdoor environment maps. The self-supervision loss is provided by the residual error in (1).

In the context of relighting, the main drawback of InverseRenderNet is that the model used for self-supervision cannot adequately describe real world appearance. So, unmodelled phenomena such as cast shadows, spatially varying illumination and specularities are baked into albedo and normal maps. Of these phenomena, the most severe are shadows. When baked into the albedo map, relit images retain the shadows of the original illumination. When baked into the normal map, relit images contain shading artefacts caused by warped normals.

We propose a novel variant of InverseRenderNet that explicitly estimates an additional channel, $s(p)$, to explain these unmodelled phenomena and avoid them being baked into the albedo or normal maps. The additional channel acts multiplicatively on the appearance predicted by the local spherical harmonics model:

$$\mathbf{i}(p) = s(p)\boldsymbol{\alpha}(p) \odot \mathbf{L}\mathbf{b}(\mathbf{n}(p)). \tag{2}$$

Without appropriate constraint, the introduction of this additional channel could lead to trivial solutions. Hence, we constrain it in two ways. First, we restrict it to the range $[0, 1]$ so that it can only downscale appearance. Second, it is a scalar quantity acting equally on all colour channels. Together, these restrictions encourage this channel to explain cast shadows and we refer to it as a shadow map. However, note that we do not expect it to be a physically valid shadow map nor that it contains only shadows. During training, we compute our self-supervised appearance loss in a shadow free space:

$$\ell_{\text{appearance}} = \sum_p \left\| \min\left(1, \frac{\mathbf{i}(p)}{s(p)}\right) - \boldsymbol{\alpha}(p) \odot \mathbf{L}\mathbf{b}(\mathbf{n}(p)) \right\|^2, \tag{3}$$

i.e. we compare the appearance predicted by the local illumination model against the original image with shadows divided out.

For training, we use the same dataset, training schedule and hyperparameters as the InverseRenderNet and retain multiview supervision. Specifically, albedo consistency and direct normal map supervision are applied in the same way while the cross-rendering loss (mixing lighting from one view with albedo and normal map from another) is formulated in the shadow free space as in (3). Explicitly modelling shadow allows us to drop generic priors used in InverseRenderNet (albedo smoothness and pseudo-supervision) such that addressing problems like

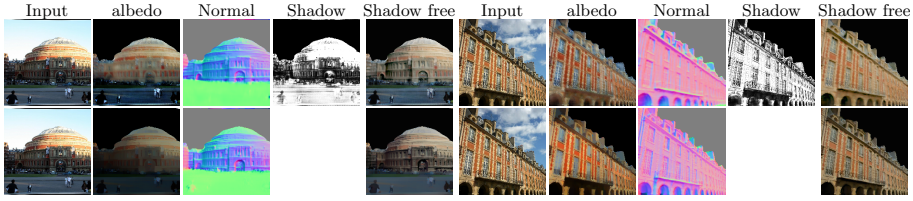

Fig. 3: Inverse rendering with shadow prediction. Rows 1: proposed variant, rows 2: original InverseRenderNet [48].

oversmoothing. We show some sample qualitative results in Figure 3. Note that, relative to InverseRenderNet, cast shadows are not baked into the albedo map such that the shadow free rendering removes their effect.

5 Neural rendering

We now describe our neural rendering network. This can be viewed as a conditional GAN [26] in which the conditioning input is the maps required for a Lambertian rendering and the latent space is the spherical harmonic lighting parameter space. The objective of the network is to generate images indistinguishable from real ones while keeping the lighting consistent with the target lighting parameters.

The input to the neural rendering network is constructed from the outputs of InverseRenderNet (see Figure. 2). The albedo and normals are taken as direct inputs from the output of InverseRendernet, because they are scene invariants. Additional inputs of a shading map and a shadow map consistent with the target illumination are constructed. The shading channel is obtained using the Lambertian spherical harmonic lighting model under the desired lighting with the estimated normal map. The shadow map for a given novel lighting condition is predicted using a separate shadow prediction network described in Section 5.2.

We concatenate the albedo prediction (3 channels), normal prediction (3 channels), shading (3 channels), shadow map (1 channel) and sky segmentation (1 channel) into an 11 dimensional tensor. In addition to this tensor, we compute another 3-channel *residual map* that contains the lost fine-scale details from original image after inverse rendering decomposition. The residual map is computed by subtracting Lambertian rendering composed by inverse rendering results from original input image. We then stack this residual map at the end of concatenated 11 dimensional tensor and feed it to the neural rendering network.

5.1 Losses

We use three classes of loss function in order to train the neural renderer. First, an adversarial loss ensures the realism of the generated images. Second, direct supervision is provided in the form of self-reconstruction and cross-projection rendering losses to ensure the images are accurate predictions of the scene appearance under desired lighting conditions. Third, this direct supervision is aided by a cycle consistency loss that uses InverseRenderNet to consistent decompositions of original and rendered images.

Adversarial loss For adversarial loss we use the multiscale LSGAN [19] architecture. Real images are true images with the sky masked out. Fake images are the neural renderings, again with all pixels in the sky region set to black.

Direct supervision Our training set provides real example images under a variety of illumination conditions. We can exploit these for direct supervision. When the chosen lighting condition for relighting is the same as the original image, we expect the neural rendering to exactly match the original image. We refer to this as self-reconstruction loss. In practice, this is computed as a sum of the VGG perceptual loss [39] (difference in VGG features from the first two layers) and ℓ_2 distance in LAB colour space. However, self-reconstruction loss does not penalise baked-in effects. To overcome this, we use multiview supervision. A mini-batch consists of a set of overlapping images with different illumination and which can be cross projected from one view to another using the multi-view stereo (MVS) reconstructed geometry and camera parameters. We use this for additional direct supervision. Within a mini-batch, we shuffle the lighting estimates from InverseRenderNet so that we relight the albedo and normal predictions from one view with the lighting from another. We rotate the spherical harmonic lighting to account for the relative pose between views. Supervision is provided by comparing the neural rendering against the cross projection of the view from which the lighting was taken, again measured in terms of VGG perceptual loss and ℓ_2 distance in LAB space. However, errors in the MVS geometry and camera poses cause slight misalignments in the cross projected images. We found that applying this loss at full resolution led to a blurry output. For this reason, before computing the cross projection loss, we downscale both the cross projected and rendered images by a factor of 4.

Cycle consistency We found that direct supervision and adversarial loss alone are insufficient for good performance and smooth relighting under smooth illumination parameter changes. This is partly due to the fact that cross projected images are incomplete and can be quite sparse when the view change is large. Therefore, to improve stability we propose to also include a cycle consistency loss. Here, we use the InverseRenderNet trained as described in Section 4 and measure the consistency between the input maps to the neural renderer and those obtained by decomposing the neural rendered image. Specifically, we penalise the difference in the albedo, normal, lighting and shadow maps. Lighting consistency is measured by the sum of VGG perceptual loss and ℓ_2 difference between the Lambertian shading maps. Normal map consistency is measured by the mean angular error between original and estimated normal maps. For albedo consistency, we weight the error by the shading map. The idea is that albedo estimates in darkly shaded regions are unlikely to be accurate and we do not wish to overemphasise errors in these regions. Again, the albedo difference is measured in terms of VGG perceptual loss and ℓ_2 distance in LAB space.

5.2 Shadow prediction network

When illumination changes, the shadowing changes. To estimate such changes in shadows, we train a separate shadow prediction network. It takes as input a

normal map and the spherical harmonic lighting vector and outputs a shadow map. In order to input the lighting vector while retaining the image-to-image architecture of the network, we replicate the 27D lighting vector (since $\mathbf{L} \in \mathbb{R}^{3\times9}$) pixel-wise and attach it to normal map such that the input is a 30D tensor. We train the shadow prediction network using illumination, normal and shadow maps predicted by our modified InverseRenderNet.

5.3 Sky GAN

Our physical illumination model is only able to describe non-sky regions of the image. Sky cannot be meaningfully represented in terms of geometry, reflectance and lighting. Moreover, sky appearance is partially stochastic (the precise arrangement of clouds is not informative). For this reason, we train a second network specifically to generate skies that are plausible given the rest of the image. For example, if the image contains strong cast shadows and shading, one would expect a clear sky with sunlight coming from an appropriate direction. If the image is highly diffuse with little discernible shading one would expect an cloudy sky.

For this purpose, we use the GauGAN architecture [29] with two semantic classes: sky and foreground. This network performs sky generation from random noise and conditional inputs of the sky segmentation mask and the foreground image with black sky. The output is the sky image which is blended with the foreground image using the binary sky mask. Such binary blended images are inputs to the discriminator along with the sky mask as a conditional input. Hence, the discriminator loss will help generate both more realistic skies but also skies that are plausible given the foreground appearance.

To train the generator, we use the adversarial loss and the feature matching loss as in [29] but remove other appearance losses. We train using real images in which sky has been masked to black. The discriminator is trained using the same loss as the original GauGAN [29]. We find that, in practice, this network generalises well to foregrounds generated using our neural rendering network.

5.4 Training

Our network graph is implemented in tensorflow. The neural rendering network and shadow prediction network are modelled as UNET architectures [33] and The skyGAN network and InverseRenderNet were modelled after ResNet architecture [8]. For details of our network architectures and training hyperparameters, please refer to the supplementary document.

The training of the networks is performed in several stage. The inverse rendering network is trained indepedently as the first step. The output of inverse rendering network is used to train the shadow prediction network. Given the well-trained shadow prediction network and inverse rendering network, the neural rendering network is trained. The training of the neural rendering network is done in two phases. In the first phase only a self-reconstruction loss is employed and this stage is stopped when the loss reaches a steady-state value. In the second phase, the cycle-consistency loss and adversarial loss are added. In the experiments, we found such pre-training step ensures fast convergence and leads to renderings containing more fine details.

Fig. 4: We present a new high-quality high-resolution outdoor relighting dataset. Our dataset consists of high-resolution HDR images of a single monument captured under several different lighting conditions from multiple views, along with the ground-truth HDR environment light maps.

Similar to Yu and Smith [48], we run our training and testing on the megaDepth dataset [16]. The dataset contains multiview stereo images, which enable us to directly train inverse rendering network and find relative rotations between image views before shuffling illumination estimates. The dataset contains a variety of outdoor scenes. All training images were resized to a size of 200×200 pixels to keep the training tractable on single-gpu hardware.

6 Results

6.1 Outdoor Relighting Bechmarking Dataset

We present a new high-quality benchmarking dataset for the evaluation of outdoor relighting techniques. The dataset consists of several sets of multi-view, multi-illumination high dynamic range (HDR) images of a single monument, along with ground-truth HDR environment maps for each illumination condition. We captured 6 different lighting conditions, including clear sky with bright sunlight, cloudy overcast sky and evening light. For each lighting condition, we capture 10 images from views around the monuments and also the ground-truth environment light map. Each image of the monument is of resolution 5184×3456, captured with a Canon 5D Mark II DSLR camera with an 18mm focal length lens. It consists of 6 multi-exposure raw captures, which are fused in Adobe Photoshop to generate an HDR image. The lowest camera exposure time is chosen to ensure that the captured image has minimal amount of pixel saturation from bright light sources such as the sun. We use constant ISO and aperture settings in the capture. The environment light map is captured using a 360 degree camera (LG360) with 6 multi-exposure shots fused to obtain the HDR image.

While the original environment maps are captured from arbitrary viewpoints, in order to perform view consistent relighting, the environment maps need to be rotated to align them to the same viewpoint as the camera images. This is achieved by performing multi-view 3D reconstruction of the monument from all the dataset images and estimating accurate camera pose for each camera view

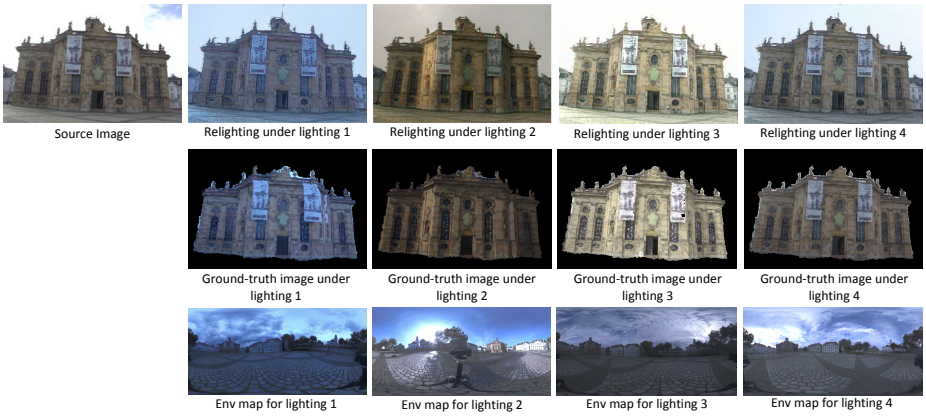

Fig. 5: Relighting result on our new high-quality outdoor religting dataset. Note the plausible shading effects obtained by our method on the surfaces of the monument compared to the ground-truth.

through bundle adjustment. The rotation between environment map and the global co-ordinate system of the monument (taken as the camera co-ordinate system of the first camera view image) is computed by performing a sparse feature match between the environment map and the 3D model and optimizing for the camera rotation between the two. This process is repeated for each of the 6 lighting conditions. In the dataset, we provide the camera pose for every image and also the rotation for each of the 6 ground-truth environment maps to the first camera view image. This provides 'aligned environment maps' for each lighting condition. Please see the supplementary document for an illustration of this alignment process. A low-frequency representation of each captured environment map is also provided by computing the 2nd order spherical harmonics co-efficients that fit the light map.

6.2 Qualitative Evaluation

On the Benchmarking Dataset Our benchmarking dataset is used for qualitative evaluation of our method. We perform cross-relighting of the monument by taking an image for a particular lighting condition as input and performing relighting to another target light condition using as input the 2nd order spherical harmonic co-efficients of the ground-truth 'aligned' environment light map. The results for such relighting is shown in Fig. 5, where an image captured under a source lighting is relighted to several target lighting conditions, and Fig 7, where source image is relighted to target image under GT illumination. As can be seen, our method is able to generate relighting result that closely resembles the ground-truth images for each target lighting condition. Our method does a particularly good job of estimating plausible color-cast and shading across various surfaces of the monument including those with intricate geometry.

On test dataset In Figure 6, we show relighting results on our test split from MegaDepth data and comparison with other single-image relighting approaches.

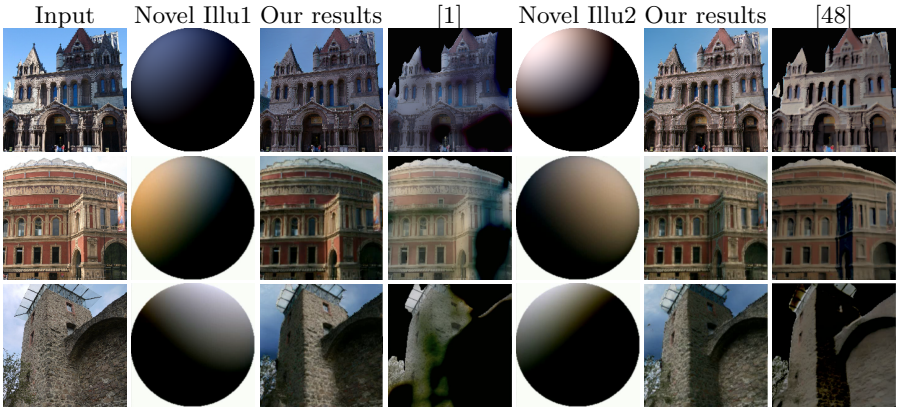

Fig. 6: Relighting results from testing data. It shows the comparison between our methods with InverseRenderNet [48] and SIRFS [1].

Fig. 7: Relighting of benchmark dataset images and comparison with Philip *et al.*[31], Yu and Smith [48] and Barron and Malik [1].

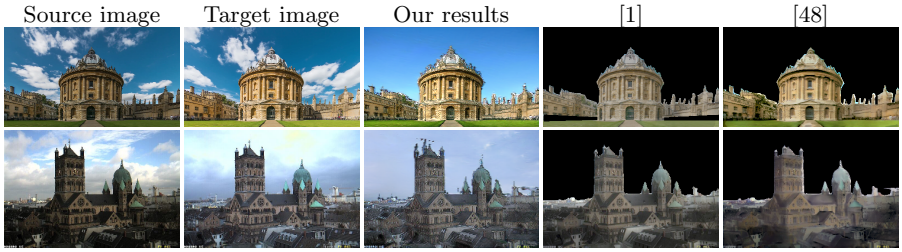

Fig. 8: Relighting of BigTime images and comparison with Yu and Smith [48] and Barron and Malik [1].

Our method results in realistic looking relighting results with shading and shadows that are very consistent with the target lighting condition, while maintaining the fine underlying reflectance details. We also generate sky regions which match the general colour tone of the relit structure. The method of Yu and Smith [48] generates non-photorealistic images due to their simple Lambertian reflectance model. The method of Barron and Malik [1] struggles with the darker sides of the target lighting conditions because they cannot account for global illumination.

On time-lapse dataset We evaluate our neural rendering network on BigTime[15] dataset, which contains approximately 200 time-lapse image sequences of indoor and outdoor scenes. For each time lapse sequence, we perform cross-rendering

by relighting each frame with lighting estimates from all the other frames in the sequence. The qualitative comparison between our method and other methods is shown in Figure 8. It is evident that our method preserves the colour-cast and the brightness scale better, and is able to generate accurate relighting effects such as consistent shading and shadows.

6.3 Quantitative Evaluation

We also perform quantitative evaluations on BigTime[15] and our benchmarking data. To evaluate the relit results on BigTime[15], we use multiple error metrics computed between the relit result and corresponding real image. The quantitative comparison, averaged over 15 sequences, is shown in Table 1. It is shown that our network can generalise well to time-lapse image sequences. Our method has the best performance on ℓ_1 error and the mean square error (mse) and is comparable to the method of Barron and Malik [1] on metrics measuring structural information like SSIM and DSSIM. Barron and Malik's [1] method seems to perform slightly better on these metrics because their method tends to baking albedo/reflectance details into shading. While this leads to preserving details in the output and better SSIM score (depending on how close the target and source lighting are), it is in general not a desirable quality (see Fig. 6 for failure cases). This issue with their method is concealed when evaluating this dataset since the relighting is based on their estimated lighting.

Figure 7 shows example of the cross-relighting that we perform across all lighting conditions in the benchmark dataset. In order to get the ground-truth image for our relighting, we project all the camera images from a given target lighting condition onto the 3D geometry of the monument and average them. This is then re-projected to the camera viewpoint of the source image to obtain the ground-truth relit image. Although this leads to the loss of view-dependent effects, it still provides a plausible ground-truth image with accurate shadows and shading. Error metric is computed as ℓ_1 error averaged over the reprojected pixels of the monument, see Table 2. Our method generates plausible relighting results close the the ground-truth image and produces the least error in most cases, while the other techniques struggle to preserve the high-frequency details, the colour-cast and the shading variations. For the method of Philip *et. al.* [31], we were able to obtain cross-relighting results only in specific cases since their sun-lighting model cannot be applied to cloud or evening skies. Only in one case, their method was able to outperform ours quantitatively. Please note that their method uses the full multi-view dataset for relighting whereas our method relights a single image.

More results and ablation study can be found in supplementary document.

7 Discussion

While our method generalizes well to various new scenes, it may be ill-posed for darker input images because sufficient information is not available due to limited photometric resolution of the camera sensor at lower light intensity levels to perform an accurate decomposition. Our method also struggles with strong cast

Method	ℓ_1	mse	SSIM	DSSIM
Proposed	**0.103**	**0.021**	0.760	0.120
[48]	0.117	0.26	0.722	0.139
[1]	0.115	0.24	**0.770**	**0.115**

Table 1: Quantitative evaluation on the BigTime [15] time-lapse dataset. The error values are computed by averaging over 15 sequences.

Method	Original lighting condition											
	1		2		3		4		5		6	
	ℓ_1	SSIM	ℓ_1	SSIM	ℓ_1	SSIM	ℓ_1	SSIM	ℓ_1	SSIM	ℓ_1	SSIM
Proposed	**0.077**	0.871	**0.078**	**0.850**	**0.074**	**0.876**	**0.075**	0.872	**0.076**	0.842	**0.073**	**0.839**
[48]	0.082	0.824	0.085	0.780	0.087	0.791	0.083	0.818	0.079	0.819	0.077	0.810
[1]	0.083	**0.879**	0.097	0.826	0.091	0.852	0.080	**0.883**	0.086	0.840	0.098	0.814
[31]									$0.095^{\dagger}$	**0.871**	$0.083^{\ddagger}$	0.834

Table 2: Mean ℓ_1 colour error (lower is better) and SSIM index (higher is better) for relit images against cross projected ground-truth. Results are averaged across all images and all target lighting conditions. (†averaged over only target lighting condition 6 because the authors of method provided their results for only one target lighting condition.)(‡averaged over only target lighting conditions 2 & 5 for the same reason.)

shadows. For similar reasons, a strong cast shadow in the input is a challenge for the inverse rendering network because it leads to non-linearity in the pixel-value vs. radiance curve which is difficult to recover. Conversely, generating strong cast shadows is also a challenge for the neural renderer. Generating such shadows involves simulating the physical ray-tracing process which requires a knowledge of the full 3D scene geometry. An interesting way of dealing with this would be to ensure that the rendering network is aware of such inaccuracies in the decomposition by training the entire pipeline end-to-end and make the network implicitly aware of the 3D scene geometry.

Our method, while capable of generating non-lambertian effects and thus relighting results with greater realism, does not explicitly model them. This may sometimes lead to incorrect specularities that are not accurate reflections based on the position of the light source, the surface normals and the viewing direction. An explicit non-lambertian reflectance model and decomposition and a corresponding neural rendering pipeline would solve such an issue.

8 Conclusion

We present a novel self-supervised single-image based relighting framework for outdoor scenes and an outdoor relighting benchmark dataset. This neural rendering framework based on self-supervision from casual photography can also be extended in the future to lighting augmentation tasks such as addition or removal of existing light sources in the scene, opening up interesting applications in augmented and virtual reality domain.

Acknowledgments

This work was funded by the ERC Consolidator Grant *4DRepLy* (770784).

References

1. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. TPAMI (2015)
2. Debevec, P.: Image-based lighting. IEEE Comput. Graph. Appl. **22**(2), 26–34 (Mar 2002)
3. Debevec, P., Hawkins, T., Tchou, C., Duiker, H.P., Sarokin, W., Sagar, M.: Acquiring the reflectance field of a human face. In: Proceedings of SIGGRAPH 2000. SIGGRAPH '00 (2000)
4. 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), 164:1–164:16 (Nov 2015)
5. Garon, M., Sunkavalli, K., Hadap, S., Carr, N., Lalonde, J.F.: Fast spatially-varying indoor lighting estimation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N.D., Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems 27, pp. 2672–2680. Curran Associates, Inc. (2014)
7. Haber, T., Fuchs, C., Bekaer, P., Seidel, H., Goesele, M., Lensch, H.P.A.: Relighting objects from image collections. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 627–634 (June 2009)
8. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. CoRR **abs/1512.03385** (2015)
9. Hold-Geoffroy, Y., Sunkavalli, K., Hadap, S., Gambaretto, E., Lalonde, J.F.: Deep outdoor illumination estimation. In: CVPR (2017), `http://vision.gel.ulaval.ca/~jflalonde/projects/deepOutdoorLight/`
10. Karsch, K., Hedau, V., Forsyth, D., Hoiem, D.: Rendering synthetic objects into legacy photographs. In: Proceedings of the 2011 SIGGRAPH Asia Conference. pp. 157:1–157:12. SA '11, ACM, New York, NY, USA (2011)
11. Kopf, J., Neubert, B., Chen, B., Cohen, M., Cohen-Or, D., Deussen, O., Uyttendaele, M., Lischinski, D.: Deep photo: Model-based photograph enhancement and viewing. ACM Trans. Graph. **27**(5), 116:1–116:10 (Dec 2008)
12. Laffont, P.Y., Bousseau, A., Paris, S., Durand, F., Drettakis, G.: Coherent intrinsic images from photo collections. ACM Trans. Graph. **31**(6), 202:1–202:11 (Nov 2012)
13. Lalonde, J.F., Efros, A.A., Narasimhan, S.G.: Webcam clip art: Appearance and illuminant transfer from time-lapse sequences. In: ACM SIGGRAPH Asia 2009 Papers. pp. 131:1–131:10. SIGGRAPH Asia '09, ACM, New York, NY, USA (2009)
14. Lalonde, J.F., Efros, A.A., Narasimhan, S.G.: Estimating the natural illumination conditions from a single outdoor image. International Journal of Computer Vision **98**(2), 123–145 (Jun 2012)
15. Li, Z., Snavely, N.: Learning intrinsic image decomposition from watching the world. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 9039–9048 (2018)
16. Li, Z., Snavely, N.: Megadepth: Learning single-view depth prediction from internet photos. In: Computer Vision and Pattern Recognition (CVPR) (2018)

17. Loscos, C., Frasson, M.C., Drettakis, G., Walter, B., Granier, X., Poulin, P.: Interactive virtual relighting and remodeling of real scenes. In: Lischinski, D., Larson, G.W. (eds.) Rendering Techniques' 99. pp. 329–340. Springer Vienna, Vienna (1999)
18. Luan, F., Paris, S., Shechtman, E., Bala, K.: Deep photo style transfer. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 6997–7005 (July 2017)
19. Mao, X., Li, Q., Xie, H., Lau, R.Y., Wang, Z., Paul Smolley, S.: Least squares generative adversarial networks. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2794–2802 (2017)
20. Martin-Brualla, R., Pandey, R., Yang, S., Pidlypenskyi, P., Taylor, J., Valentin, J., Khamis, S., Davidson, P., Tkach, A., Lincoln, P., Kowdle, A., Rhemann, C., Goldman, D.B., Keskin, C., Seitz, S., Izadi, S., Fanello, S.: Lookingood: Enhancing performance capture with real-time neural re-rendering. ACM Trans. Graph. **37**(6), 255:1–255:14 (Dec 2018)
21. Masselus, V., Peers, P., Dutré, P., Willems, Y.D.: Relighting with 4d incident light fields. ACM Trans. Graph. **22**(3), 613–620 (Jul 2003)
22. Meka, A., Fox, G., Zollhöfer, M., Richardt, C., Theobalt, C.: Live user-guided intrinsic video for static scene. IEEE Transactions on Visualization and Computer Graphics **23**(11) (NOVEMBER 2017)
23. Meka, A., Häne, C., Pandey, R., Zollhöfer, M., Fanello, S., Fyffe, G., Kowdle, A., Yu, X., Busch, J., Dourgarian, J., Denny, P., Bouaziz, S., Lincoln, P., Whalen, M., Harvey, G., Taylor, J., Izadi, S., Tagliasacchi, A., Debevec, P., Theobalt, C., Valentin, J., Rhemann, C.: Deep reflectance fields: High-quality facial reflectance field inference from color gradient illumination. ACM Trans. Graph. **38**(4), 77:1–77:12 (Jul 2019)
24. Meka, A., Maximov, M., Zollhoefer, M., Chatterjee, A., Seidel, H.P., Richardt, C., Theobalt, C.: LIME: Live intrinsic material estimation. In: Proceedings of Computer Vision and Pattern Recognition (CVPR) (June 2018)
25. Meshry, M.M., Goldman, D.B., Khamis, S., Hoppe, H., Pandey, R.K., Snavely, N., Brualla, R.M.: Neural rerendering in the wild. In: Computer Vision and Pattern Recognition (CVPR) (2019)
26. Mirza, M., Osindero, S.: Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784 (2014)
27. Nam, S., Ma, C., Chai, M., Brendel, W., Xu, N., Joo Kim, S.J.: End-to-end time-lapse video synthesis from a single outdoor image. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1409–1418 (June 2019). https://doi.org/10.1109/CVPR.2019.00150
28. Okabe, M., Zeng, G., Matsushita, Y., Igarashi, T., Quan, L., yeung Shum, H.: Single-view relighting with normal map painting (2006)
29. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019)
30. Patow, G., Pueyo, X.: A survey of inverse rendering problems. Computer Graphics Forum **22**(4), 663–687 (2003)
31. Philip, J., Gharbi, M., Zhou, T., Efros, A.A., Drettakis, G.: Multi-view relighting using a geometry-aware network. ACM Trans. Graph. **38**(4), 78:1–78:14 (Jul 2019)
32. Ramamoorthi, R., Hanrahan, P.: An efficient representation for irradiance environment maps. In: Proc. SIGGRAPH. pp. 497–500. ACM (2001)

33. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) Proceedings of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI). pp. 234–241 (2015)
34. Sengupta, S., Gu, J., Kim, K., Liu, G., Jacobs, D.W., Kautz, J.: Neural inverse rendering of an indoor scene from a single image. CoRR **abs/1901.02453** (2019)
35. 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)
36. Shan, Q., Adams, R., Curless, B., Furukawa, Y., Seitz, S.M.: The visual turing test for scene reconstruction. In: Proceedings of the 2013 International Conference on 3D Vision. pp. 25–32. 3DV '13, IEEE Computer Society, Washington, DC, USA (2013)
37. 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), 200:1–200:11 (Nov 2013)
38. Shu, Z., Yumer, E., Hadap, S., Sunkavalli, K., Shechtman, E., Samaras, D.: Neural face editing with intrinsic image disentangling. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 5541–5550 (2017)
39. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
40. Stumpfel, J., Jones, A., Wenger, A., Tchou, C., Hawkins, T., Debevec, P.: Direct hdr capture of the sun and sky. In: ACM SIGGRAPH 2006 Courses. SIGGRAPH '06, ACM, New York, NY, USA (2006)
41. Sunkavalli, K., Matusik, W., Pfister, H., Rusinkiewicz, S.: Factored time-lapse video. In: ACM SIGGRAPH 2007 Papers. SIGGRAPH '07, ACM, New York, NY, USA (2007)
42. Tchou, C., Stumpfel, J., Einarsson, P., Fajardo, M., Debevec, P.: Unlighting the parthenon. In: ACM SIGGRAPH 2004 Sketches. SIGGRAPH '04, ACM, New York, NY, USA (2004)
43. Thies, J., Zollhöfer, M., Nießner, M.: Deferred neural rendering: Image synthesis using neural textures. ACM Transactions on Graphics 2019 (TOG) (2019)
44. Troccoli, A., Allen, P.: Building illumination coherent 3d models of large-scale outdoor scenes. International Journal of Computer Vision **78**(2), 261–280 (Jul 2008)
45. Wenger, A., Gardner, A., Tchou, C., Unger, J., Hawkins, T., Debevec, P.: Performance relighting and reflectance transformation with time-multiplexed illumination. ACM Trans. Graph. **24**(3), 756–764 (2005)
46. Xing, G., Zhou, X., Peng, Q., Liu, Y., Qin, X.: Lighting simulation of augmented outdoor scene based on a legacy photograph. Computer Graphics Forum **32**(7), 101–110 (2013)
47. Xu, Z., Sunkavalli, K., Hadap, S., Ramamoorthi, R.: Deep image-based relighting from optimal sparse samples. ACM Trans. Graph. **37**(4), 126:1–126:13 (Jul 2018)
48. Yu, Y., Smith, W.A.P.: InverseRenderNet: Learning single image inverse rendering. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2019)
49. Yu, Y., Debevec, P., Malik, J., Hawkins, T.: Inverse global illumination: recovering reflectance models of real scenes from photographs. In: Proc. SIGGRAPH. pp. 215–224 (1999). https://doi.org/10.1145/311535.311559

50. Yu, Y., Malik, J.: Recovering photometric properties of architectural scenes from photographs. In: Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques. pp. 207–217. SIGGRAPH '98, ACM, New York, NY, USA (1998)
51. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). pp. 2881–2890 (2017)
52. Zhou, H., Hadap, S., Sunkavalli, K., Jacobs, D.W.: Deep single portrait image relighting. In: International Conference on Computer Vision (ICCV) (2019)