Eurographics Symposium on Rendering (DL-only Track) (2021)
A. Bousseau and M. McGuire (Editors)

Single-image Full-body Human Relighting

Manuel Lagunas[1,2], Xin Sun[2], Jimei Yang[2], Ruben Villegas[2], Jianming Zhang[2], Zhixin Shu[2], Belen Masia[1], and Diego Gutierrez[1]

[1]Universidad de Zaragoza, I3A
[2]Adobe Research

Abstract

We present a single-image data-driven method to automatically relight images with full-body humans in them. Our framework is based on a realistic scene decomposition leveraging precomputed radiance transfer (PRT) and spherical harmonics (SH) lighting. In contrast to previous work, we lift the assumptions on Lambertian materials and explicitly model diffuse and specular reflectance in our data. Moreover, we introduce an additional light-dependent residual term that accounts for errors in the PRT-based image reconstruction. We propose a new deep learning architecture, tailored to the decomposition performed in PRT, that is trained using a combination of L1, logarithmic, and rendering losses. Our model outperforms the state of the art for full-body human relighting both with synthetic images and photographs.

CCS Concepts

• ***Computing methodologies*** → ***Rendering;*** *Neural networks;* ***Image-based rendering;***

1. Introduction

The growth in mobile computing, together with the increasing demand for visual social media has led to a tremendous rise in the popularity of consumer digital photography. In full-body photographs lighting plays an important role in transmitting the desired appearance of the subject, and changes in the illumination can lead to drastically different renditions. However, these photographs usually lack controlled illumination conditions.

We present a single-image relighting method that acts as a post-processing step, allowing a casual user to plausibly change and manipulate the illumination on a subject in a photograph. Human relighting usually benefits from multiple images as input, and requires solving an inverse rendering problem; in the general case, illumination information needs to be disambiguated from geometry and material appearance, based on simple pixel values. This is a well-studied but ill-posed problem, for which no definite solution exists. This paper takes a data-driven approach to the problem, requiring only one photograph and a user-specified target illumination as input (see Figure 1). Our method relies on precomputed radiance transfer [SKS02] (PRT) and spherical harmonics lighting [RH01] (SH). Based on this, a convolutional neural network (CNN) decomposes the image into its albedo, illumination, and light transport components; from which the shading can be easily computed. Disentangling the illumination from all other factors in the scene allows for effective relighting, while the PRT-based scheme enables fast, efficient rendering. Our work lifts the assumption of Lambertian materials present in previous single-image human relighting methods [SKCJ18; KE18]. We model the PRT decomposition in our framework by approximating material reflectance using an Oren-Nayar [ON94] and GGX microfacet model [WMLT07] for the diffuse and specular components, respectively. In addition, we extend the image reconstruction formulation by adding a *residual* term learned by our model, which accounts for errors in image reconstruction that would be obtained using only the terms proposed by PRT.

Figure 1: *Relighted results given a single image as input for different illumination maps. Please refer to Figure 12 for more details about the reconstructions.*

To train our model, we create a synthetic dataset containing almost 140,000 images with a rich variety of humans (105), poses (5), and illumination maps (266). We quantitatively and qualitatively evaluate relighting results on both synthetic images and real photographs, and perform extensive ablation experiments to validate our design choices in the model architecture, reflectance model

DOI: 10.2312/sr.20211300

for data generation, and loss functions. Compared with the current state of the art in full-body single-image human relighting [KE18], our model yields more accurate reconstructions of relighted images for both synthetic images and real photographs. We will make our code publicly available.

2. Related Work

Inverse rendering Single-image physically-based relighting typically requires solving an *inverse rendering* problem where shape, reflectance, and illumination need to be inferred from a single image. This is a highly ill-posed problem, with infinite solutions, classically solved assuming that some information is known beforehand. *Shape from shading* [Ram88; IH81] is one of the earliest methods, estimating shape from shading under a known illumination. Other methods estimate shape relying on simple illumination models such as directional, point, or area light sources [CK97; OD97; LHRG09], or environmental lighting encoded into spherical harmonics [JA11]. Reflectance and illumination can be estimated from a known convex shape [CR11], a shape with occluding contours [LGH*13], or just an approximated geometry [KSES14]. A similar line of research has focused on *intrinsic images* [BM12; YGL*14; GEZ*17; GMLG12; Wei01], which aims to decompose a scene into its shading and albedo components [LM71]. Our method draws inspiration from intrinsic images, and we estimate albedo and shading from a single input image; however, we additionally decompose shading into shape and illumination by developing a framework inspired by precomputed radiance transfer (PRT) [Ram09; Leh07; SKS02]. In addition, our decomposition also takes into account diffuse and specular material reflections, thus producing more realistic results.

Image-based rendering A classic application of image-based rendering (IBR [SK00]) allows to take several pictures of a subject from the same viewpoint under different illuminations, and relight it using a weighted linear combination of those images [DHT*00; DYB98]. More sophisticated approaches optimize energy functions [LI07], work with layered decompositions [SKG*12], or employ RGB-D cameras [HRDB16]. However, those techniques require a large number of input images, as well as precise control over the lighting, making them unfeasible for single-image, in-the-wild applications. Recent work exploits the potential of implicit representations and Fourier mappings of the input to learn high-quality 3D scene representations using one multi-layer perceptron (MLP) per scene and several hundreds of images as the input [MST*20; ZRSK20; BBJ*20], although these methods do not generalize across scenes. The work of Wang et al. [WWG*21] addresses this by combining implicit models with IBR to generate novel views without relighting. In contrast, our work develops a general framework that takes a single RGB image of a human as input, and outputs an intermediate representation suitable for relighting.

Data-driven methods Recent techniques leverage deep learning to predict illumination [HSH*17; HAL19; GSY*17; LSGM21], estimate specular reflectance and illumination [GRR*17; LN15], devise material reflectance metrics [DLG*20; LMS*19], or perform intrinsic image decomposition [MCZ*18; LS18; BM14]. For the particular case of humans, many relighting approaches rely on complex hardware setups [CCS*15; GLD*19; ZFT*21], which are not widely available; we instead focus on single RGB images as input.

Single-image human relighting approaches have been proposed for faces [WYL*20]: Sengupta et al. [SKCJ18] show how we can relight faces using convolutional neural networks and spherical harmonics, later extended with more complex model architectures [ZHSJ19], or by directly fitting encoder-decoder architectures to light-stage portrait data [SBT*19; NLML20]. Closer to ours is the work of Kanamori and Endo [KE18], performing full-body relighting. They use an encoder-decoder architecture to perform a single-image decomposition of the scene that is suitable for full-body human relighting. Our work lifts their assumption of materials being Lambertian by explicitly modeling the diffuse and specular reflectance in our data. We also add a residual term to the image reconstruction equation that allows to better model errors in the PRT image reconstruction.

3. Background

In this section, we briefly review the building blocks of our technique: Spherical harmonics (SH) lighting [RH01], and precomputed radiance transfer (PRT) [SKS02].

PRT [SKS02] and SH lighting [RH01] enable rendering dynamic low-frequency environments with realistic highlights and real-time shading. They estimate the amount of radiance reflected at a point in the scene by solving a simplified version of the rendering equation:

$$R(x) = \int_{\mathbb{S}^2} L(x,\omega_i)\, T(x,\omega_i)\, \mathrm{d}\omega_i, \tag{1}$$

where R is the reflected radiance or image intensity computed over the sphere $\mathbb{S}^2$ of incoming directions ω_i, L is the incoming light at point x from direction ω_i, and T is a transport function computed for each vertex that includes the material reflectance f_r, visibility term V that is 1 if the point is not occluded and 0 otherwise, and the cosine term which uses the normal $\mathbf{n}$ at point x. The function T can be expressed as:

$$T(x,\omega_i) = f_r(x,\omega_i)\, V(x,\omega_i)\, (\omega_i \cdot \mathbf{n}). \tag{2}$$

The formulation presented by PRT expands the illumination L and the transport T using (real) spherical harmonics basis functions $Y_{l,m}$, such that:

$$\begin{aligned} L(x,\omega_i) &= \sum_{l=0}^{\infty} \sum_{m=-l}^{l} L_{l,m}(x) Y_{l,m}(\omega_i), \\ T(x,\omega_i) &= \sum_{l=0}^{\infty} \sum_{m=-l}^{l} T_{l,m}(x) Y_{l,m}(\omega_i), \end{aligned} \tag{3}$$

where $L_{l,m}$ and $T_{l,m}$ are the corresponding coefficients for illumination and transport, respectively (see [Ram09, Sections 3 and 4] for additional details on how to obtain $T_{l,m}$ and $L_{l,m}$). The integral in Equation 1 then becomes:

$$R(x) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} L_{l,m}(x)\, T_{l,m}(x). \tag{4}$$

This formulation has two advantages: It allows to approximate the rendering equation as a fast dot product, and it disentangles the illumination and the transport in the scene. In this way, relighting a scene only requires computing the coefficients of the new illumination $L'_{l,m}$, while keeping $T_{l,m}$ fixed.

Traditionally, relighting methods based on the estimation of illumination and transport coefficients from a single image soften the problem by assuming that the scene has a light source at a sufficient distance to neglect the angular variation between points, i.e., $L(x,\omega_i) \approx L(\omega_i)$. They also estimate a transport function T encoding only the cosine term [SKCJ18], or the cosine term together with the visibility function [KE18]. These methods assume all materials to be Lambertian, removing the reflectance term from the transport $T_{l,m}(x)$, and modeling it as a constant for each point of the scene represented by the albedo $\rho(x)$. With this, expressing $L_{l,m}$ as a vector $\mathbf{L}$ and $T_{l,m}(x)$ as a vector per point of the scene $\mathbf{T}(x)$, $R(x)$ can be approximated as (hereafter, we omit the dependency on x for clarity):

$$R \approx \underbrace{\rho}_{albedo} \cdot \underbrace{(\mathbf{T}^T \cdot \mathbf{L})}_{shading\ S}, \tag{5}$$

where the dot product between transport and illumination yields the shading $S(x)$ of a point in the scene, then scaled by the albedo ρ. The error of the approximation will be related to the number of coefficients used to estimate the illumination and transport in Equations 3 and 4; this number of coefficients depends on the number of terms used to approximate the infinite term summation of Equation 4, $l = [0..N]$.

To increase realism in the inferred and rendered images, we lift the Lambertian material assumption of previous work and include a better approximation of material reflectance in the transport function $\mathbf{T}$. We approximate the reflectance term in Equation 2 by keeping the albedo ρ as a constant and using a white material with an Oren-Nayar [ON94] for the diffuse component, and a GGX model with Smith shadowing factor and Fresnel [WMLT07] for the specular reflection. Then, we use Equation 3 to encode such reflectance in a new transport function $\mathbf{T}$, later used to render new images with Equation 5. As Figure 2 shows, this allows to better capture the directionality of specular reflections. Our reflectance model employs the following parameters: albedo, roughness, metallic, and transparency (refer to Section 5 for additional details). Both the Oren-Nayar and the GGX models share the same roughness parameter. The final reflectance model is defined as a combination of up to seven BSDFs, which can be either a diffuse Oren-Nayar microfacets model or a specular GGX model.

4. Our Image Reconstruction Formulation

This section describes our image reconstruction formulation, including the motivation behind the addition of a new residual term.

Since using a large number of basis coefficients in Equation 5 to approximate R with a low error is computationally expensive, we introduce an additional residual vector $\mathbf{E}$, leading to:

$$R \approx \underbrace{\rho}_{albedo} \cdot \underbrace{(\mathbf{T}^T \cdot \mathbf{L})}_{shading\ S} + \underbrace{(\mathbf{E}^T \cdot \mathbf{L})}_{residual\ E}, \tag{6}$$

Figure 2: *Comparison between the data generated with our framework and that of the recent work by Kanamori and Endo [KE18], used to train the respective models. Our transport function* $\mathbf{T}$ *takes into account angular dependencies in the reflectance term, better capturing specular reflections and improving high-frequency details in the shading.*

where the dot product between the residual vector $\mathbf{E}$ and the illumination $\mathbf{L}$ yields a residual value per point $E(x)$. Again, the dependency on x is omitted for clarity, but Equation 6 applies to each point in the scene, yielding the corresponding images; in the following, we will use S to denote the shading image, and E for the residual image. The residual vector $\mathbf{E}$ does not have a physical meaning; instead, it is a set of learned coefficients that aim to model the errors in image reconstruction that we would obtain using only the terms (albedo, transport, and illumination) with a limited number of coefficients.

4.1. Problem Formulation

Our main goal is to relight an image ψ with a full-body human in it, given a user-specified target illumination $\mathbf{L}'$:

$$\hat{\psi} = \mathcal{R}(\psi, \mathbf{L}'), \tag{7}$$

where $\mathcal{R}$ is a relighting function, and $\hat{\psi}$ is the resulting relighted image with target illumination $\mathbf{L}'$.

Using a model such as the one in Equation 6, one can change $\mathbf{L}$ to $\mathbf{L}'$ to obtain the relighted image. However, given a single image as input, the transport $\mathbf{T}$, illumination $\mathbf{L}$, residual $\mathbf{E}$, and albedo ρ are unknown. To obtain an approximation of $\mathbf{T}$, $\mathbf{L}$, $\mathbf{E}$, and ρ, we introduce the parametric function $\mathcal{G}$, which takes as input the image ψ and a set of parameters β:

$$\{\mathbf{T}, \mathbf{L}, \mathbf{E}, \rho\} \approx \mathcal{G}(\psi, \beta). \tag{8}$$

In particular, we model $\mathcal{G}$ using a convolutional neural network whose parameters are represented by β. Note that $\mathcal{G}$ tries to approximate each of the terms $\{\mathbf{T}, \mathbf{L}, \mathbf{E}, \rho\}$ irrespective of the underlying reflectance model previously used to generate them. With the output of $\mathcal{G}$ and a given user-specified illumination $\mathbf{L}'$, we can use Equation 6 to directly approximate the relighting function $\mathcal{R}$.

5. Dataset

To learn the parametric function $\mathcal{G}$ introduced in Section 4 we have created a synthetic human image dataset of almost 140,000 images including a rich variety of humans, poses, and illuminations, which we describe in this section.

Human 3D models Existing models captured using photogrammetry mostly consist only of diffuse and normal maps. To fully exploit the capabilities of our framework and go beyond Lambertian materials, we purchase rigged 3D human models and clothing from the *DAZ* website [DAZ], which include realistic materials and texture maps for diffuse color, specular, opacity, roughness, metallic, translucency, and normals. In total, we collected 105 different clothed models; augmented with five poses each, this yields a total of 525 different renditions. For each pose we simulate cloth interaction after posing the model, and, to foster diversity, perform subtle random changes to the hue of the diffuse color.

Illumination maps We used freely-available spherical high-dynamic range images (HDRIs) from HDRIHaven [HDR], corresponding to both indoor and outdoor scenarios. To normalize the HDRIs, we compute a *reference radiance* for each image by obtaining the mean shading in Equation 5, where $\mathbf{L}$ are the coefficients of the HDRI, and $\mathbf{T}$ is obtained analytically by sampling all unit directions in the sphere. We scale all the illuminations $\mathbf{L}$ to have a reference radiance in the range $[0.7, 0.9]$. In total we gathered 266 different HDRIs.

Rendering We used Monte Carlo path tracing to render realistic images and to obtain the transport vector $\mathbf{T}$ for each scene. To generate $\mathbf{L}$ for each illumination, we integrate over the unit sphere of directions. We fix $N = 4$ ($l = [0..4]$), which leads to 25 spherical harmonics coefficients in $\mathbf{T}$ and $\mathbf{L}$ (in contrast, the work of Kanamori and Endo [KE18] estimates only Lambertian materials and uses $N = 2$). Among all the available maps defining reflectance for each purchased model, during rendering we employ the albedo (diffuse color), roughness, metallic, and transparency maps. In total, we render 139,650 different scenes. For each scene, we generate: Its path-traced (PT) image, the PRT image computed using Equation 5, an alpha mask of the human, the shading, the normals, the albedo, and a material map containing the roughness, transparency, and metallic, each of them encoded in separate channel of an RGB image. All images are rendered with a resolution of 768×768 pixels; using 256 samples per pixel for the PT image, and 1,024 for the transport $\mathbf{T}$ and all other scene properties. Figure 3 shows two samples from our dataset, cropped down from the squared aspect ratio.

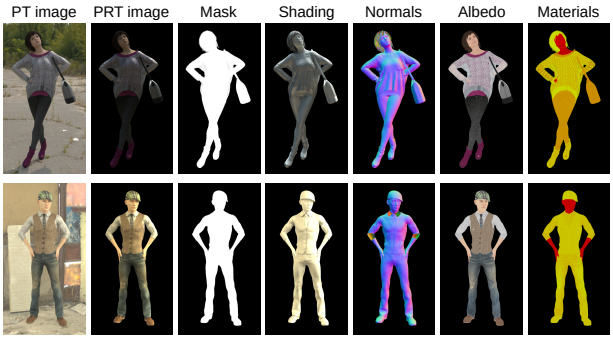

Figure 3: *Two examples in our dataset. For each scene we obtain its path-traced (PT) rendered image, its PRT image rendered with our image reconstruction formulation, the alpha mask, the shading, the normals, the albedo, and a material map describing the roughness, transparency, and metallic (each encoded in a separate channel of an RGB image).*

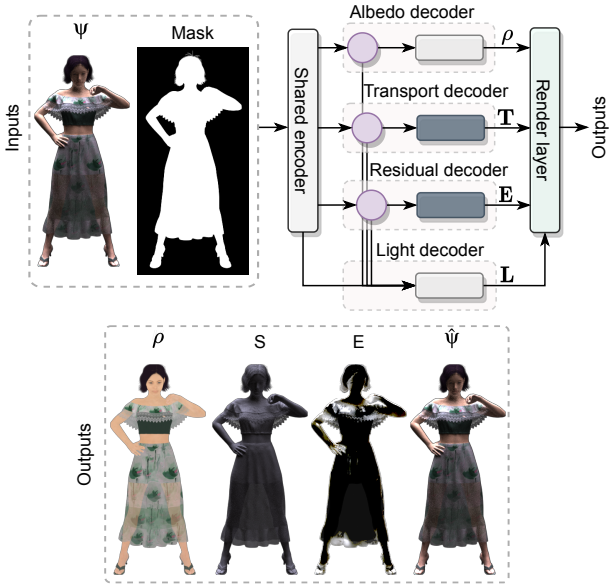

Figure 4: *Our model architecture. The masked input image goes through a shared encoder that converts it into a feature map. Such feature map simultaneously serves as an input to the albedo, transport, and residual decoders. The three decoders output the albedo* ρ*, transport* ***T****, and residual coefficients* ***E****, respectively. The features from these three decoders and from the shared encoder are concatenated and fed to the light decoder, which outputs the illumination coefficients* ***L****. Last, the rendering layer outputs the albedo (equal to the output from the albedo decoder), shading, residual image, and the final relighted image.*

6. Our Model

In this section we explain our model architecture and its components, together with an intuition behind our design choices; in addition, we provide details on our training, hyper-parameters, and loss function.

6.1. Model Architecture

To represent our parametric function $\mathcal{G}$ we use a convolutional neural network based on a UNet-like model [RFB15]. Figure 4 shows an overview. It consists of a shared encoder that receives the input image ψ, and several decoders responsible for estimating albedo ρ, transport $\mathbf{T}$, residual coefficients $\mathbf{E}$, and the illumination of the input image $\mathbf{L}$. We add skip-connections between the shared encoder and each decoder to encourage better reconstructions, except for the light decoder. Last, we have a rendering layer based on Equation 6 that generates the shading S, the residual E, and the final relighted image $\hat{\psi}$.

Shared encoder Our encoder has a standard architecture consisting of several convolutional blocks with batch-normalization (BN) that sequentially decrease the resolution of the features by a factor of two. The features between convolutional blocks are used as skip-connections with the decoders.

Decoders Each decoder has a *residual block* (similar to ResNet [HZRS16]), and a *generator block* except for the light decoder that only has a generator block. The generator block varies between decoders. The output of the albedo, transport, and residual coefficients decoder has the same spatial resolution as the input image. We only add batch-normalization to the albedo decoder. The architecture of each generator is as follows (see also Figure 5):

- The albedo decoder has several convolutional blocks with skip-connections. In each convolutional block features are scaled by a factor of two. The output of the albedo decoder is clamped to lie in the range $[0,1]$.
- To properly capture geometry and material reflectance in the scene, a good estimation of the transport matrix $\mathbf{T}$ is needed. The transport and residual decoders feature a generator tailored for the PRT decomposition in Equation 6. Deep neural networks, by design, add non-linear functions that clamp negative values. However, the transport coefficients are defined with both positive and negative values. Thus, we would rely on the last convolution without non-linearities to generate all the negative content in the coefficients. To give additional degrees of freedom to the decoders, we decompose the coefficients as $\mathbf{T} = \mathbf{T}^+ - |\mathbf{T}^-|$ where $\mathbf{T}^+$ corresponds to the positive part and $|\mathbf{T}^-|$ is the absolute value of the negative part. Instead of directly predicting $\mathbf{T}$, we add two generators (similar to the albedo one) to predict $\mathbf{T}^+$ and $|\mathbf{T}^-|$, respectively, and later we reconstruct the coefficients $\mathbf{T}$. We apply a similar strategy to the residual coefficients $\mathbf{E}$.
- The light decoder differs from the previous as its input is the output of the shared encoder and the residual blocks of the albedo, transport, and residual decoders. Those features go straight to a generator that follows a similar decomposition as for the transport and residual decoder, however, the generator architectures differ. The generator has several convolutional blocks that reduce the spatial dimensions of the features by a factor of two. After the convolutions, we perform an average pooling making the features one-dimensional, and a fully-connected layer outputs the positive and negative illumination coefficients in each generator, with shape $3 * 25$ (25 being the total number of coefficients when $N = 4$). Then, we reconstruct $\mathbf{L}$ using the positive and negative part.

6.2. Training

The dataset in Section 5 is split into training and validation, where we select 7 clothed models (with all their poses) that are representative of challenging scenes as the validation set. The rest of the humans with their poses are used for training. The input to our model are images rendered with PRT, where we crop the human using the bounding-box defined by the mask with a padding of 20 pixels. Since our network is fully-convolutional it allows inputs of arbitrary resolution. We normalize the image pixels to lie in the range $[-1,+1]$ and multiply it by the alpha mask before forwarding it through the model. For training we use the Adam optimization algorithm [KB14] with the decoupled weight decay regularization [LH17]. The learning rate has a value of $5 \cdot 10^{-5}$. We set an effective batch size of 16. We use the PyTorch framework [PGM*19] with PyTorch-Lightning [Fal] to design our model and experiments. The model is trained for 25 epochs on eight Tesla V100-SXM2-16GB, lasting 55 hours approximately.

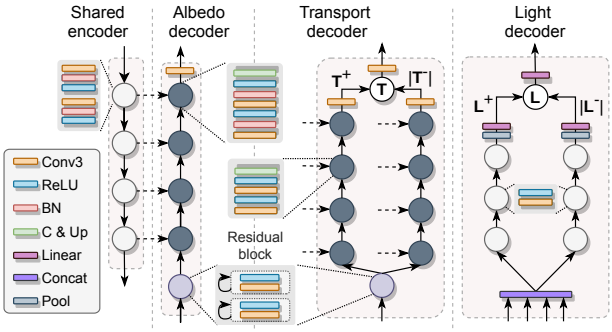

Figure 5: *Workflow of each component of our model. The shared encoder contains several convolutional blocks that reduce the spatial dimensions by two and output a feature map of the input. Such feature map goes to the albedo, transport, and residual decoders. Each decoder (except for the light one) has a residual block and a generator. The generator concatenates skip-connections and upscales (C & Up) the spatial resolution of the features. The output of the decoders has the same spatial resolution as the input image. Last, the light decoder uses the features of the encoder, together with the features from the residual block of each decoder, to predict the illumination in the scene.*

6.3. Loss Functions

Our loss function $\mathcal{L}$ can be expressed as:

$$\mathcal{L} = \mathcal{L}_\rho + \mathcal{L}_\mathbf{T} + \mathcal{L}_\mathbf{L} + \mathcal{L}_\mathrm{S} + \mathcal{L}_{\hat{\psi}}. \tag{9}$$

where each term supervises the prediction of albedo, transport, illumination, shading, and the final relighted image. Note that the residual coefficients are not directly supervised. Instead, we let the network freely learn a set coefficients $\mathbf{E}$ that aim to improve the quality of the rendered images. Each of the terms in $\mathcal{L}$ is additionally composed of different losses. We linearly combine the different terms using a weight of 1 for all of them.

- **Reconstruction loss** ($\mathcal{L}_{L1}$) We apply an L1 loss function to each predicted map with respect to ground-truth data. To encourage a better reconstruction, we leverage the architecture tailored for PRT rendering, and additionally include an L1 loss between the positive and negative coefficients in $\mathcal{L}_\mathbf{T}$ and $\mathcal{L}_\mathbf{L}$.
- **Render loss** ($\mathcal{L}_r$) The terms in Equation 6 are computed using the albedo, transport, illumination, and residual vectors. For each of those vectors (except the residual $\mathbf{E}$), there is both a predicted (which is being learned) and a ground truth vector. To increase robustness, we introduce in $\mathcal{L}_\mathrm{S}$ and $\mathcal{L}_{\hat{\psi}}$ an L1 error term for each possible way of generating the shading and relighted image in Equation 6 from the predicted and ground truth vectors.

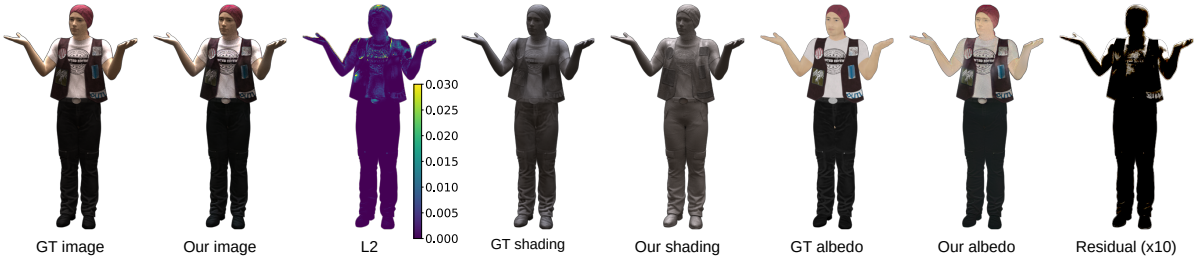

Figure 6: *Example result of our model for a synthetic image (see also Table 1, synthetic images). Neither the human nor the illumination were used for training. We show direct comparisons with the ground truth (GT), the L2 error in the final image, and our residual term scaled by a factor of 10 for visualization purposes.*

Figure 7: *Comparison between our model and the model provided by Kanamori and Endo [KE18] in two examples of the validation dataset. We can see how our model outperforms them rendering the input image, albedo, and shading. Note that the shading encodes both the transport and the illumination of the scene.*

- **Log loss** ($\mathcal{L}_{\log}$) The transport, and the illumination coefficients have an unbounded range. To compress it, we apply a logarithmic loss of the following form:

$$\mathcal{L}_{\log} = ||\log(|x|+1) - \log(|\hat{x}|+1)||_2^2$$

in $\mathcal{L}_{\mathbf{T}}$ and $\mathcal{L}_{\mathbf{L}}$. We apply $|x|$ in the logarithmic loss to avoid errors on the negative values of the coefficients. We leverage the PRT decomposition to apply the logarithmic loss also to the positive and negative decomposition of transport and illumination.

7. Results

We show and evaluate results of our model on both synthetic images, where ground truth data is available, and real photographs. Throughout the evaluation, we show the reconstructed albedo ρ, shading S (resulting from the combination of transport $\mathbf{T}$ and target illumination $\mathbf{L}'$, see Equation 6), the final rendered result $\hat{\psi}$, and the residual image E. We also include ablation studies to clearly demonstrate the influence of each component in the final relighted images.

Synthetic images We use the validation subset of our dataset (see Section 6.2) rendered with six new illuminations not used for training: *ennis*, *grace*, *pisa*, *doge*, *glacier* and *uffizi* [Lab]. We render the final relighted (target) image using the predicted illumination of the scene to reconstruct the shading and the residual. Since ground truth data is available, we also compute quantitative error measures for the albedo, shading, and final rendered image. Specifically, we compute the L1 and L2 distances, as well as PSNR, averaged across the dataset. Table 1 (synthetic images) shows the results, including a comparison with the pretrained model of the recent work by Kanamori and Endo [KE18]. Our more complete material reflectance formulation, together with our residual term (see ablation studies in Subsection 7.1) lead to significantly lower L1 and L2 values, and higher PSNR for the albedo and shading, as well as the final relighted image. Figure 6 shows a direct comparison of our reconstructed image with the ground truth; both images match with a very small L2 error. Figure 7 shows a comparison between our model and the pretrained model given in the work of Kanamori and Endo on synthetic images. We can see how our model better estimates the shading and albedo, leading to more accurate results where directional effects are better reproduced (see the highlights in the face of the first image, for instance).

Real photographs To test our model on real photographs we use free-license images downloaded from Unsplash [Uns]. To obtain the alpha mask we rely on freely available APIs [Rem]. In total we collected 10 different images with a single human in them. Error metrics for the resulting rendered images, averaged over the 10 photos, can be found in Table 1 (real photographs). As with synthetic images, our results significantly outperform previous work [KE18]. Maybe surprisingly, the error metrics indicate better results with real photographs (both for our method and using

Figure 8: *Example results of our model on real photographs (see also Table 1, real photographs). For each image, from left to right: Ground truth input image, resulting image relighted with our model, L2 error, albedo, shading, and residual term scaled by a factor of 10 for visualization purposes.*

Table 1: *Quantitative results of our model for synthetic images and real photographs, measured with three metrics: L1 and L2 distances, and PSNR. Note that the L1 and L2 metrics have been scaled by a factor of 100. We also include a comparison to the model of Kanamori and Endo [KE18], which our model consistently outperforms. Boldface highlights the best result in each case.*

	Synthetic images									Real photographs		
	Albedo			Shading			Image			Image		
Model	L1 (x100)	L2 (x100)	PSNR	L1 (x100)	L2 (x100)	PSNR	L1 (x100)	L2 (x100)	PSNR	L1 (x100)	L2 (x100)	PSNR
Ours	**2.88**	**0.44**	**24.18**	**3.77**	**0.71**	**24.05**	**1.64**	**0.19**	**28.94**	**1.17**	**0.08**	**31.42**
Kanamori and Endo	4.95	1.19	20.68	6.75	1.90	18.29	2.94	0.47	26.06	2.14	0.20	28.38

Table 2: *Quantitative results of our model and all the ablation experiments for synthetic images and real photographs. We measure three different metrics: L1 and L2 distances, and PSNR. Note that the L1 and L2 metrics have been scaled by a factor of 100. Our model outperforms all other experiments. Boldface highlights the best result in each case.*

	Synthetic images									Real photographs		
	Albedo			Shading			Image			Image		
Model	L1 (x100)	L2 (x100)	PSNR	L1 (x100)	L2 (x100)	PSNR	L1 (x100)	L2 (x100)	PSNR	L1 (x100)	L2 (x100)	PSNR
Ours	**2.88**	**0.44**	**24.18**	**3.77**	**0.71**	**24.05**	**1.64**	**0.19**	**28.94**	**1.17**	**0.08**	**31.42**
Without **E**	3.67	0.66	23.05	6.71	2.74	17.89	1.97	0.24	27.13	2.64	0.29	26.24
Without PRT decomposition	4.54	1.00	21.08	10.57	5.69	14.43	2.13	0.24	26.80	2.55	0.30	26.66
Without $\mathcal{L}_{\log}$	4.02	0.83	21.84	10.34	5.35	14.55	2.14	0.24	27.82	2.31	0.25	26.85
With $N = 2$	3.31	0.58	23.33	8.60	4.21	16.18	1.83	0.22	28.33	2.08	0.18	27.76
With $\mathbf{T}^*$	3.68	0.76	21.74	7.53	3.54	16.65	2.25	0.31	27.29	1.75	0.14	29.43
Lambertian materials	3.58	0.68	22.66	7.22	2.91	17.09	1.91	0.21	27.92	1.82	0.18	29.15

the pretrained model of Kanamori and Endo) than using synthetic images. This is possibly due to the fact that the synthetic validation dataset contains some quite extreme illuminations (e.g., *glacier* or *grace*), while the photographic dataset has more natural illuminations that the two models are able to reproduce better. Figure 8 shows the reconstruction performed by our model for two different input photographs, including albedo and shading components, while a direct comparison with previous work is shown in Figure 9. Again, we see how our model is able to better capture directional effects (see, e.g., the faces or the highlights in the jackets) and overall produce more accurate reconstructions.

Finally, in Figure 12 we show a variety of relighting results under different illuminations (refer to the supplemental material for the full set, a table with quantitative metrics, as well as a video when rotating the illumination maps). For each input photo and illumination map we show the final relighted image, and the reconstructed shading and residual terms.

7.1. Ablation Studies

We evaluate the contribution of our design choices with a series of ablation experiments performed on both the synthetic images and the real photographs. In particular, we first compare the performance of our model (*Ours*) without the residual generator predicting **E** (*Without* $\boldsymbol{E}$) and without including the PRT decomposition in the architecture of the generators (*Without PRT decomposition*). Then, we evaluate the impact of the logarithmic loss $\mathcal{L}_{\log}$ in the prediction of **T** and **L** (*Without* $\mathcal{L}_{\log}$), as well as the performance of our model when using only nine coefficients (*With* $N = 2$). To avoid using a constant albedo in Equation 6, we combine the different terms that define reflectance (*With* $\boldsymbol{T}^*$) into a single vector

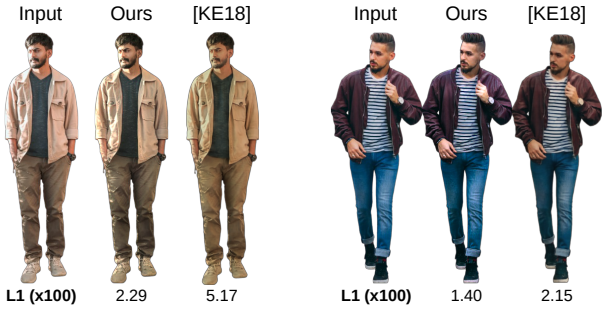

Figure 9: *Image reconstructions obtained by our model, and the model provided by Kanamori and Endo [KE18]. We can see how our model outperforms them in the three metrics (see also Table 1, real photographs). Note that the L1 and L2 metrics have been scaled by a factor of 100. In addition, our model better captures skin and cloth albedo, and the directionality of the illumination.*

Figure 10: *Reconstruction results obtained on the different ablation experiments. We can clearly observe how our full model better captures the appearance of the input photograph.*

$\mathbf{T}^* = (\rho * \mathbf{T} + \mathbf{E})$. Last, to showcase the benefit of our reflectance, we have trained a model using purely Lambertian materials in our data (*Lambertian materials*).

Table 2 shows the results (including albedo and shading for synthetic images) for the L1, L2, and PSNR metrics for all the ablation studies. All options yield significantly inferior results when compared with our full model. Figure 10 further illustrates this on an example using a real photograph. One could think that the model *With* $\boldsymbol{T}^*$ would obtain better performance since it does not need to assume a constant albedo ρ in the reflectance. However, $\mathbf{T}^*$ requires estimating 25 different RGB maps (with $N = 4$), leading to additional complexity that hinders convergence and produces higher errors.

8. Discussion

We have presented a model for human relighting that requires a single image as input. We lift the assumption on Lambertian materials and include a better approximation of material reflectance in our transport function. Moreover, we introduce an additional residual term which further mitigates errors in the PRT-based final reconstruction. This additional term becomes increasingly relevant for challenging illuminations, such as backlighting, where the overall dark appearance of the image does not allow for an accurate estimation of the PRT terms. The resulting errors are absorbed by our residual, helping to produce good final reconstructions. Overall our results show compelling estimations of albedo and shading (transport and illumination), leading to accurate relighting reconstructions for both synthetic images and real photographs.

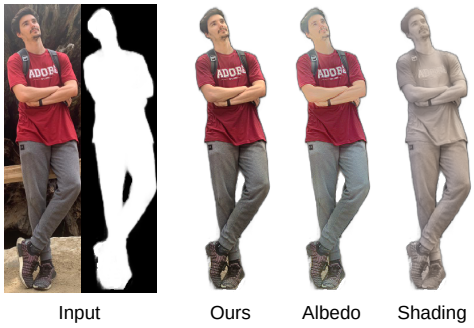

Figure 11: *Example of the limitations of our model. The strong presence of stray light in the input image leads to an excessively flat albedo, seen especially in the head and shoulders area, while some texture details appear in the shading image.*

Nevertheless, our work is not free of limitations. Figure 11 shows a difficult case with a real photograph as input. While our reconstruction is still plausible, the strong presence of stray light (especially on top) leads to an excessively flat, milky estimation of the albedo in the head and shoulders area. Also, our shading reconstruction carries traces of texture details in the T-shirt, which remains an open problem in intrinsic images decomposition.

Human relighting poses many challenges not fully investigated in this work. Besides making the model more robust to poorly lit input images, being able to take into account other lighting effects such as subsurface scattering [JZJ*15], anisotropy in cloth materials [ACG*17], or more complex reflectance models, remain interesting open problems. Moreover, one implicit problem of SH-based lighting is the need for a large number of coefficients to reconstruct high-frequency details. While we mitigate this problem by introducing the residual term, complex high-frequency effects are still an open challenge. Another exciting avenue of future work is to extend the potential of our approach, for instance by using contrastive loss functions, or proposing self-supervised schemes that would avoid having to generate additional synthetic data.

Acknowledgements

We want to thank the anonymous reviewers for their feedback on the manuscript; also, thanks to Ibon Guillen, and Adrian Jarabo for the occasional discussions about the paper. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (CHAMELEON project, grant agreement No 682080), from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreements No 765121 and 956585, from the Spanish Ministry of Economy

and Competitiveness (project PID2019-105004GB-I00), and generous gifts from Adobe Systems.

References

[ACG*17] ALIAGA, CARLOS, CASTILLO, CARLOS, GUTIERREZ, DIEGO, et al. "An appearance model for textile fibers". *Computer Graphics Forum*. Vol. 36. 4. 2017 8.

[BBJ*20] BOSS, MARK, BRAUN, RAPHAEL, JAMPANI, VARUN, et al. "NeRD: Neural Reflectance Decomposition from Image Collections". *arXiv preprint arXiv:2012.03918* (2020) 2.

[BM12] BARRON, JONATHAN T and MALIK, JITENDRA. "Color constancy, intrinsic images, and shape estimation". *Proc. of ECCV*. 2012, 57–70 2.

[BM14] BARRON, JONATHAN T and MALIK, JITENDRA. "Shape, illumination, and reflectance from shading". *IEEE Trans. on PAMI* 37.8 (2014) 2.

[CCS*15] COLLET, ALVARO, CHUANG, MING, SWEENEY, PAT, et al. "High-quality streamable free-viewpoint video". *ACM Trans. on Graphics* 34.4 (2015) 2.

[CK97] CHRISTOU, C. G. and KOENDERINK, J. J. "Light source dependence in shape from shading". *Vision Res.* 37.11 (1997) 2.

[CR11] CHANDRAKER, MANMOHAN and RAMAMOORTHI, RAVI. "What an image reveals about material reflectance". *Proc. of ICCV*. 2011 2.

[DAZ] DAZ. URL: https://www.daz3d.com/ 4.

[DHT*00] DEBEVEC, PAUL, HAWKINS, TIM, TCHOU, CHRIS, et al. "Acquiring the reflectance field of a human face". *Proc. of SIGGRAPH*. 2000 2.

[DLG*20] DELANOY, JOHANNA, LAGUNAS, MANUEL, GALVE, IGNACIO, et al. "The role of objective and subjective measures in material similarity learning". *ACM SIGGRAPH 2020 Posters*. 2020, 1–2 2.

[DYB98] DEBEVEC, PAUL, YU, YIZHOU, and BORSHUKOV, GEORGE. "Efficient view-dependent image-based rendering with projective texture-mapping". *Proc. of EGSR*. 1998 2.

[Fal] FALCON, WA ET AL. *PyTorch Lightning*. URL: https://github.com/PyTorchLightning/pytorch-lightning 5.

[GEZ*17] GARCES, ELENA, ECHEVARRIA, JOSE I, ZHANG, WEN, et al. "Intrinsic light field images". *Computer Graphics Forum*. Vol. 36. 8. 2017 2.

[GLD*19] GUO, KAIWEN, LINCOLN, PETER, DAVIDSON, PHILIP, et al. "The relightables: Volumetric performance capture of humans with realistic relighting". *ACM Trans. on Graphics* 38.6 (2019) 2.

[GMLG12] GARCES, ELENA, MUNOZ, ADOLFO, LOPEZ-MORENO, JORGE, and GUTIERREZ, DIEGO. "Intrinsic images by clustering". *Computer Graphics Forum* 31.4 (2012) 2.

[GRR*17] GEORGOULIS, STAMATIOS, REMATAS, KONSTANTINOS, RITSCHEL, TOBIAS, et al. "Reflectance and natural illumination from single-material specular objects using deep learning". *IEEE Trans. on PAMI* 40.8 (2017) 2.

[GSY*17] GARDNER, MARC-ANDRÉ, SUNKAVALLI, KALYAN, YUMER, ERSIN, et al. "Learning to predict indoor illumination from a single image". *arXiv preprint arXiv:1704.00090* (2017) 2.

[HAL19] HOLD-GEOFFROY, YANNICK, ATHAWALE, AKSHAYA, and LALONDE, JEAN-FRANÇOIS. "Deep sky modeling for single image outdoor lighting estimation". *Proc. of CVPR*. 2019 2.

[HDR] HDRIHAVEN. URL: https://www.hdrihaven.com/ 4.

[HRDB16] HEDMAN, PETER, RITSCHEL, TOBIAS, DRETTAKIS, GEORGE, and BROSTOW, GABRIEL. "Scalable inside-out image-based rendering". *ACM Trans. on Graphics* 35.6 (2016) 2.

[HSH*17] HOLD-GEOFFROY, YANNICK, SUNKAVALLI, KALYAN, HADAP, SUNIL, et al. "Deep outdoor illumination estimation". *Proc. of CVPR*. 2017 2.

[HZRS16] HE, KAIMING, ZHANG, XIANGYU, REN, SHAOQING, and SUN, JIAN. "Deep residual learning for image recognition". *Proc. of CVPR*. 2016, 770–778 5.

[IH81] IKEUCHI, KATSUSHI and HORN, BERTHOLD KP. "Numerical shape from shading and occluding boundaries". *Artificial Intelligence* 17.1-3 (1981), 141–184 2.

[JA11] JOHNSON, MICAH K and ADELSON, EDWARD H. "Shape estimation in natural illumination". *Proc. of CVPR*. 2011 2.

[JZJ*15] JIMENEZ, JORGE, ZSOLNAI, KÁROLY, JARABO, ADRIAN, et al. "Separable subsurface scattering". *Computer Graphics Forum*. Vol. 34. 6. 2015 8.

[KB14] KINGMA, DIEDERIK P and BA, JIMMY. "Adam: A method for stochastic optimization". *arXiv preprint arXiv:1412.6980* (2014) 5.

[KE18] KANAMORI, YOSHIHIRO and ENDO, YUKI. "Relighting humans: occlusion-aware inverse rendering for full-body human images". *ACM Trans. on Graphics (SIGGRAPH Asia)*. 2018 1–4, 6–8.

[KSES14] KHOLGADE, NATASHA, SIMON, TOMAS, EFROS, ALEXEI, and SHEIKH, YASER. "3D object manipulation in a single photograph using stock 3D models". *ACM Trans. on Graphics* 33.4 (2014) 2.

[Lab] LAB, ICT VISION & GRAPHICS. URL: https://vgl.ict.usc.edu/Data/HighResProbes/ 6.

[Leh07] LEHTINEN, JAAKKO. "A framework for precomputed and captured light transport". *ACM Trans. on Graphics* 26.4 (2007), 13–es 2.

[LGH*13] LOPEZ-MORENO, JORGE, GARCES, ELENA, HADAP, SUNIL, et al. "Multiple light source estimation in a single image". *Computer Graphics Forum*. Vol. 32. 8. 2013 2.

[LH17] LOSHCHILOV, ILYA and HUTTER, FRANK. "Decoupled weight decay regularization". *arXiv preprint arXiv:1711.05101* (2017) 5.

[LHRG09] LOPEZ-MORENO, JORGE, HADAP, SUNIL, REINHARD, ERIK, and GUTIERREZ, DIEGO. "Light Source Detection in Photographs." *CEIG*. 2009, 161–167 2.

[LI07] LEMPITSKY, VICTOR and IVANOV, DENIS. "Seamless mosaicing of image-based texture maps". *Proc. of CVPR*. 2007 2.

[LM71] LAND, EDWIN H and MCCANN, JOHN J. "Lightness and retinex theory". *JOSA* 61.1 (1971) 2.

[LMS*19] LAGUNAS, MANUEL, MALPICA, SANDRA, SERRANO, ANA, et al. "A Similarity Measure for Material Appearance". *ACM Trans. on Graphics (SIGGRAPH)* 38.4 (2019) 2.

[LN15] LOMBARDI, S. and NISHINO, K. "Reflectance and illumination recovery in the wild". *IEEE Trans. on PAMI* 38.1 (2015) 2.

[LS18] LI, Z. and SNAVELY, N. "Learning intrinsic image decomposition from watching the world". *Proc. of CVPR*. 2018, 9039–9048 2.

[LSGM21] LAGUNAS, MANUEL, SERRANO, ANA, GUTIERREZ, DIEGO, and MASIA, BELEN. "The joint role of geometry and illumination on material recognition". *Journal of Vision (JoV)* 21 (2021) 2.

[MCZ*18] MA, WEI-CHIU, CHU, HANG, ZHOU, BOLEI, et al. "Single image intrinsic decomposition without a single intrinsic image". *Proc. of ECCV*. 2018, 201–217 2.

[MST*20] MILDENHALL, BEN, SRINIVASAN, PRATUL P, TANCIK, MATTHEW, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis". *Proc. of ECCV*. 2020 2.

[NLML20] NESTMEYER, THOMAS, LALONDE, JEAN-FRANÇOIS, MATTHEWS, IAIN, and LEHRMANN, ANDREAS. "Learning physics-guided face relighting under directional light". *Proc. of CVPR*. 2020 2.

[OD97] OKATANI, TAKAYUKI and DEGUCHI, KOICHIRO. "Shape reconstruction from an endoscope image by shape from shading technique for a point light source at the projection center". *Computer vision and image understanding* 66.2 (1997) 2.

[ON94] OREN, M. and NAYAR, S. K. "Generalization of Lambert's reflectance model". *Proc. of SIGGRAPH*. 1994 1, 3.

Figure 12: *Relighting results for three different illuminations (*doge, ennis, *and* pisa*) and five different input images. Last two columns feature the same illumination under two different rotations. In each case, we show the relighted image, and the reconstructed shading and residual terms. Our model is capable of producing a compelling relighting result for a varied set of input images and illuminations, including both indoors and outdoors cases. The residual term has been scaled by a factor of 10 for visualization purposes.*

[PGM*19] PASZKE, ADAM, GROSS, SAM, MASSA, FRANCISCO, et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". *Advances in Neural Information Processing Systems 32*. 2019 5.

[Ram09] RAMAMOORTHI, RAVI. *Precomputation-based rendering*. 2009 2.

[Ram88] RAMACHANDRAN, VILAYANUR S. "Perception of shape from shading". *Nature* 331.6152 (1988) 2.

[Rem] REMOVEBG. URL: https://www.remove.bg/ 6.

[RFB15] RONNEBERGER, OLAF, FISCHER, PHILIPP, and BROX, THOMAS. "U-net: Convolutional networks for biomedical image segmentation". *Intnl. Conf. on Medical Image Computing and Computer-assisted Intervention*. 2015, 234–241 4.

[RH01] RAMAMOORTHI, RAVI and HANRAHAN, PAT. "An efficient representation for irradiance environment maps". *Proc. of SIGGRAPH*. 2001 1, 2.

[SBT*19] SUN, T., BARRON, J. T., TSAI, Y.-T., et al. "Single image portrait relighting." *ACM Trans. on Graphics* 38.4 (2019) 2.

[SK00] SHUM, HARRY and KANG, SING BING. "Review of image-based rendering techniques". *Visual Communications and Image Processing 2000*. Vol. 4067. 2000 2.

[SKCJ18] SENGUPTA, SOUMYADIP, KANAZAWA, ANGJOO, CASTILLO, CARLOS D, and JACOBS, DAVID W. "SfSNet: Learning Shape, Reflectance and Illuminance of Faces in the Wild'". *Proc. of CVPR*. 2018, 6296–6305 1–3.

[SKG*12] SINHA, SUDIPTA N, KOPF, JOHANNES, GOESELE, MICHAEL, et al. "Image-based rendering for scenes with reflections". *ACM Trans. on Graphics* 31.4 (2012) 2.

[SKS02] SLOAN, PETER-PIKE, KAUTZ, JAN, and SNYDER, JOHN. "Precomputed Radiance Transfer for Real-time Rendering in Dynamic, Low-frequency Lighting Environments". *ACM Trans. on Graphics* 21.3 (2002) 1, 2.

[Uns] UNSPLASH. URL: https://unsplash.com/ 6.

[Wei01] WEISS, YAIR. "Deriving intrinsic images from image sequences". *Proc. of ICCV*. 2001 2.

[WMLT07] WALTER, BRUCE, MARSCHNER, STEPHEN R., LI, HONGSONG, and TORRANCE, KENNETH E. "Microfacet Models for Refraction through Rough Surfaces". *Proc. of EGSR*. 2007 1, 3.

[WWG*21] WANG, QIANQIAN, WANG, ZHICHENG, GENOVA, KYLE, et al. "IBRNet: Learning Multi-View Image-Based Rendering". *arXiv preprint arXiv:2102.13090* (2021) 2.

[WYL*20] WANG, ZHIBO, YU, XIN, LU, MING, et al. "Single image portrait relighting via explicit multiple reflectance channel modeling". *ACM Trans. on Graphics* 39.6 (2020), 1–13 2.

[YGL*14] YE, GENZHI, GARCES, ELENA, LIU, YEBIN, et al. "Intrinsic video and applications". *ACM Trans. on Graphics* 33.4 (2014) 2.

[ZFT*21] ZHANG, XIUMING, FANELLO, SEAN, TSAI, YUN-TA, et al. "Neural light transport for relighting and view synthesis". *ACM Trans. on Graphics* 40.1 (2021) 2.

[ZHSJ19] ZHOU, HAO, HADAP, SUNIL, SUNKAVALLI, KALYAN, and JACOBS, DAVID W. "Deep single-image portrait relighting". *Proc. of ICCV*. 2019 2.

[ZRSK20] ZHANG, KAI, RIEGLER, GERNOT, SNAVELY, NOAH, and KOLTUN, VLADLEN. "Nerf++: Analyzing and improving neural radiance fields". *arXiv preprint arXiv:2010.07492* (2020) 2.