Complementary Intrinsics from Neural Radiance Fields and CNNs for Outdoor Scene Relighting

Siqi Yang[1,2,3#] Xuanning Cui[1,2,4#] Yongjie Zhu[5] Jiajun Tang[1,2] Si Li[5] Zhaofei Yu[3] Boxin Shi[1,2,3,4*]

[1] National Key Laboratory for Multimedia Information Processing, School of Computer Science, Peking University
[2] National Engineering Research Center of Visual Technology, School of Computer Science, Peking University
[3] Institute for Artificial Intelligence, Peking University
[4] AI Innovation Center, School of Computer Science, Peking University
[5] School of Artificial Intelligence, Beijing University of Posts and Telecommunications
{yousiki, cxn, jiajun.tang, yuzf12, shiboxin}@pku.edu.cn
yongjie.zhu.96@gmail.com lisi@bupt.edu.cn

Abstract

Relighting an outdoor scene is challenging due to the diverse illuminations and salient cast shadows. Intrinsic image decomposition on outdoor photo collections could partly solve this problem by weakly supervised labels with albedo and normal consistency from multi-view stereo. With neural radiance fields (NeRF), editing the appearance code could produce more realistic results without interpreting the outdoor scene image formation explicitly. This paper proposes to complement the intrinsic estimation from volume rendering using NeRF and from inversing the photometric image formation model using convolutional neural networks (CNNs). The former produces richer and more reliable pseudo labels (cast shadows and sky appearances in addition to albedo and normal) for training the latter to predict interpretable and editable lighting parameters via a single-image prediction pipeline. We demonstrate the advantages of our method for both intrinsic image decomposition and relighting for various real outdoor scenes.

1. Introduction

The same landmark may appear with drastically varying appearances in different photos, even if they are taken from the same viewpoint with the same camera parameters, *e.g.*, the Taj Mahal may look golden or white at sunset or in the afternoon[1]. For a set of photos containing the same landmark captured in different seasons and times, their "dynamic" lighting changes (compared with the relatively "stable" geometry and reflectance) play a vital role in explaining such great appearance variations. If we can independently manipulate lighting in these photos, the relighted outdoor scenes could substantially improve experiences for taking digital photographs.

Outdoor scene relighting could be realized by learning a style transfer procedure [1, 5]. Such a process only requires a single reference image for editing a target image, but it apparently retouches the image to "look like" each other, without explicitly modeling the lighting changes. Intrinsic image decomposition, which inversely decomposes the photometric image formation model, has been extended to work with outdoor photo collections [32–34]. The common geometry/reflectance (for the whole collection) and distinctive lighting components (for each image) are estimated using deep convolutional neural networks (CNNs), so that relighting could be achieved by keeping the former while editing the latter in a physics-aware manner. These methods explicitly conduct computationally expensive multi-view reconstruction of the scene at the training stage. The weakly supervised constraints built upon albedo and normal consistency via multi-view correspondence cannot support the handling of cast shadows [33] or still struggle with strong cast shadows [32].

Recently, the emerging of neural radiance fields (NeRF) [23] has not only boosted the performance of novel view synthesis with significantly better quality for outdoor scene photo collections [22], but has also been demonstrated to be capable of transferring the lighting appearance across the image set using hallucination [4] or a parametric lighting model [28]. However, these existing NeRF methods in the outdoor scene either miss the explanation to some important intrinsics for relighting such as cast shadows (except for [28]) or ignore distinctive characteristics between the non-sky and sky regions, in a physically interpretable manner.

In this paper, we hope to conduct outdoor scene relighting by mutually complementing intrinsics estimated from NeRF and CNN and taking advantages of the comprehensive representation power of NeRF and physics interpretability of CNN-based single-image decomposition, in a *single-image* inference pipeline as shown in Figure 1. We formulate the color formation of a pixel as a combination of objects and the sky by tracing rays from camera origins to the furthest plane and accumulating the voxel intrin-

[#] Contributed equally to this work as first authors
[*] Corresponding author
[1] Changing colours of Taj Mahal: agratajcitytour.com

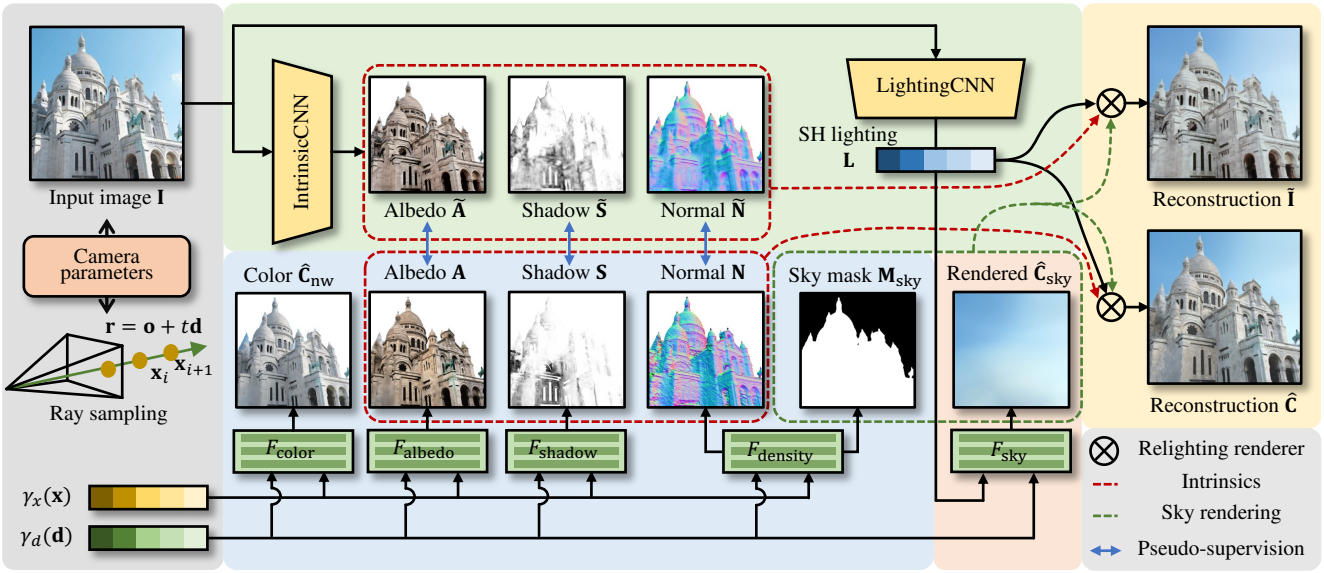

Figure 1. Illustration of our overall pipeline. For each image, the corresponding camera rays and sampled 3D points are positionally-encoded for MLPs (gray block). Our IntrinsicCNN and LightingCNN modules derive lighting-(in)dependent intrinsic components and second-order spherical harmonics (SH) lighting coefficients (green block), while Our NeRF module provides the pseudo labels of intrinsics and sky mask by volumetric rendering through MLPs (blue block). Our SkyMLP module renders the sky with viewing direction and extracted lighting (orange block). Given lighting extracted from the input/reference image, the reconstructed/relighted image is rendered by photometric image formation along with the rendered sky (yellow block).

sics. We then propose a modified NeRF system to estimate diffuse albedo, surface normal, cast shadow, and illumination parameters. Our NeRF rendering naturally shares the albedo and normal of one point across all images and interprets the geometry from voxel density, which provides more accurate pseudo labels for identifying shadows than purely CNN-based approaches [32–34]. We finally predict the intrinsics and lighting parameters by designing two separate CNN modules based on the NeRF-produced pseudo labels with a clearer separation of lighting-dependent and independent intrinsic components to achieve high-quality relighting via single-image prediction.

Hence, our contribution becomes clear in three folds: Based on 1) the newly proposed "object-sky" hybrid image formation, 2) the intrinsics estimated from NeRF provide more accurate pseudo labels to complement 3) the intrinsics estimated from CNNs, for conducting outdoor scene relighting using a single image in a physically interpretable manner and with a visually pleasing appearance, which is demonstrated by our experimental results.

2. Related work

Style transfer is a possible way to relight an outdoor scene, by formulating it as an image-to-image translation problem [9, 12, 17, 20, 22]. By appropriately selecting a reference photo with desired illumination and transferring its style to the content image, the target image may look like being taken under the specified illumination [20, 29]. This could be achieved by feeding illumination representation as input conditions of conditional generative adversarial networks (CGANs) via an image-to-image translation framework [12]. With cycle-consistent supervision, the lighting condition can be transferred from the reference image to the target image even in the absence of paired training examples [37]. The task of image style transfer can be also solved using CNNs [8, 9, 12, 20]. CNN-based models are first used to extract object features and texture features, the relighted image is then generated with the combination of object features from the original image and texture features from the target image [9, 20]. Although approaches in this category can achieve photo-realistic effects as relighting in some certain scale, their lacks of physical representation of lighting condition prevent them from controlling illuminated appearance in an interpretable manner. This paper pays attention to intrinsics with clear physical meaning to deal with this issue.

Intrinsic decomposition provides another solution for relighting an outdoor scene, by inversely analyzing the photometric image formation model [7, 11, 16, 24] and independently editing the lighting component, while leaving other lighting-irrelevant components (such as albedo and normal) unchanged [14, 15, 32–34]. Collections of outdoor photographs could be used to provide constraints for model training [6, 19]. Due to the difficulty of obtaining ground

truth intrinsic components in real scenes, synthetic datasets are often used for model pre-training [2, 7, 16, 25], but the large gap between synthetic and real datasets makes it still difficult to achieve photo-realistic results in real scenes. Another way is to use self-supervised or semi-supervised learning methods based on a large number of images of the same scene [32–34], so that auxiliary supervisions and constraints to improve the model performance in real scenes can be acquired by exploring the consistent geometry and reflectance constraints in the multiview stereo context. This paper aims to distinguish lighting-(in)dependent intrinsics like [32–34], but pays more attention to reliably constrain the decomposition process with more accurate pseudo labels.

Neural scene representation is applied to modify the lighting-relevant properties of the outdoor scenes since the booming of NeRF [23] and its variants [4, 21, 28, 30, 35, 36]. Benefiting from a more comprehensive way of scene representation, these methods can produce photo-realistic renderings from any viewpoint. The vanilla NeRF [23] only works on static scenes and assumes all input images share the same lighting condition, which makes it unsuitable for relighting. The fixed lighting condition is relieved in NeRV [30], which represents the scene as a continuous volumetric function of albedo, normal, and other scene properties parameterized as multi-layer perceptrons (MLPs). To further deal with the influence of transient parts during scene reconstruction, such as pedestrians and cars, semantic segmentation [27] and transient modules [4, 21] are introduced. NeRF-W [21] proposes a latent appearance modeling method, which adopts the approach of generative latent optimization [3]. The latent appearance code is learned for each input image to represent the appearance separately. However, the latent appearance model is not supported by a physics-based image formation model, so it cannot generate physically meaningful shadows or conduct controllable appearance editing.

Ha-NeRF [4] proposes a CNN-based appearance encoder to encode the appearance of each image, which is constrained by a dedicated loss. Although Ha-NeRF [4] can render images in any new lighting conditions, the relighted scenes are hallucinated and not physically interpretable, as the name says. The physically interpretable lighting is supported by NeRF-OSR [28], which uses spherical harmonic coefficients for lighting representation [26], but it can only render images in lighting conditions that are observed in the training set or extracted from given environment maps. This paper integrates the intrinsics from both NeRF- and CNN-based methods in a complementary manner to balance relighting fidelity and controllability.

3. Method

We first introduce relevant background knowledge about neural radiance fields (Section 3.1). Based on which, we propose the complementary intrinsic estimations using both NeRF (Section 3.2) and CNNs (Section 3.3) for outdoor scene relighting, whose joint interaction enables single-image inference. The illustration of our overall pipeline is shown in Figure 1.

3.1. Preliminaries

We briefly review the formulation of neural radiance fields, especially its application in outdoor scenes, to make this paper self-contained. NeRF [23] uses volumetric functions to represent the scene, which is modeled by two multilayer perceptrons (MLPs). The first MLP F_{density} models the density function, which takes the 3D location of point $\mathbf{x}$ as input and estimates the volume density σ as

$$(\sigma(\mathbf{x}), \mathbf{z}) = F_{\text{density}}(\gamma_x(\mathbf{x})), \tag{1}$$

where F stands for multilayer perception[2] throughout this paper, γ_x is the positional encoding function for 3D locations, and $\mathbf{z}$ is the latent features used for color prediction.

The second MLP F_{color} models the color function, which depends on the latent features $\mathbf{z}$ together with the viewing direction $\mathbf{d}$. For outdoor scenes in a landmark photo collection, the camera parameters and environment lightings are usually different for each image. Thus, a learnable appearance embedding l is assigned to each image to model these per-image discrepancies [21]. Then, the color prediction via NeRF for outdoor images in the wild c_{nw} is formulated as:

$$c_{\text{nw}}(\mathbf{x}, \mathbf{d}, l) = F_{\text{color}}(\gamma_d(\mathbf{d}), \mathbf{z}, l), \tag{2}$$

where γ_d is the positional encoding function for viewing diretions.

NeRF independently renders each ray $\mathbf{r} \in \mathcal{R}$ casted from the camera origin $\mathbf{o}$ in direction $\mathbf{d}$ corresponding to each pixel. For the ray $\mathbf{r} = \mathbf{o} + t\mathbf{d}$, N points in total are sampled from the nearest plane to the furthest plane, and then the accumulated color of the ray is approximated as:

$$\begin{aligned} \hat{\mathbf{C}}_{\text{nw}}(\mathbf{r}) &= \sum_{i=1}^{N} T_i (1 - \exp(-\sigma_i \delta_i)) c_{\text{nw}}(\mathbf{o} + t_i \mathbf{d}, \mathbf{d}, l), \\ T_i &= \exp(-\sum_{j=1}^{i-1} \sigma_j \delta_j), \end{aligned} \tag{3}$$

where $\delta_i = t_{i+1} - t_i$ is the distance between two adjacent sampling points.

The optimization goal of NeRF is simply minimizing the total squared error between the rendered images and the ground truth [23]. However, in-the-wild images usually contain moving objects, such as tourists and vehicles. This could be dealt with by introducing uncertainty of transient objects [21]. A recent study suggests that such an approach may lead to ghostly figures and blurry results [27], and a pre-trained semantic segmentation model [31] could

[2] Later, we will use G for CNNs.

be used to mark out these regions using a mask for moving object $\mathbf{M}_{\text{mov}}$ and eliminate the effects of transient objects during training:

$$\mathcal{L}_{\text{nw}} = \sum_{\mathbf{r}\in\mathcal{R}} (1 - \mathbf{M}_{\text{mov}}(\mathbf{r})) \|\hat{\mathbf{C}}_{\text{nw}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|_2^2. \quad (4)$$

3.2. Intrinsic estimation using NeRF

By optimizing Equation (4), relighting features could be obtained by interpolating between appearance embeddings from two images via NeRF [4, 21], but the relighted scenes in this way cannot maintain semantic meanings and the lighting is not physically editable. In this paper, we hope to conduct interpretable single-image relighting for outdoor scenes, and we achieve this by estimating the intrinsic components of the scenes and rendering pixels that satisfy the physics-based photometric image formation model. We formulate the formation of the color of pixels as NeRF-OSR [28], *i.e.*, by using the combination of diffuse albedo $\mathbf{A}$, shadow $\mathbf{S}$, and shading, where the shading is derived from the surface normal $\mathbf{N}$ and spherical harmonics lighting $\mathbf{L}$ under the Lambertian model. Thus, we could approximate the appearance color along the ray $\mathbf{r}$ as:

$$\hat{\mathbf{C}}_{\text{obj}}(\mathbf{r}) = \mathbf{A}(\mathbf{r}) \odot \mathbf{S}(\mathbf{r}, \mathbf{L}) \odot \mathbf{Lb}(\mathbf{N}(\mathbf{r})), \quad (5)$$

where $\odot$ is Hadamard product, $\mathbf{L} \in \mathbb{R}^{3\times 9}$ is the second-order spherical harmonics coefficients for each color channel, $\mathbf{b}(\mathbf{n}) \in \mathbb{R}^9$ is the second-order spherical harmonics basis corresponding to normal vector $\mathbf{n}$. Albedo $\mathbf{A}$ and surface normal $\mathbf{N}$ are lighting-independent intrinsics and the others are lighting-dependent.

We modify the NeRF system introduced in Section 3.1 to estimate the outdoor scene intrinsics in Equation (5), as shown in the blue block of Figure 1. Diffuse albedo $\mathbf{A}(\mathbf{r})$, surface normal $\mathbf{N}(\mathbf{r})$, and shadow $\mathbf{S}(\mathbf{r})$ are accumulated from corresponding terms $\alpha(\mathbf{x})$, $n(\mathbf{x})$, and $s(\mathbf{x})$, in a similar way as Equation (3). The shading term is derived from $\mathbf{N}(\mathbf{r})$ at the pixel level instead of the voxel level. To calculate intrinsic components at each 3D location $\mathbf{x}$, the normal term $n(\mathbf{x})$ at point $\mathbf{x}$ can be directly derived from the density function $\sigma(\mathbf{x})$, by calculating the normalized negative derivative of σ with respect to $\mathbf{x}$:

$$n(\mathbf{x}) = -\frac{\partial\sigma(\mathbf{x})}{\partial\mathbf{x}} / \|\frac{\partial\sigma(\mathbf{x})}{\partial\mathbf{x}}\|_2. \quad (6)$$

Diffuse albedo is very similar to the color function of NeRF, except that the albedo is a direction-irrelevant term and it does not depend on the viewing direction $\mathbf{d}$ of each ray. Thus, we use a new MLP F_{albedo} which takes only the latent features $\mathbf{z}$ as input to calculate the diffuse albedo term $\alpha(\mathbf{x})$ at point $\mathbf{x}$ as:

$$\alpha(\mathbf{x}) = F_{\text{albedo}}(\mathbf{z}). \quad (7)$$

Predicting the cast shadow accurately is important for realistically relighting an outdoor scene. This could be achieved by using another MLP F_{shadow}. The cast shadow term $s(\mathbf{x})$ at each point $\mathbf{x}$ depends on the direction of lighting and the geometry of outdoor scenes, so we calculate it as

$$s(\mathbf{x}) = F_{\text{shadow}}(\gamma_x(\mathbf{x}), \mathbf{L}). \quad (8)$$

We use fewer fully-connected layers and neurons in F_{shadow} than other MLPs, to ensure the continuity and low frequency of estimated shadow.

The image formation in Equation (5) is only valid for opaque objects, which means it does not apply to the sky – an indispensable factor for modeling the outdoor scene appearance. To overcome this issue, a separate sky modeling is proposed to render the color of sky pixels, as shown in the orange block of Figure 1.

For the sky region, we formulate the color of the ray $\mathbf{r} = \mathbf{o} + t\mathbf{d}$ as a function F_{sky} that only depends on its viewing direction $\mathbf{d}$ and the global lighting representation $\mathbf{L}$. We then design a new MLP module to represent F_{sky}, denoted as *SkyMLP*:

$$\hat{\mathbf{C}}_{\text{sky}}(\mathbf{d}, \mathbf{L}) = F_{\text{sky}}(\gamma_d(\mathbf{d}), \mathbf{L}). \quad (9)$$

It is vital that these two variables are represented in the same coordinate system, no matter whether it is the camera or the world coordinate system. The shared coordinate system enables us to render the sky region during single-image inference, as the directions of rays cast in the camera coordinate system are computable. Moreover, the proposed SkyMLP module together with CNN modules (which will be introduced in Section 3.3) provides us a way to render the outdoor sky from a given reference image.

By merging the terms for objects $\hat{\mathbf{C}}_{\text{obj}}$ and for the sky $\hat{\mathbf{C}}_{\text{sky}}(\mathbf{r})$ together, we can predict color of ray $\mathbf{r}$ with our NeRF module as:

$$\hat{\mathbf{C}}(\mathbf{r}) = \mathbf{M}_{\text{obj}}(\mathbf{r})\hat{\mathbf{C}}_{\text{obj}}(\mathbf{r}) + \mathbf{M}_{\text{sky}}(\mathbf{r})\hat{\mathbf{C}}_{\text{sky}}(\mathbf{r}). \quad (10)$$

To bridge these two terms seamlessly, we need something like the semantic segmentation mask for the sky region. Actually, $T_N = \exp(-\sum_{j=1}^{N} \sigma_j\delta_j)$ (recall that N is the total number of sampling points in a ray) denotes the accumulated transmittance along the ray from the nearest plane to the furthest plane, which is the probability that the ray hitting nothing but the sky which is infinitely far away. Then the masks for the object and sky terms can be calculated as:

$$\mathbf{M}_{\text{obj}} = 1 - T_N \quad \text{and} \quad \mathbf{M}_{\text{sky}} = T_N. \quad (11)$$

Ideally, both $\hat{\mathbf{C}}_{\text{obj}}$ and $\hat{\mathbf{C}}_{\text{sky}}$ can be simply supervised by the total squared error between rendered images and ground truth, which can be formulated as:

$$\mathcal{L}_{\text{nerf}} = \sum_{\mathbf{r}\in\mathcal{R}} (1 - \mathbf{M}_{\text{mov}}(\mathbf{r})) \|\hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|_2^2. \quad (12)$$

However, without additional constraints, NeRF sometimes bakes $\hat{\mathbf{C}}_{\text{obj}}$ into $\hat{\mathbf{C}}_{\text{sky}}$, and T_N falls to 0 for all rays at the early stage of training, which leads to the collapse for

SkyMLP. To overcome this issue, we randomly add gaussian noise to T_N and append the total squared error between rendered sky and image to the optimization objectives:

$$\mathcal{L}_{\text{sky}} = \lambda_{\text{sky}} \|\hat{\mathbf{C}}_{\text{sky}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})\|_2^2. \tag{13}$$

This won't bake the color of the foreground objects into SkyMLP because SkyMLP is shared by multiple scenes in our dataset and the number of layers and neurons is especially limited. Similarly as conducted by Rematas *et al*. [27], we apply an additional loss at the early stage of training to decrease the density of sky region:

$$\mathcal{L}_{\text{seg}} = \lambda_{\text{seg}} \sum_{\mathbf{r} \in \mathcal{R}_{\text{sky}}} \sum_{i=1}^{N} \sigma(\mathbf{o} + t_i \mathbf{d}), \tag{14}$$

where $\mathcal{R}_{\text{sky}}$ stands for rays of the sky region and the sky mask is generated by pre-trained semantic segmentation model.

While using the rendering procedure described above is sufficient to reconstruct the outdoor scenes, combining it with the one described in Equation (3) and sharing the density function σ across them leads to more precise geometry reconstruction and better intrinsics estimation. Thus, the overall loss function for NeRF training state is:

$$\mathcal{L} = \mathcal{L}_{\text{nw}} + \mathcal{L}_{\text{nerf}} + \mathcal{L}_{\text{sky}} + \mathcal{L}_{\text{seg}}. \tag{15}$$

3.3. Intrinsic estimation using CNN

Using the estimated intrinsics from NeRF in Section 3.2, we can perform outdoor scene relighting using the pre-trained model of the specific scene, *i.e*., replace the lighting $\mathbf{L}$ with another one $\mathbf{L}'$ and leave other terms unchanged in Equation (5) and Equation (9). However, single-image relighting at inference time is still impossible because the lighting and camera extrinsic parameters are unknown. And the recovered noisy geometry will lead to unpleasant rendering results. Therefore, we further introduce two CNN modules to address these problems, as shown in the green block of Figure 1.

We use the first CNN $G_{\text{intrinsic}}$ to extract key intrinsic components corresponding to Equation (5), including diffuse albedo, surface normal, and cast shadow, and we call it *IntrinsicCNN*. It adopts a U-Net-like architecture, which consists of 9 residual blocks and 30 convolution layers in total, and takes the images as input and predicts each intrinsic components at the pixel level:

$$(\widetilde{\mathbf{A}}, \widetilde{\mathbf{S}}, \widetilde{\mathbf{N}}) = G_{\text{intrinsic}}(\mathbf{I}). \tag{16}$$

Here, we use $\widetilde{\cdot}$ to distinguish these components from the previous ones estimated by NeRF.

We design the second CNN G_{lighting} to estimate the lighting from a single image, which is called *LightingCNN*. It consists of 4 convolution layers and takes an image as input and predicts its lighting represented by second-order spherical harmonics, whose coefficients are $\widetilde{\mathbf{L}} \in \mathbb{R}^{3\times 9}$:

$$\widetilde{\mathbf{L}} = G_{\text{lighting}}(\mathbf{I}). \tag{17}$$

As shown in the green and orange blocks of Figure 1, to re-render the image, $\widetilde{\mathbf{L}}$ is fed back to the SkyMLP for reconstructing the sky background, and then the foreground color is composed with the sky using the predicted sky mask $\mathbf{M}_{\text{sky}}$:

$$\begin{aligned} \widetilde{\mathbf{I}} = (1 - \mathbf{M}_{\text{sky}})(\widetilde{\mathbf{A}}\widetilde{\mathbf{S}} \odot \widetilde{\mathbf{L}}\mathbf{b}(\widetilde{\mathbf{N}})) \\ + \mathbf{M}_{\text{sky}} \hat{\mathbf{C}}_{\text{sky}}(\gamma_d(\mathbf{d}), \widetilde{\mathbf{L}}). \end{aligned} \tag{18}$$

IntrinsicCNN obtains pseudo-supervision from pre-trained NeRF, as shown by the blue double arrows between the green and blue blocks of Figure 1. We train the CNN by minimizing the total squared error between albedo, normal, and shadow maps, together with the reconstruction loss:

$$\begin{aligned} \mathcal{L}_{\text{cnn}} = (1 - \mathbf{M}_{\text{mov}}) \|\widetilde{\mathbf{A}} - \hat{\mathbf{A}}\|_2^2 \\ + (1 - \mathbf{M}_{\text{mov}}) \|\widetilde{\mathbf{S}} - \hat{\mathbf{S}}\|_2^2 \\ + (1 - \mathbf{M}_{\text{mov}}) \|\widetilde{\mathbf{N}} - \hat{\mathbf{N}}\|_2^2 \\ + \lambda_{\text{recon}} \|\widetilde{\mathbf{I}} - \mathbf{I}\|_2^2. \end{aligned} \tag{19}$$

Once the intrinsic components are known, we can perform lighting editing by replacing $\widetilde{\mathbf{L}}$ of the target image with new lighting parameters to produce the re-rendered result.

4. Experiment

We use landmark image collections from the Phototourism dataset [13] and MegaDepth dataset [18] to evaluate the performance of our method. Since there is no ground truth for each estimated intrinsic component in real data, we mainly evaluate these tasks qualitatively. For quantitative evaluation, we follow [32] to measure the consistency between the input image and the reconstructed counterpart from intrinsic components. We demonstrate the benefits of our design in the ablation study.

4.1. Data preparation

Unlike NeRF-W [21] and Ha-NeRF [4], our method requires an existing semantic segmentation algorithm to exclude the transient objects from training. Because handling the dynamic occlusion is not the main task of this paper, we experimentally find that using segmentation masks makes the training procedure of our NeRF module much more stable, also as suggested in [27]. We use SegFormer [31] to generate high-quality semantic labels of images. All tourists and vehicles are excluded from training and qualitative evaluation. We further clean our dataset by manually discarding photos with obviously bad quality, such as the ones with strong over-exposure, shaking, *etc*. Finally, we randomly pick up 1000 images from about 1300 images of each scene to form the training set and leave the others as the test set. Our dataset contains 12 landmark scenes in total, among which 2 scenes are used for evaluation only.

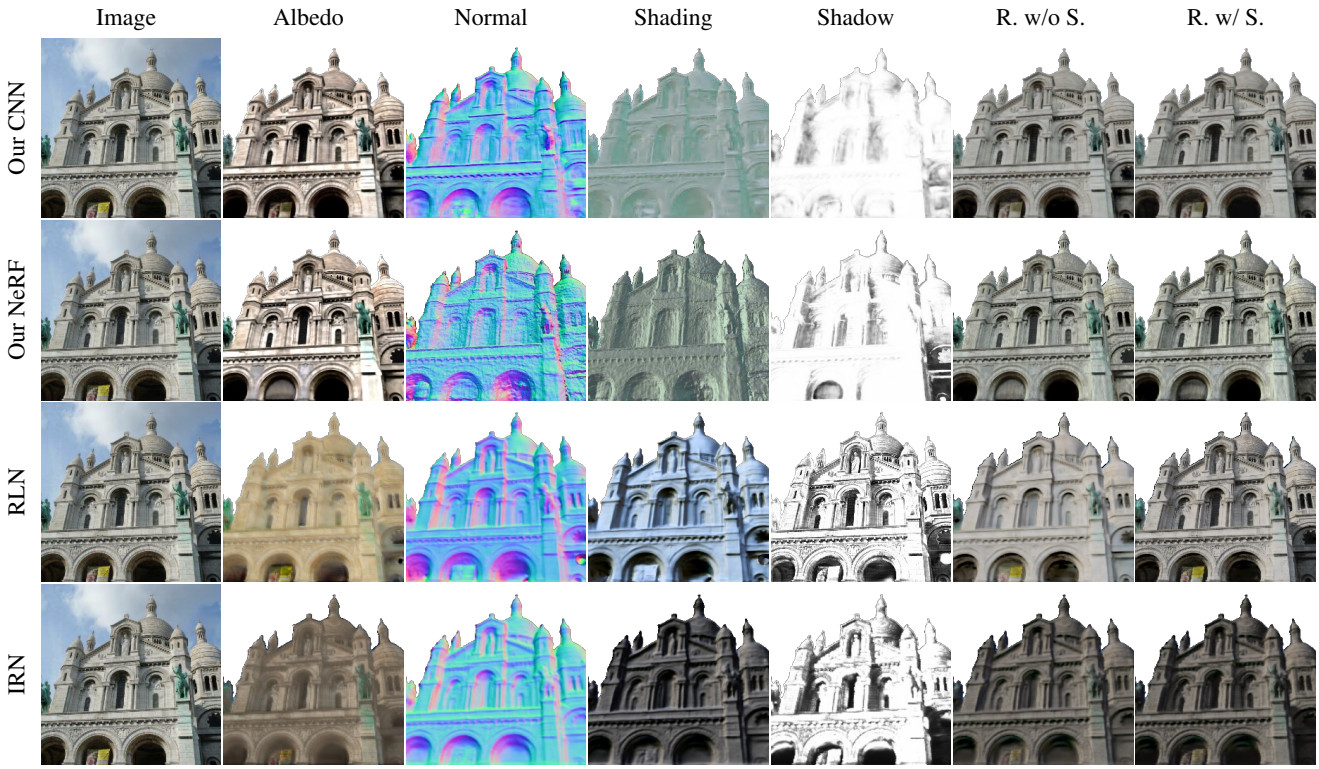

Figure 2. Estimated intrinsics (albedo, normal, shading, and shadow) and reconstruction results without and with shadow (R. w/o S. and R. w/ S.) from our CNN (Section 3.3) module, our NeRF (Section 3.2) module, RelightingNet (RLN) [32], and InverseRenderNet (IRN) [34].

4.2. Implementation details

Our training procedure can be divided into two stages, corresponding to our NeRF and CNN module. LightingCNN is jointly trained with our NeRF and SkyMLP modules at the first stage. To accelerate this training procedure, we use the learnable lighting code $\mathbf{L}$ assigned to each image, together with $\widetilde{\mathbf{L}}$ estimated by LightingCNN, to complete the subsequent pipeline. The duplicate results are both supervised in the same way as Equation (15). The involvement of $\mathbf{L}$ contributes to the converging of geometry.

To relight an outdoor scene using a single image, we need to first decompose the input image into intrinsic components using Equation (16), re-render the foreground objects by Equation (5), generate an appropriate sky image as the background by Equation (9), and finally combine these together according to Equation (10). The direction $\mathbf{d}$ of each ray and the lighting representation $\mathbf{L}$ are both aligned to the camera coordinate system, and are generated by camera intrinsic parameters and our LightingCNN respectively.

As for the hyperparameters, we select $\lambda_{\text{sky}} = 0.1$ and $\lambda_{\text{recon}} = 0.5$ during training, and gradually decrease λ_{seg} from 0.2 to 0.001. Please refer to our supplementary material for more implementation details.

4.3. Performance evaluation

Intrinsic decomposition. We qualitatively evaluate the intrinsic components estimated by our NeRF and CNN, and compare them with RelightingNet [32] and InverseRenderNet [34]. As shown in Figure 2, though the geometry directly derived from the density function σ has a rough surface, our IntrinsicCNN complements it by suppressing artifacts and providing smooth predictions. Our method shows stronger capability in distinguishing shape-independent shadow and shape-dependent shading than purely CNN-based methods [32, 34], while they suffer from the ambiguity of shading as well as cast shadow and bake all these factors into shadow maps due to the distinctive pseudo-supervision labels provided by our NeRF. As the ground truth of intrinsics in the wild is not available, we only evaluate the reconstruction results, as shown in Table 1.

Relighting. We evaluate the single-image relighting performance[3] of our CNN (trained with labels from our NeRF) in Figure 3 by comparing with RelighitngNet [32] in both shadow-aware (left of the target image) domain and

[3]We cannot compare with NeRF-based methods like [21] for this task, since they only relight with known lighting from multi-view images set.

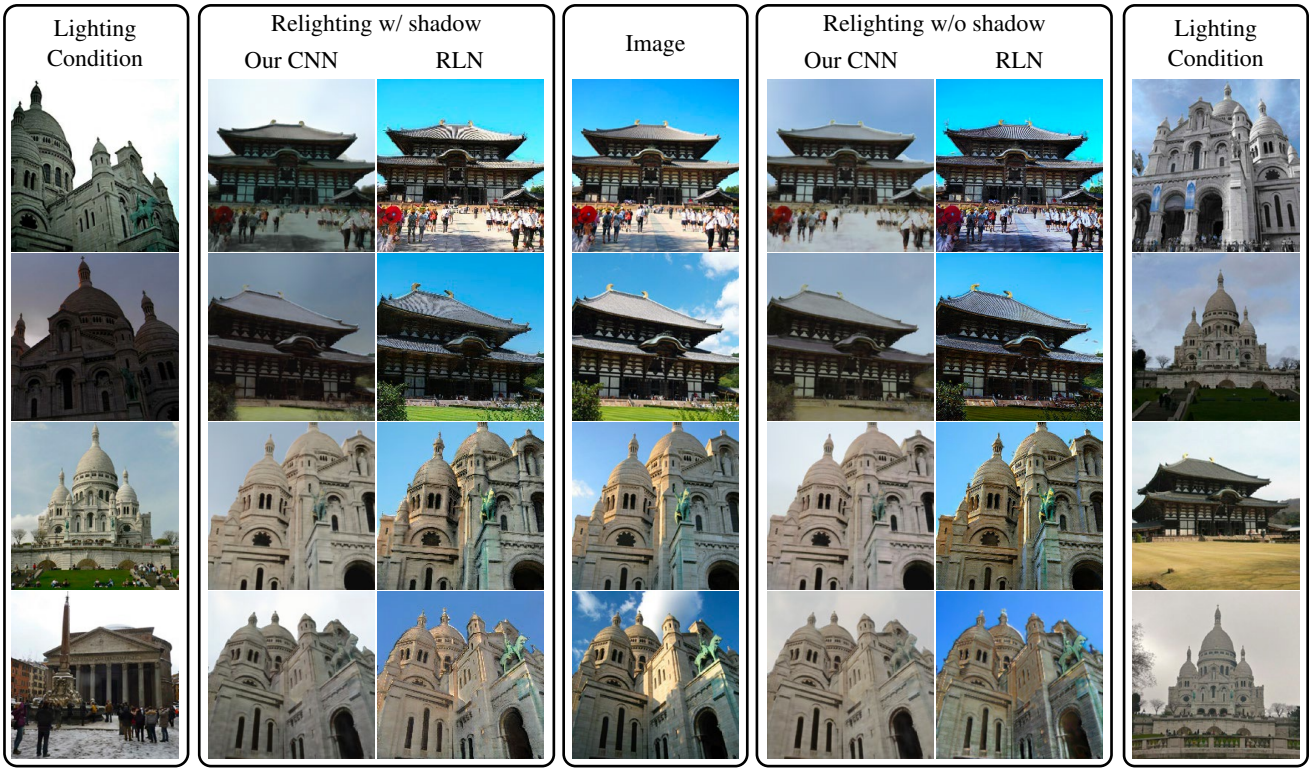

Figure 3. Relighting results (the target image in the middle rendered with lighting condition of the reference image in the leftmost and rightmost column) with cast shadows from our CNN and RelightingNet (RLN) [32].

Table 1. Comparison of reconstructed image quality. The metrics are averaged across the test set of our image collections. ↑ (↓) means higher (lower) is better.

Method	PSNR ↑	SSIM ↑	MSE ↓	MAE ↓	LPIPS ↓
Our CNN	**24.2387**	0.8558	**0.0038**	**0.0430**	0.1354
RelightingNet [32]	23.4755	**0.8595**	0.0051	0.0543	**0.1336**
InverseRenderNet [34]	17.3381	0.5999	0.0204	0.1176	0.1512

shadow-free (right of the target image) domain. Our method generates a more realistic and appropriate sky region during relighting, partly due to the effectiveness of hybrid image formation and SkyMLP, and it also prevents the incorrect appearing of highlight regions in the third and fourth row (under uniform lighting). We also correctly remove the shadow on the ground in the first row, and the shadow under the eave in the second row in the shadow-free domain.

We pretrain our CNN module using a large-scale dataset of 10 landmark scenes. The high-quality pseudo labels provided by our NeRF module effectively shorten the learning curve of CNN, hence our CNN module can relight images from both seen and unseen landmark scenes. We demonstrate such generalization capability of our method in Figure 4.

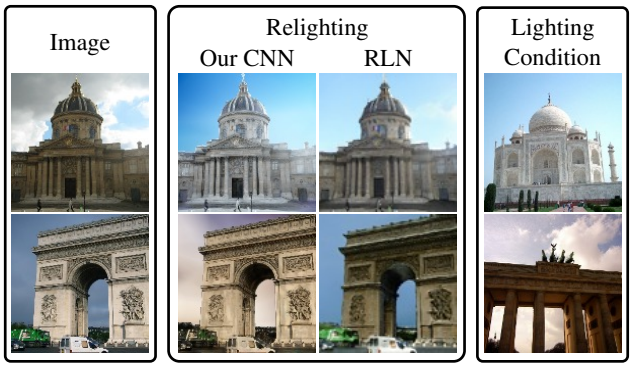

Figure 4. Relighting results (the target image in the leftmost rendered with lighting condition from the reference image in the rightmost column) from our CNN and RelightingNet (RLN) [32] on unseen scenes for validating the generalization capability.

4.4. Ablation study

Effectiveness of hybrid formulation. Although our NeRF design shares some similar structures with NeRF-OSR [28], it is non-trivial to adapt their model to our formulation due to that 1) we handle the sky region with a separate color function and 2) we combine the photometric image formulation ($\hat{\mathbf{C}}$) together with the uninterpretable one ($\hat{\mathbf{C}}_{\text{nw}}$). We use our implementation of NeRF-OSR [28] and

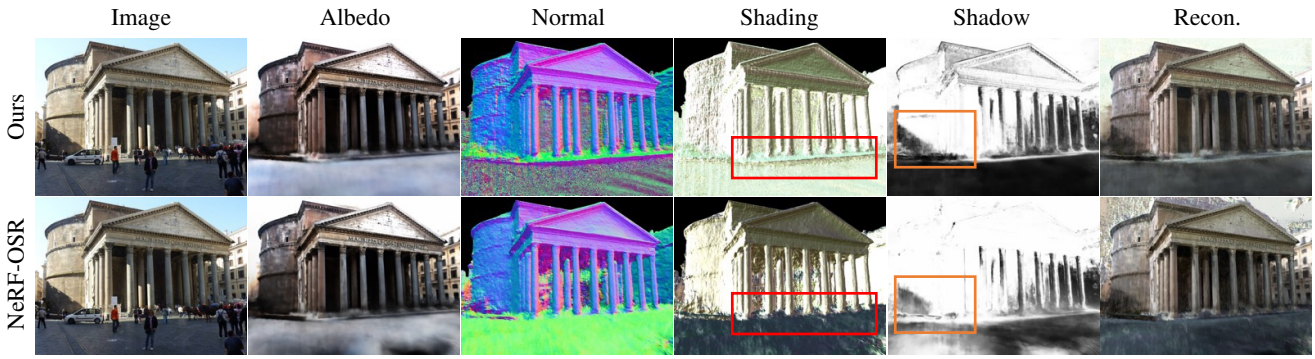

Figure 5. Estimated intrinsic components and reconstruction results (recon.) of our NeRF and NeRF-OSR [28].

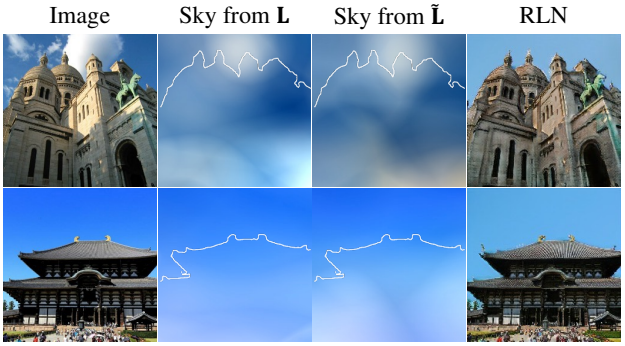

Figure 6. Comparison of the sky reconstructed by our method and RelightingNet (RLN) [32]. $\mathbf{L}$ and $\tilde{\mathbf{L}}$ are lighting parameters learned by our NeRF and the LightingCNN module, respectively. The white contour on each sky image is generated from the sky mask to indicate the sky region (above the contour) for better visualization.

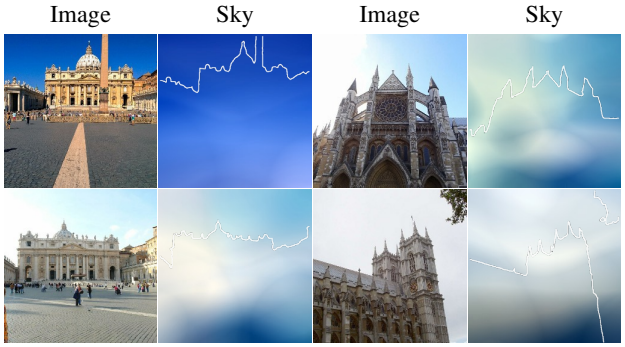

Figure 7. Sky estimated and rendered from unseen outdoor scenes in MegaDepth dataset [18].

compare with their intrinsic estimation performances. As shown in Figure 5, our method produces better shadow and geometry (disparity). The foundation of building is recovered with $\hat{\mathbf{C}}_{\text{nw}}$, but infused into the ground without $\hat{\mathbf{C}}_{\text{nw}}$.

Effectiveness of SkyMLP. We evaluate the performance of our sky generation (using LightingCNN together with SkyMLP), and compare it with RelightingNet [32]. As shown in Figure 6, Our SkyMLP can render a more realistic sky. As shown in Figure 7, with our LightingCNN capturing the lighting coefficients from images that are not included in our NeRF training set, our SkyMLP can also retain the major characteristic of the sky, *i.e.*, the primary color and relative brightness.

5. Conclusion

We present a novel approach for outdoor scene relighting by complementing the intrinsics from both neural radiance fields and convolutional neural networks. We propose a new image formation model for NeRF volume rendering, which handles static scenes and sky separately. Our method recovers accurate intrinsic components from NeRF, and then these pseudo labels enable our CNNs to estimate intrinsics from a single image. Outdoor scene relighting is conducted by editing the lighting coefficients in the physics-aware rendering procedure. Our method is completely driven by real-world data and demonstrates noticeable improvements over previous methods.

Limitations. In this paper, we either use the unchanged cast shadow or evaluate in the shadow-free domain (Figure 3). This is because predicting highly accurate and physics-aware cast shadow boundaries from a single image is a rather challenging problem, as it requires a precise understanding of the entire scene geometry and sunlight direction. In our future work, we hope to resolve this issue by integrating with user-interaction or graphics techniques [10].

Acknowledgement

This work was supported by National Natural Science Foundation of China (Grant No. 62136001, 62088102). The AI training platform supporting this work was provided by High-Flyer AI (Hangzhou High-Flyer AI Fundamental Research Co., Ltd.).

References

[1] Ivan Anokhin, Pavel Solovev, Denis Korzhenkov, Alexey Kharlamov, Taras Khakhulin, Aleksei Silvestrov, Sergey Nikolenko, Victor Lempitsky, and Gleb Sterkin. High-resolution daytime translation without domain labels. In *CVPR*, 2020. 1

[2] Sai Bi, Nima Khademi Kalantari, and Ravi Ramamoorthi. Deep hybrid real and synthetic training for intrinsic decomposition. 2018. 3

[3] Piotr Bojanowski, Armand Joulin, David Lopez-Paz, and Arthur Szlam. Optimizing the latent space of generative networks. 2018. 3

[4] Xingyu Chen, Qi Zhang, Xiaoyu Li, Yue Chen, Feng Ying, Xuan Wang, and Jue Wang. Hallucinated neural radiance fields in the wild. *arXiv preprint arXiv:2111.15246*, 2021. 1, 3, 4, 5

[5] Chia-Chi Cheng, Hung-Yu Chen, and Wei-Chen Chiu. Time flies: Animating a still image with time-lapse video as reference. In *CVPR*, 2020. 1

[6] Sylvain Duchêne, Clement Riant, Gaurav Chaurasia, Jorge Lopez-Moreno, Pierre-Yves Laffont, Stefan Popov, Adrien Bousseau, and George Drettakis. Multi-view intrinsic images of outdoors scenes with an application to relighting. *ACM TOG*, 2015. 2

[7] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. Revisiting deep intrinsic image decompositions. In *CVPR*, 2018. 2, 3

[8] Jacob R Gardner, Paul Upchurch, Matt J Kusner, Yixuan Li, Kilian Q Weinberger, Kavita Bala, and John E Hopcroft. Deep manifold traversal: Changing labels with convolutional features. *arXiv preprint arXiv:1511.06421*, 2015. 2

[9] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *CVPR*, 2016. 2

[10] David Griffiths, Tobias Ritschel, and Julien Philip. OutCast: Outdoor single-image relighting with cast shadows. In *Computer Graphics Forum*, volume 41, pages 179–193. Wiley Online Library, 2022. 8

[11] Guangyun Han, Xiaohua Xie, Jianhuang Lai, and Wei-Shi Zheng. Learning an intrinsic image decomposer using synthesized RGB-D dataset. *IEEE Sign. Process. Letters*, pages 753–757, 2018. 2

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, 2017. 2

[13] 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. *IJCV*, 129(2):517–547, 2021. 5

[14] Pierre-Yves Laffont, Adrien Bousseau, and George Drettakis. Rich intrinsic image decomposition of outdoor scenes from multiple views. *IEEE transactions on visualization and computer graphics*, 19(2):210–224, 2012. 2

[15] Pierre-Yves Laffont, Adrien Bousseau, Sylvain Paris, Frederic Durand, and George Drettakis. Coherent intrinsic images from photo collections. *ACM TOG*, 2012. 2

[16] Louis Lettry, Kenneth Vanhoey, and Luc Van Gool. DARN: a deep adversarial residual network for intrinsic image decomposition. 2018. 2, 3

[17] Chuan Li and Michael Wand. Combining Markov random fields and convolutional neural networks for image synthesis. In *CVPR*, 2016. 2

[18] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, pages 2041–2050, 2018. 5, 8

[19] Andrew Liu, Shiry Ginosar, Tinghui Zhou, Alexei A Efros, and Noah Snavely. Learning to factorize and relight a city. In *ECCV*, 2020. 2

[20] Fujun Luan, Sylvain Paris, Eli Shechtman, and Kavita Bala. Deep photo style transfer. In *CVPR*, 2017. 2

[21] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 3, 4, 5, 6

[22] Moustafa Meshry, Dan B Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. In *CVPR*, 2019. 1, 2

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

[24] Takuya Narihira, Michael Maire, and Stella X Yu. Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In *ICCV*, 2015. 2

[25] Julien Philip, Michaël Gharbi, Tinghui Zhou, Alexei A Efros, and George Drettakis. Multi-view relighting using a geometry-aware network. *ACM TOG*, 38(4):78–1, 2019. 3

[26] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. 2001. 3

[27] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *CVPR*, 2022. 3, 5

[28] Viktor Rudnev, Mohamed Elgharib, William Alfred Peter Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *ECCV*, 2022. 1, 3, 4, 7, 8

[29] Yichang Shih, Sylvain Paris, Frédo Durand, and William T Freeman. Data-driven hallucination of different times of day from a single outdoor photo. *ACM TOG*, 2013. 2

[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 *CVPR*, 2021. 3

[31] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M. Alvarez, and Ping Luo. SegFormer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 3, 5

[32] Ye Yu, Abhimitra Meka, Mohamed Elgharib, Hans-Peter Seidel, Christian Theobalt, and William AP Smith. Self-supervised outdoor scene relighting. In *ECCV*, 2020. 1, 2, 3, 5, 6, 7, 8

[33] Ye Yu and William AP Smith. InverseRenderNet: Learning single image inverse rendering. In *CVPR*, 2019. 1, 2, 3

[34] Ye Yu and William AP Smith. Outdoor inverse rendering from a single image using multiview self-supervision. *IEEE TPAMI*, 44(7):3659–3675, 2021. 1, 2, 3, 6, 7

[35] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. NeRF++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 3

[36] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. NeRFactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM TOG*, 2021. 3

[37] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 2