Total Relighting: Learning to Relight Portraits for Background Replacement

ROHIT PANDEY*, SERGIO ORTS ESCOLANO*, CHLOE LEGENDRE*, CHRISTIAN HÄNE, SOFIEN BOUAZIZ, CHRISTOPH RHEMANN, PAUL DEBEVEC, and SEAN FANELLO, Google Research

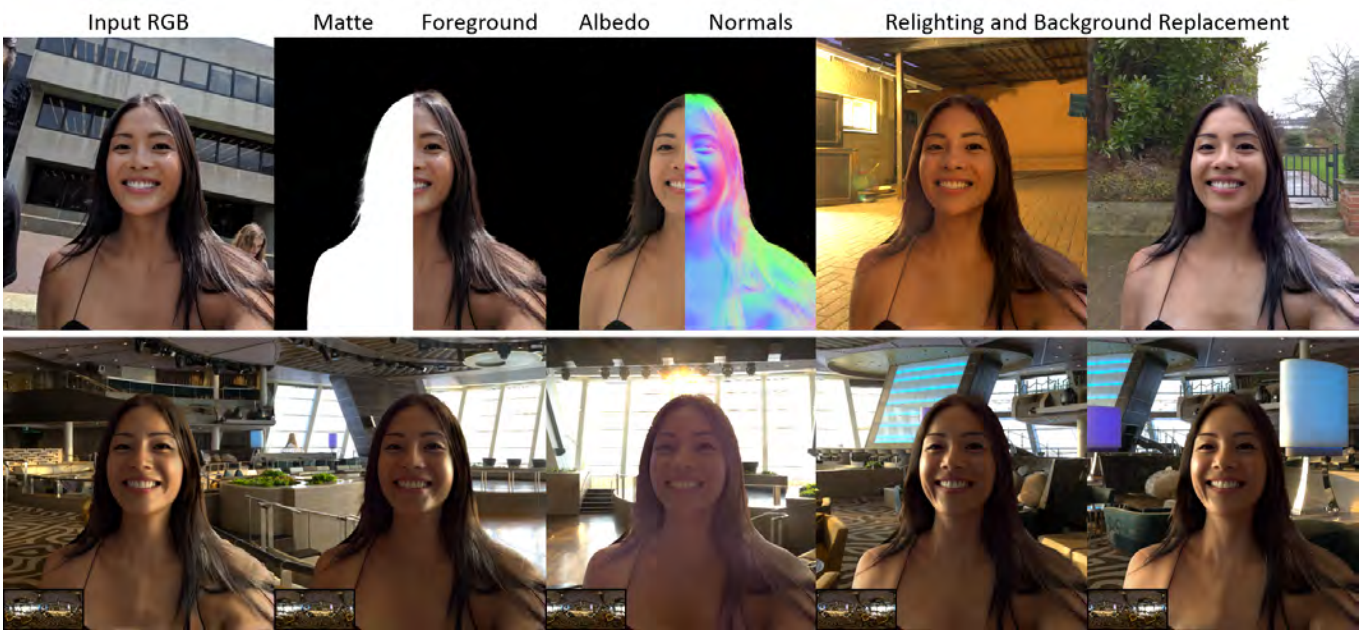

Fig. 1. Given a portrait and an arbitrary high dynamic range lighting environment, our framework uses machine learning to composite the subject into a new scene, while accurately modeling their appearance in the target illumination condition. We estimate a high quality alpha matte, foreground element, albedo map, and surface normals, and we propose a novel, per-pixel lighting representation within a deep learning framework.

We propose a novel system for portrait relighting and background replacement, which maintains high-frequency boundary details and accurately synthesizes the subject's appearance as lit by novel illumination, thereby producing realistic composite images for any desired scene. Our technique includes foreground estimation via alpha matting, relighting, and compositing. We demonstrate that each of these stages can be tackled in a sequential pipeline without the use of priors (e.g. known background or known illumination) and with no specialized acquisition techniques, using only a single RGB portrait image and a novel, target HDR lighting environment as inputs. We train our model using relit portraits of subjects captured in a light stage computational illumination system, which records multiple lighting conditions, high quality geometry, and accurate alpha mattes. To perform realistic relighting for compositing, we introduce a novel per-pixel lighting representation in a deep learning framework, which explicitly models the diffuse and the specular components of appearance, producing relit portraits with convincingly rendered non-Lambertian effects like specular highlights. Multiple experiments and comparisons show the effectiveness of the proposed approach when applied to in-the-wild images.

CCS Concepts: • **Computing methodologies → Computer vision**; **Machine learning**; **Rendering**.

Additional Key Words and Phrases: Compositing, Relighting, Neural Rendering

ACM Reference Format:
Rohit Pandey, Sergio Orts Escolano, Chloe LeGendre, Christian Häne, Sofien Bouaziz, Christoph Rhemann, Paul Debevec, and Sean Fanello. 2021. Total Relighting: Learning to Relight Portraits for Background Replacement. *ACM Trans. Graph.* 40, 4, Article 43 (August 2021), 21 pages. https://doi.org/10.1145/3450626.3459872

*Authors contributed equally to this work.

Authors' address: Rohit Pandey, rohitpandey@google.com; Sergio Orts Escolano, sorts@google.com; Chloe LeGendre, chlobot@google.com; Christian Häne, chaene@google.com; Sofien Bouaziz, sofien@google.com; Christoph Rhemann, crhemann@google.com; Paul Debevec, debevec@google.com; Sean Fanello, seanfa@google.com, Google Research.

0730-0301/2021/8-ART43
https://doi.org/10.1145/3450626.3459872

1 INTRODUCTION

Compositing a person into a scene to look like they are really there is a fundamental technique in visual effects, with many other applications such as smartphone photography [Tsai and Pandey 2020] and video conferencing [Hou and Mullen 2020]. The most common practice in film-making has been to record an actor in front of green or blue screen and use chroma-keying [Wright 2013] to derive an alpha matte and then change the background to a new one. However, this does nothing to ensure that the lighting on the subject appears consistent with the lighting in the new background environment, which must be solved with laborious lighting placement or elaborate LED lighting reproduction systems [Bluff et al. 2020; Debevec et al. 2002; Hamon et al. 2014]. Our goal is to design a system that allows for automated portrait relighting and background replacement.

There is a significant body of work both in relighting, e.g. [Barron and Malik 2015; Debevec et al. 2000; Nestmeyer et al. 2020; Sun et al. 2019; Wang et al. 2020; Zhou et al. 2019], and in determining alpha mattes and foreground colors, e.g. [Cai et al. 2019; Forte and Pitié 2020; Hou and Liu 2019; Lutz et al. 2018; Xu et al. 2017]. A few techniques simultaneously consider foreground estimation and compositing in a unified framework [Wang and Cohen 2006; Zhang et al. 2020b] and produce convincing composites when the input and target lighting conditions are similar. However, the absence of an explicit relighting step limits realism when the input and target illumination conditions are different.

To generate convincing relit composites, Einarsson et al. [2006] and Wenger et al. [2005] captured reflectance field basis images using time-multiplexed lighting conditions played back at very high frame rates ($\sim$ 1000 Hz) in a computational illumination system, leveraging image-based relighting [Debevec et al. 2000] to match the lighting of the subject to the target background. Both methods also employed a simple ratio matting technique [Debevec et al. 2002] used to derive the alpha channel, based on infrared or time-multiplexed mattes *and* recording a "clean plate". These hardware-based systems produced realistic composites by handling matting, relighting, and compositing in one complete system. However, their specialized hardware makes these techniques impractical in casual settings, such as for mobile phone photography and video conferencing.

Inspired by these approaches, we propose a system for realistic portrait relighting and background replacement, starting from just a single RGB image and a desired target high dynamic range (HDR) lighting environment [Debevec 1998]. Our approach relies on multiple deep learning modules trained to accurately detect the foreground and alpha matte from portraits and to perform foreground relighting and compositing under a target illumination condition.

We train our models using data from a light stage computational illumination system [Guo et al. 2019] to record reflectance fields and alpha mattes of 70 diverse individuals in various poses and expressions. We process the data to estimate useful photometric information such as per-pixel surface normals and surface albedo, which we leverage to help supervise the training of the relighting model. We extrapolate the recorded alpha mattes to all of the camera viewpoints using a deep learning framework that leverages clean plates of the light stage background, extending ratio matting to unconstrained backgrounds without the need for specialized lighting. With these reflectance fields, alpha mattes, and a database of high resolution HDR lighting environments, we use image-based relighting [Debevec et al. 2000] to generate composite portraits to simulate in-the-wild photographs, and we use these for training both a relighting and an alpha matting model.

While previous deep portrait relighting techniques either inject target HDR lighting into the relighting network at the bottleneck of an encoder-decoder architecture [Sun et al. 2019], or extract features from HDR lighting for modulating the features in the decoder blocks [Wang et al. 2020], we instead employ a novel, pixel-aligned, and rendering-based in-network lighting representation. This is based on the insight that U-Net architectures [Ronneberger et al. 2015] are best at leveraging extra inputs that are spatially or pixel-aligned to the original input [Isola et al. 2017].

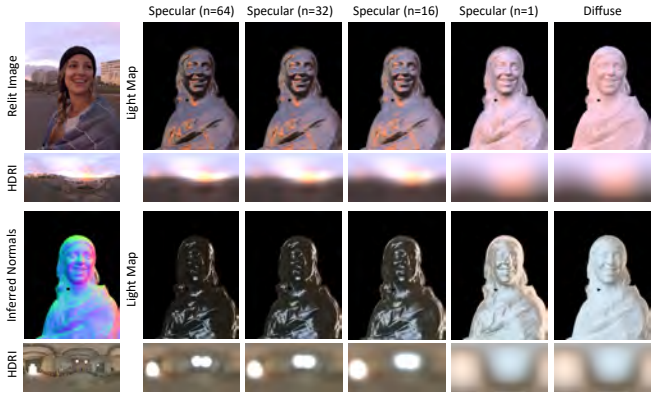

Fig. 2. We use *light maps* as a pixel-aligned lighting representation in our relighting framework. Here we show several diffuse irradiance maps and prefiltered HDR environment maps (panoramas), and their corresponding light maps computed using our framework's inferred surface normals. The light maps have been scaled for display.

We generate our per-pixel lighting representation by preconvolving the target HDR illumination with Lambertian and Phong reflectance lobes [Phong 1975] to generate a set of prefiltered environment maps with different specular exponents [Cabral et al. 1999; Greene 1986; Miller and Hoffman 1984; Ramamoorthi and Hanrahan 2001] (see filtered HDR maps in Fig. 2). Our trained model infers per-pixel surface normals, which we use as indices into the prefiltered lighting environments to form diffuse and specular *light maps* [Ramamoorthi and Hanrahan 2001] (see Fig. 2), and we inject these into the relighting model as pixel-aligned representations of the target illumination. We demonstrate through experimentation that this representation allows our relighting model to generate complex non-Lambertian reflectance while correctly inferring lower-frequency color and shading under the target illumination.

We finally demonstrate that this approach generalizes to in-the-wild portrait images, relighting and compositing subjects captured via mobile phone photography into novel backgrounds.

In summary, our main contributions are:

- A complete system – from data generation to in-the-wild inference – for portrait relighting and background replacement.

- A novel per-pixel lighting representation within a deep learning based relighting framework, which produces state-of-the-art portrait relighting results.
- Photorealistic relighting and compositing results for in-the-wild portraits, demonstrating the effectiveness and generalization of the proposed approach and the importance of high quality ground truth data.

2 RELATED WORK

Image-Based Relighting. Leveraging the linearity of light transport, relighting can be computed as a linear combination of images in different basis lighting conditions [Dorsey et al. 1995]. Debevec et al. [2000] realistically relit faces in novel HDR lighting environments from reflectance fields recorded one light at a time (OLAT) in a spherical lighting rig. This *image-based relighting* approach has been successfully used to render realistic digital actors in films [Sagar 2005], and has been applied to moving subjects [Einarsson et al. 2006; Meka et al. 2019; Wenger et al. 2005] using time-multiplexed lighting and high frame rate cameras. However, these relighting approaches require custom capture hardware which precludes using them in the context of casual photography.

Recent advances in deep learning have enabled machine learning solutions for relighting objects [Meka et al. 2018; Ren et al. 2015; Sang and Chandraker 2020; Xu et al. 2018] and people [Kanamori and Endo 2018; Meka et al. 2019; Sun et al. 2019]. Kanamori and Endo [2018] enabled inverse rendering for the human body by inferring albedo, illumination, and a light transport map that encodes occlusion. However, the results of this method were restricted to Lambertian surfaces and relighting only with low-frequency illumination. Nalbach et al. [2017] showed that appearance synthesis can be solved by learning screen-space shading, using deferred shading buffers with per-pixel scene attributes such as position, normal, and material parameters. Xu et al. [2018] trained a neural network to relight a scene under novel illumination based on a set of five jointly selected OLAT images. Meka et al. [2019] learn a full 4D reflectance field from two colored, spherical gradient illuminations captured in a light stage. While these results are compelling, the need for controllable lighting again prevents using the techniques for casually shot photos.

Portrait Relighting. Several recent works [Nestmeyer et al. 2020; Shu et al. 2017; Sun et al. 2019; Wang et al. 2020; Zhang et al. 2020a; Zhou et al. 2019] do address portrait relighting for consumer photography. Sun et al. [2019] proposed a self-supervised method to estimate an input portrait's current illumination and to relight the subject in a novel, target lighting environment. This work was the first to apply deep learning to the single image portrait relighting problem, achieving state-of-the-art results compared with earlier techniques such as the mass transport approach of Shu et al. [2017]. While Sun et al. [2019] achieved realistic results for low-frequency lighting, the network was less successful at rendering the hard shadows and specular highlights appropriate for lighting with high-frequency detail.

To better handle directional illumination, Nestmeyer et al. [2020] trained a relighting network using physics-guided supervision with per-pixel normals and albedo, a technique that we leverage in our work. However, this method required the input portrait's lighting direction to be known at inference time, and both input and output relit images were constrained to directional lighting conditions only. In contrast, our technique works for arbitrary omnidirectional input and target lighting environments. Zhou et al. [2019] developed a deep learning approach to relight in-the-wild portraits under novel spherical harmonic illumination, but the representation limited the relighting to relatively low-frequency illumination.

Leveraging the physics-based supervision approach of Nestmeyer et al. [2020], but for arbitrary input and output lighting, Wang et al. [2020] recently used synthetic renderings of 3D photogrammetry scans to supervise relighting explicitly using the diffuse and specular components of reflectance. While this method captures non-Lambertian effects to some degree, the final renderings suffer from artifacts due to the synthetic nature of the training data. Whereas this approach injects illumination into the relighting decoder using learned features concatenated along channels, we try to leverage rendering-based insights with our pixel-aligned lighting representation. In our evaluation, we show that our method outperforms both of the state-of-the-art techniques for single image portrait relighting with arbitrary inputs and target lighting [Sun et al. 2019; Wang et al. 2020], generates high-frequency self-shadowing effects and non-Lambertian effects such as specularities, and generalizes well to real-world portraits. Furthermore, in contrast to our approach, none of the portrait relighting techniques in the literature explicitly build a complete system for background replacement, which involves not only relighting, but also integration with a robust matting module.

Alpha Matting, Foreground Estimation, and Compositing. With significant progress in deep learning, many new matting methods have been proposed that improve upon the state-of-the-art results on classical benchmarks [Rhemann et al. 2009]. One such method is that of Cai et al. [2019], which we leverage in our work. This work showed the importance of an accurate input *trimap*, a partitioning of the image into a definite foreground, a definite background, and a boundary area where pixels are an unknown blend of foreground and background colors. As in Cai et al. [2019], we disentangle the matting estimation problem into two sub-tasks: trimap refinement and matting estimation, although we add foreground prediction.

Image compositing is the combining of one or more images seamlessly to form a new image. Alpha-based linear blending [Porter and Duff 1984] is possibly the simplest approach to solve this problem, although the method often blurs high frequency details around the boundaries and cannot handle complex illumination effects like refraction [Zongker et al. 1999]. More sophisticated deep-learning based methods have been recently proposed to learn a compositing function directly from data [Lin et al. 2020; Sengupta et al. 2020; Tsai et al. 2017; Zhang et al. 2020b], rather than relying on alpha-based linear blending. Despite producing excellent results, these methods fail to produce photorealistic composites when the target illumination differs substantially from the input lighting condition, leading to uncanny renderings. Tsai et al. [2017] proposed a technique to blend a given foreground in a new background using scene semantic information. While the results look visually pleasing, they mostly capture the diffuse component of the lighting and transfer

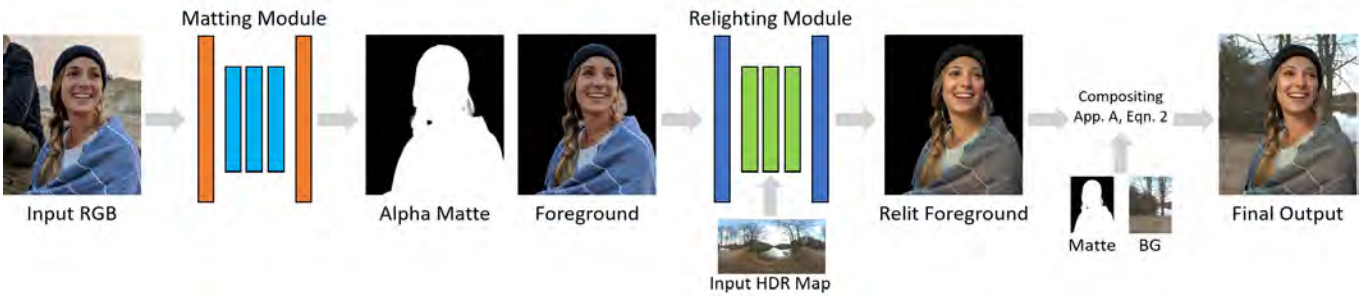

Fig. 3. Proposed Framework. Our method starts with a single portrait and estimates an alpha matte and foreground image using a deep learning module. The estimated foreground image is passed through a relighting network, which uses a target HDR lighting environment to relight the subject. Finally, a composite module produces the output rendering.

high-frequency light transport effects directly from the input image, which are inaccurate when the target illumination differs.

Our Approach. We propose a complete system for in-the-wild portrait relighting and background replacement, sequentially tackling the problems of foreground estimation, relighting, and compositing. Our method allows the portrait's subject to be relit and convincingly composited into into any HDR lighting environment (and if only a background photo is available, its HDR lighting can be estimated with a technique such as LeGendre et al. [2019]). To realistically relight the foreground subject, we propose a novel per-pixel lighting representation that models the diffuse and specular reflection components of appearance. Each module of our system is trained directly from photorealistic synthetic renderings from light stage reflectance field data.

3 FRAMEWORK

Our system (Fig. 3) consists of three sequential steps. First, a matting module estimates the alpha matte and foreground from a given RGB image. The estimated foreground and a target HDR lighting environment are then provided to a relighting module, which infers surface geometry and albedo and uses a per-pixel lighting representation to explicitly model the diffuse and specular reflection components of rendered appearance. The relit result, the alpha matte, and the new background are finally composited together, producing a relit portrait with a new background, where the lighting conditions of the portrait match that of the novel background.

3.1 Matting Module

A learned matting module is used to extract a high-quality segmentation mask as well as the color of the foreground from the input portrait. In particular, we employ a deep convolutional network and extend the state-of-the-art method of Cai et al. [2019] to predict both alpha matte and foreground color F. Details regarding the specific architecture, implementation, and training are provided in Appendix A. This matting model is trained specifically to work for portraits, rather than generic objects, using data captured in the light stage system. Compared with previous works, our matting training dataset is more realistic, as relighting allows the the illumination of the foreground subject to match the background. We show in Sec. 6.2.1 that this improves our matting model's performance.

3.2 Relighting Module

The relighting module regresses from an input foreground F to a geometry image N, encoding the per-pixel surface normals, and then to an approximate diffuse albedo image A. These intrinsic image features have been previously shown to assist in neural relighting [Nestmeyer et al. 2020; Wang et al. 2020]. Differently from previous work, we introduce a novel per-pixel lighting representation or *light maps* L, which encode the specular S and diffuse D components of surface reflection for a given omnidirectional target HDR lighting environment and inferred surface geometry. Finally, a neural shading module performs the final foreground rendering. This proposed system is shown in Fig. 4. In the following subsections, we describe the components of our full relighting module.

3.2.1 Geometry Net. The input to the relighting module is the predicted foreground F generated by the matting network, resized to our inference size of 1024×768. Our goal is to infer the geometry image N, represented as per-pixel surface normals, and the per-pixel albedo image A. Although a single architecture with multiple branches could predict these components at the same time [Nestmeyer et al. 2020; Wang et al. 2020], we found that surface normals were easier for the network to learn with high-quality ground truth supervision (see Sec. 5), consistent with recent works on single image geometry estimation [Saito et al. 2020]. Hence, the first network is used to perform image-to-image translation of the input RGB foreground to an image of surface normals using a U-Net architecture (Figure 4, upper left) with 13 encoder-decoder layers and skip connections. Each layer is run through 3×3 convolutions followed by Leaky ReLU activations and the number of filters are 32, 64, 128, 256, 512, 512 for the encoder, 512 for the bottleneck, and 512, 512, 256, 128, 64, 32 for the decoder respectively. The encoder uses blur-pooling [Zhang 2019] layers for down-sampling, whereas the decoder uses bilinear resizing followed by a 3×3 convolution for upsampling. The output of the module is a surface normal image N in camera space coordinates.

3.2.2 Albedo Net. The surface normals N and the input foreground F are concatenated to form a $1024 \times 768 \times 6$ tensor and passed as input to another U-Net with the same architecture as the Geometry Net. The output of this architecture is an image of the diffuse albedo A of the subject.

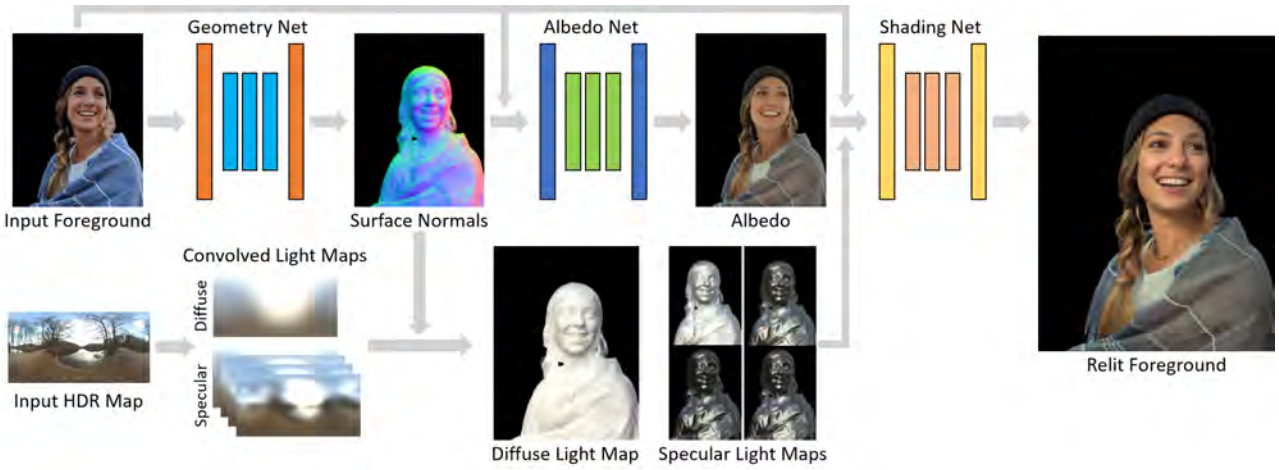

Fig. 4. The relighting module is divided into three sequential steps. A first *Geometry Network* estimates per-pixel surface normals N from the input foreground. The surface normals and foreground F are used to generate the albedo A. The target HDR lighting environment is prefiltered using diffuse and specular convolution operations, and then these prefiltered maps are sampled using surface normals or reflection vectors, producing a per-pixel representation of diffuse and specular reflectance for the target illumination (light maps). Finally, a *Shading Network* (Figure 5) produces the final relit foreground.

3.2.3 Light Maps as a Lighting Representation. Next, we represent target HDR illumination in a pixel-aligned format suitable for concatenation along channels, for input to the U-Net based shading network. Prior relighting works do not perform an explicit sampling of the light directions in the input HDR environment map based on surface geometry, despite also relying on U-Net architectures; hence they must learn a difficult mapping function of panoramic lighting image coordinates to portrait feature coordinates [Sun et al. 2019; Wang et al. 2020].

Given the input geometry, and the desire to produce appearance under a target HDR lighting environment while assuming a distant lighting model, one could envision treating each lighting pixel as a unique light source, and then integrating the shading contribution of each source for each pixel in N given its surface normal and a presumed bidirectional reflectance distribution function (BRDF). However, this approach is computationally prohibitive, especially when performed at training time for millions of images. In a similarly compute-constrained setting, real-time graphics practitioners have demonstrated that prefiltering or preconvolving an HDR lighting environment by cosine lobe functions representing Lambertian or Phong specular BRDFs allows this integration to happen offline [Greene 1986; Miller and Hoffman 1984; Ramamoorthi and Hanrahan 2001], which is useful for real-time rendering of both diffuse and specular materials. After precomputing a diffuse irradiance map and several prefiltered HDR environment maps with different Phong exponents ($n = 1, 16, 32, 64$) (examples in Fig. 2), at training or inference time, diffuse and specular reflectance images or so-called *light maps* can be easily computed by indexing into these prefiltered maps using the normal or reflection vectors. We show example light maps in Fig. 2. In our ablation study, we demonstrate that our network trained with this lighting representation outperforms those trained using prior representations. The proposed approach also provides some physically-based control over the final relit appearance, as we can artificially manipulate the diffuse and specular light maps provided to the shading network, for example to make the skin appear shinier or more matte.

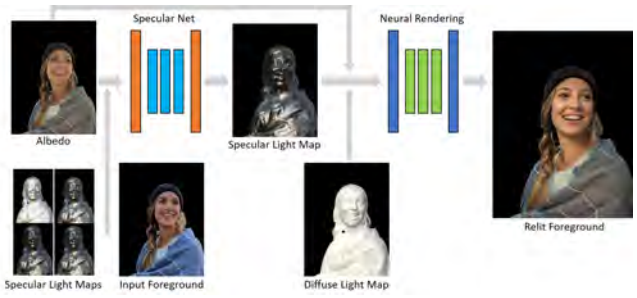

Fig. 5. Shading Network. A first *Specular Network* is used to predict a single specular light map, taking as input: multiple specular light map candidates computed using different Phong exponents, the albedo, and the input foreground (see text for details). The predicted specular light map is then concatenated with the diffuse component and the albedo and passed through a final *Neural Rendering Network* to produce the relit foreground.

3.2.4 Shading Net. This module consists of two sequential networks. The first, *Specular Net*, attempts to model the uncertainty in the material properties of the input image. It takes as input specular light maps S_n generated with multiple Phong exponents n, along with the predicted albedo A and the input foreground F. A lighter weight U-Net with 13 layers with 8, 16, 32, 64, 128, 256 filters for the encoder, 256 filters for the bottleneck, and 256, 128, 64, 32, 16, 8 filters for the decoder runs 3x3 convolutions with Leaky ReLU activations, and predicts a four channel per-pixel weight image W. Finally, a single specular light map $\hat{S}$ is obtained by taking a weighted sum of the candidates S_n using the per-pixel weights of W, i.e. given a pixel u, v: $\hat{S}(u, v) = \sum_n W_n(u, v) S_n(u, v)$. This is physically motivated as faces exhibit spatially-varying specular roughness [Debevec et al. 2000; Ghosh et al. 2010].

A final *Neural Renderer* performs the actual image synthesis. This architecture takes as input the albedo A, the diffuse light map D and the blended specular light map $\hat{S}$ and returns the final relit image. The specific architecture is a U-Net with the same architecture as the Geometry Net and the Albedo Net. We use a neural renderer to compensate for approximations employed (e.g. a relatively simple shading model) and any residual error in the predicted intermediate images. In Fig. 6, for a given input image (Fig. 6a), we show an example predicted albedo A (Fig. 6b) and diffuse light map D (Fig. 6c). Multiplying these two images together yields an image that approximates diffuse reflectance (Fig. 6d), although without self-shadowing. We also show an example specular light map S (Fig. 6f), and the result of adding the approximate diffuse reflectance and the specular light map images together, essentially shading the geometry according to the Phong model [Phong 1975], in Fig. 6g. While this Phong-shaded model is clearly not realistic, we show this overlay to demonstrate how a specular light map S supplies the neural renderer with clues about the likely location of strong highlights, given the inferred geometry. The neural renderer synthesizes high-frequency details including cast shadows and any non-Lambertian effects not captured in Fig. 6d, as shown in Fig. 6h.

Fig. 6. For an input portrait (**a**), we first infer surface normals (**e**) and albedo (**b**). Given a target HDR lighting environment and the normals, we compute the diffuse and specular light maps (**c, f**). Multiplying (**b**) by (**c**) yields approximate diffuse reflectance, without self-shadowing, while (**f**) suggests the locations of specular reflections for the new scene, as shown in the Phong-shaded subject in (**g**). Our neural relighting module learns to recover self-shadowing and specularities, as shown in result (**h**), beyond what can be rendered using the simple shading model of (**g** or **d**).

3.3 Compositing

We composite the relit foreground into a new background extracted as an oriented view into the lighting panorama, using the matting equation (See Appendix A, Eq. 2). We also trained an additional deep network to learn the compositing function directly from the data, providing it with the alpha matte, relit foreground, background, and original RGB input image, hoping that the network would learn to correct residual compositing errors. However, we experimentally observed that improvements from this approach were marginal and insufficient to justify the added compute and memory requirements.

4 LOSS FUNCTIONS AND TRAINING

Our framework predicts multiple intermediate outputs. In our experience, supervising training using ground truth imagery for intermediate quantities (e.g. normals and albedo) facilitated network convergence and led to higher quality results. We chose to supervise training using intermediate stage imagery only when it was feasible to photograph or generate high-quality, robust ground truth without relying on significant approximations or strong assumptions. In particular, we generated accurate supervision imagery for the following intermediate and final outputs: the trimap, alpha matte, foreground, albedo, surface normals, relit, and composited images.

4.1 Relighting Module Loss Functions

The relighting module minimizes the following loss terms:

Geometry L1 Loss $\mathcal{L}_{\text{geo}}$: $\ell_1(N_{\text{gt}}, N)$. The ℓ_1 loss between the ground truth surface normals N_{gt} and the predicted normals N encourages the network to learn the geometry of the subject.

Albedo VGG loss $\mathcal{L}_{\text{vgg}_{\text{alb}}}$: $\ell_2(\text{vgg}(A_{\text{gt}}), \text{vgg}(A))$. In order to preserve sharp details, we follow previous work [Martin-Brualla et al. 2018; Meka et al. 2019, 2020; Pandey et al. 2019], and use the squared ℓ_2 distance between features extracted from the target albedo A_{gt} and the predicted albedo A images using a VGG network pre-trained on the ImageNet classification task [Zhang et al. 2018].

Albedo L1 Loss $\mathcal{L}_{\text{alb}}$: $\ell_1(A_{\text{gt}}, A)$. In addition to $\mathcal{L}_{\text{vgg}_{\text{alb}}}$, we also add a small ℓ_1 loss between the ground truth albedo A_{gt} and the predicted albedo A to speed up color convergence.

Shading VGG loss $\mathcal{L}_{\text{vgg}_{\text{shad}}}$: $\ell_2(\text{vgg}(R_{\text{gt}}), \text{vgg}(R))$. Similar to $\mathcal{L}_{\text{vgg}_{\text{alb}}}$ we also use the squared ℓ_2 distance between features extracted from the target relit R_{gt} and the predicted relit R images using a VGG network pre-trained on the ImageNet classification task.

Shading L1 Loss $\mathcal{L}_{\text{shad}}$: $\ell_1(R_{\text{gt}}, R)$. Again, we add a small ℓ_1 loss between the ground truth relit image R_{gt} and the predicted relit image R to speed up color convergence.

Specular Loss $\mathcal{L}_{\text{spec}}$: $\ell_1(\hat{S} \odot R_{\text{gt}}, \hat{S} \odot R)$. Due to the lack of explicit supervision to separate the diffuse and specular components of reflection, we propose a self-supervised loss that encourages the network to preserve specular highlights and view-dependent effects. In particular, similarly to Meka et al. [2020], we compute two saliency terms as $L_1 = R \odot \hat{S}$ and $L_2 = R_{\text{gt}} \odot \hat{S}$, where $\hat{S}$ is the specular component computed following Section 3.2.4, R is the predicted relit image, R_{gt} is the ground truth relit image and $\odot$ indicates element-wise multiplication. Finally, ℓ_1 between the two saliency terms is minimized during the training. While Meka et al. [2020] compute this type of loss for a single directional light, our proposed pixel-aligned lighting representation allows us to extend this technique to the omnidirectional HDR illumination case.

Albedo Adversarial Loss $\mathcal{L}_{\text{adv}_{\text{alb}}}$: $\text{disc}_{\text{alb}}(A_{\text{crop\_gt}}, A_{\text{crop}})$. For the Albedo Net, we add an adversarial loss on the face region to help the network learn to plausibly remove high-frequency shading effects

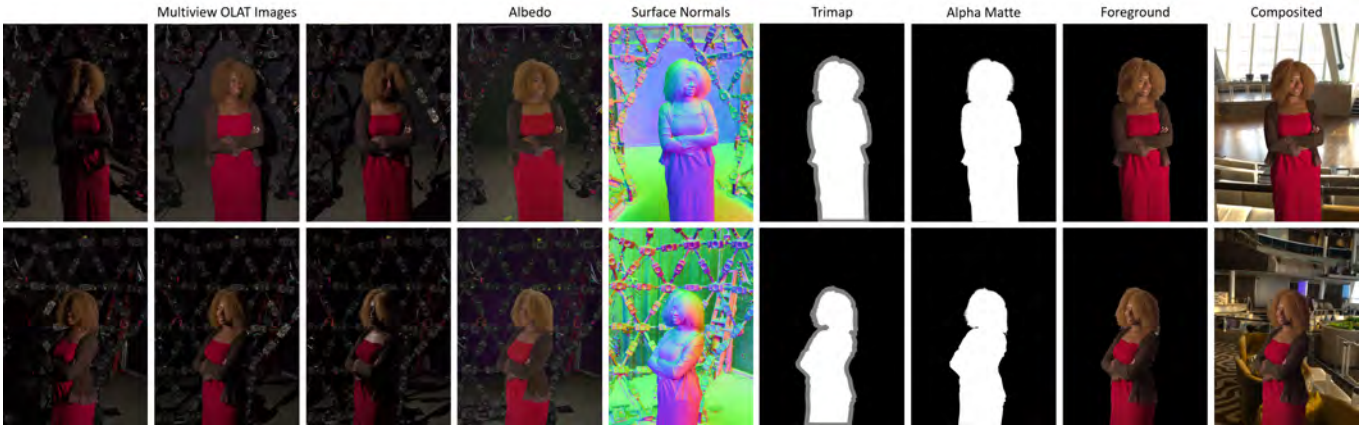

Fig. 7. Ground Truth Data Generation. To supervise training, we acquire OLAT images from multiple viewpoints using a light stage (col. 1-3). From these captures, we estimate Albedo and Surface Normals (col. 4 and 5). Albedo has been brightened for display. A deep learning framework is used to estimate an alpha matte (col. 7) from a rough segmentation trimap (col. 6). Finally, the OLAT images are linearly combined [Debevec et al. 2000] to produce relit images according to a target HDR lighting environment (col. 8). The relit foreground is then composited into a corresponding view of the panoramic lighting, using the estimated matte (col. 9).

from the input image while maintaining image detail. We use a least squares discriminator [Mao et al. 2017] disc_{alb} to add a loss between a crop of the face from the ground truth albedo $A_{\text{crop\_gt}}$ and a matching crop of the face from the predicted albedo A_{crop}.

Shading Adversarial Loss $\mathcal{L}_{\text{adv}_{\text{shad}}}$: $\text{disc}_{\text{shad}}(R_{\text{crop\_gt}}, R_{\text{crop}})$. Similarly, it is crucial for the Shading Net to synthesize realistic high-frequency shading effects on the face. We use another least squares discriminator $\text{disc}_{\text{shad}}$ to add a loss between a crop of the face from the ground truth relit image $R_{\text{crop\_gt}}$ and a matching crop of the face from the predicted relit image R_{crop}. In addition to the relit images, we also concatenate crops of the saliency terms $R_{\text{crop}} \odot \hat{S}_{\text{crop}}$ and $R_{\text{crop\_gt}} \odot \hat{S}_{\text{crop}}$ to act as an attention mechanism for the discriminator to focus on facial reflections.

The relighting module was trained end-to-end using a weighted sum of the above losses:

$$\begin{aligned}\mathcal{L}_{\text{relighting}} = \lambda_{\text{geo}} * \mathcal{L}_{\text{geo}} + \lambda_{\text{vgg}_{\text{alb}}} * \mathcal{L}_{\text{vgg}_{\text{alb}}} + \lambda_{\text{alb}} * \mathcal{L}_{\text{alb}} \\ + \lambda_{\text{vgg}_{\text{shad}}} * \mathcal{L}_{\text{vgg}_{\text{shad}}} + \lambda_{\text{shad}} * \mathcal{L}_{\text{shad}} + \lambda_{\text{spec}} * \mathcal{L}_{\text{spec}} \\ + \lambda_{\text{adv}_{\text{alb}}} * \mathcal{L}_{\text{adv}_{\text{alb}}} + \lambda_{\text{adv}_{\text{shad}}} * \mathcal{L}_{\text{adv}_{\text{shad}}}\end{aligned} \quad (1)$$

For our experiments we empirically determined these weight to be $\lambda_{\text{geo}} = \lambda_{\text{vgg}_{\text{alb}}} = \lambda_{\text{vgg}_{\text{shad}}} = 1$, $\lambda_{\text{alb}} = \lambda_{\text{shad}} = 0.1$, and $\lambda_{\text{spec}} = \lambda_{\text{adv}_{\text{alb}}} = \lambda_{\text{adv}_{\text{shad}}} = 2$.

4.2 Training Details

We implemented our training pipeline in TensorFlow, distributing the training across 8 NVIDIA Tesla V100 GPUs with 16GB of memory. Each iteration randomly picks 8 images of subjects relit with random lighting environments for both inputs and target. We use the ADAM optimizer [Kingma and Ba 2015] with a learning rate of 10^{-5}. We optimized our system for $1M$ iterations for the training to converge, taking seven days. For faster convergence and to reduce memory usage, we trained matting and relighting separately. We also trained the complete end-to-end architecture (with both matting and relighting) on Nvidia P6000 GPUs with 24GB memory, but found the improvement in the results to be marginal. We suspect that other factors, such as ground truth data quality and the overall architecture design, are more important for overall quality.

5 DATA ACQUISITION AND GROUND TRUTH GENERATION

To train our models using supervision, we require many paired portraits of different subjects lit in various lighting environments, with ground truth illumination for the target relit image. We also require the estimated foreground and alpha matte components used to generate the final composites into new scenes and the required intermediate components as previously outlined, such as per-pixel albedo and normals. To generate this data, we relight reflectance fields of a variety of different people recorded by multiple cameras in a light stage [Debevec et al. 2000]. The relit images are composited onto backgrounds using mattes derived from the light stage data using a deep learning model. This process produces more realistic training data than Wang et al. [2020], which trains on synthetic renderings of facial photogrammetry scans with approximated BRDF's. As a result, our relighting technique does a better job of reproducing complex light transport phenomena such as sub-surface scattering and spatially-varying specular reflections on skin and hair.

5.1 Reflectance Field Acquisition

Following LeGendre et al. [2020], we photographed the reflectance field (OLAT images) of 70 diverse individuals, each performing nine different facial expressions and wearing different clothing and accessories (hats, scarves, etc.). The subjects were chosen to span a wide range of ages and skin tones to support model generalization. The light stage has 331 programmable LED-based lights and 58 RGB cameras, recording video at 60 Hz with 12.4 megapixel resolution [Guo et al. 2019]. In total, we generated ~ 700 OLAT sequences, with

Fig. 8. Ratio Matting (RM) results, including an inset region to show the recovery of fine structures such as hair strands.

each sequence viewed from multiple camera viewpoints. 10% of the sequences were recorded with 58 cameras, covering the full 360° of possible vantage points to provide training examples from arbitrary viewpoints, covering the full body. We recorded the remaining 90% of the data using a subset of 6 frontal viewpoints aiming to simulate the framing of casual portrait photography. We thus recorded ∼ 7,560 unique sequences for the 58 cameras, which we then relit and composited using ∼ 200 HDR panoramic lighting environments sourced from www.HDRIHaven.com [Zaal et al. 2020] using random rotations, generating 8M training examples. For evaluation purposes, we divided the dataset into training and testing, manually selecting seven subjects with diverse skin tones for the test set, along with ten lighting environments. We show examples of generated ground truth images in Fig. 7.

5.2 Matting and Foreground Acquisition

We next describe our method to calculate accurate alpha mattes for the captured subjects.

5.2.1 Classical Ratio Matting. We directly measured ground truth alpha mattes for the two most frontal camera viewpoints in the light stage using the *ratio matting* technique [Debevec et al. 2002; Wenger et al. 2005]. This method works by recording an image of the subject silhouetted against an illuminated background (in our case, a flat grey cloth) as one of the lighting conditions in the OLAT data. In addition, we record an OLAT in the light stage without the subject after each session, which includes a clean plate of the illuminated background. The silhouetted image, divided by the clean plate image, provides a ground truth alpha channel. The background cloth is not illuminated while the rest of the OLAT sequence is captured, but some of the OLAT lighting spills onto it. The clean plate OLATs tell us how much background spill light B there is for each lighting direction, so we can use alpha and B to compute the foreground color F for each OLAT image using the matting equation (Appendix A, Eq. 2). Fig. 8 shows alpha mattes obtained using this technique, with insets that show fine details in regions with hair strands.

5.2.2 Extending Ratio Matting with Deep Learning. Unfortunately, only two cameras in the light stage see the subject against the grey backing cloth at the back of the stage. The majority see the subject

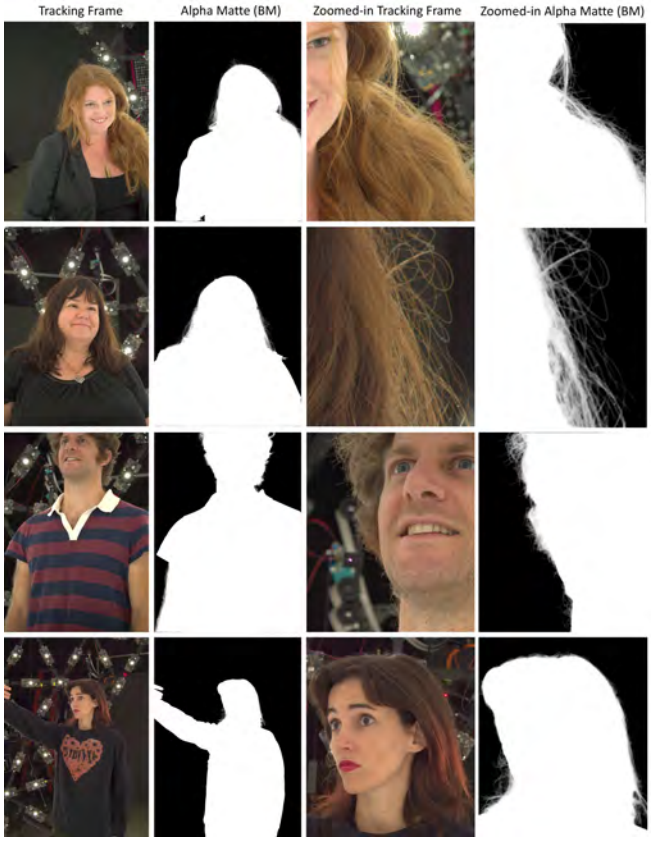

Fig. 9. Proposed Background Matting (BM) results in the light stage, with inset regions showing the recovery of fine structures.

in front of the struts and wires and cameras and light sources of the apparatus, as shown in the first column of Fig. 9. To generate alpha mattes for these other viewpoints, we make use of the clean plate image for each such view. Next, similarly to Sengupta et al. [2020], we trained a deep learning based alpha matting model that takes as inputs the clean plate (the cluttered background image B), a coarse segmentation mask computed using an off-the-shelf segmenter [Chen et al. 2018b], and the input image, and infers an alpha matte. The specific architecture and training procedure is the same used in Lutz et al. [2018]. Note that this ground truth alpha generating model is different from our in-the-wild alpha matting model of Appendix A, since during the ground truth capture, the clean plate B is known and supplied as additional input to the network.

We trained this ground truth alpha generation model with supervision, using a dataset created with our high quality mattes obtained from the frontal cameras with ratio matting. To simulate the cluttered backgrounds for this dataset, we composited foregrounds obtained from the frontal camera viewpoints over the cluttered clean plate images B acquired for the other non-frontal views using the matting equation (Appendix A, Eq. 2). Thus we were able to generate a training dataset of images with ground truth α that mostly represented the light stage imagery captured for our non-frontal viewpoints, where ratio matting was impossible (see Fig. 9). To make

the model robust to slight misalignments between the cluttered clean plate and input images, we added slight spatial perturbations to the backgrounds during training, and added background images with slight adjustments (say, for example, including clean plates captured across different days, where light stage cabling could subtly move in the field-of-view). We also used standard data augmentation techniques to improve generalization (e.g. cropping, rotation, adjustments in exposure, and adding Gaussian noise).

5.3 Albedo and Geometry Acquisition

To generate per-pixel surface normals, we follow the technique of Wenger et al. [2005], which is based on solving an over-determined system of linear equations at each pixel with photometric stereo [Woodham 1989]. Using all 331 OLAT images, we first convert the images to grayscale and, for a given pixel location (u, v), we sort all the pixel intensities across the images representing different lighting directions. As there are many more equations than unknowns with such a large lighting basis, we can discard pixels that represent the lowest 50% of values, which are likely to be noisy or in shadow, and the top 10%, which are likely to be specular reflections. This increases the chances that the pixel values used for photometric stereo represent unoccluded observations of the Lambertian component of the surface reflectance as required for classical photometric stereo. Fig. 7 shows examples of per-pixel surface normal images generated using this technique.

Although the photometric stereo equation also yields per-pixel estimates of diffuse albedo, we decided to use an image of the subject in flat omnidirectional lighting instead. Such images are readily available as the tracking frames used to align the OLAT sequences, and include the useful shading cue of ambient occlusion.

5.4 Ground Truth Compositing

By leveraging the reflectance field for each subject and the alpha matting achieved with our ground truth matte generation system, we can relight each portrait according to a given HDR lighting environment. We composite relit subjects into backgrounds corresponding to the target illumination following the matting equation (Appendix A Eq. 2). The background images are generated from the HDR panoramas by positioning a virtual camera at the center of the panorama, and ray-tracing into the panorama from the camera's center of projection with super-sampling. We ensure that the projected view into the panorama matches its orientation as used for relighting. We also use only high-resolution panoramas (16k resolution) to ensure sharp features for the background imagery. We use virtual cameras with different focal lengths to simulate the different fields-of-view of consumer cameras. Fig. 7 (right) shows several composite training images made with this process.

6 EVALUATION

In this section, we analyze the performance of our system by comparing with prior state-of-art approaches, and justify our design decisions through an ablation study. As described in Sec. 5, we manually selected seven subjects with diverse skin tones to be held out from training for evaluation purposes, along with ten lighting environments. To assess how well our *complete* system generalizes to real world imagery, we run our full pipeline on portraits captured in-the-wild under arbitrary illumination conditions. All results in this section are on subjects unseen during training.

6.1 Comparisons with State-of-the-Art: Relighting

We compare the relighting module of our framework with the two most recent state-of-the-art approaches for single image portrait relighting: Sun et al. [2019] and Wang et al. [2020]. Both of these approaches demonstrated superior performance compared with earlier non deep learning based techniques; hence we use these two methods as our baselines. Furthermore, both methods require a single RGB input image and target HDR lighting environment as input, as in our approach, though each uses a different lighting representation. While the approach of Wang et al. [2020] relies on regressing to intrinsic image components such as geometry, Sun et al. [2019] treats the entire rendering process as a black box. As the method of Wang et al. [2020] is designed to work on a crop of the face region, we used this region-of-interest for all methods including ours for a fair comparison.

For the comparison with Wang et al. [2020], the authors generously computed results for test images that we provided, as this approach requires ground truth labels for various passes from a synthetic renderer, unavailable for our light stage data. While the authors of Sun et al. [2019] also generously computed results for our provided evaluation dataset, we found in practice that their model retrained on our larger dataset demonstrated superior performance, so we refer to our implementation for a fair comparison.

Fig. 10. Qualitative comparisons between our method and the previous state-of-the-art relighting techniques [Sun et al. 2019; Wang et al. 2020]. We use our evaluation set of light stage subjects not seen during training for whom we can generate ground truth relit images. The first column has been brightened for display.

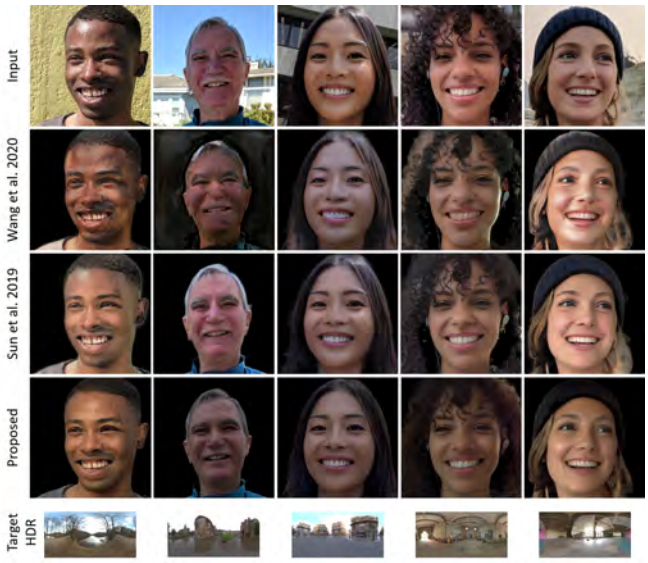

Fig. 11. Qualitative comparisons between the proposed and previous state-of-the-art relighting methods [Sun et al. 2019; Wang et al. 2020] on in-the-wild portraits.

For the evaluation subjects photographed in the light stage, we have ground truth relighting results, enabling both qualitative and quantitative comparisons among the proposed and state-of-the-art techniques. We show qualitative results in Fig. 10. The proposed method outperforms the previous state-of-the-art methods across a diverse set of subjects, demonstrating increased photorealism. In particular, when compared with Wang et al. [2020], our approach is able to accurately relight subjects of different skin tones, whereas the relighting of Wang et al. [2020] fails for subjects whose skin tones were not well-represented in the training data (see Figure 10, far left and far right). The method of Sun et al. [2019] has particular challenges with both synthesizing and removing high-frequency details like specular highlights and hard shadows, as its simple encoder-decoder style architecture does not leverage each subject's intrinsic properties e.g. albedo and geometry.

The quantitative evaluation is presented in Table 1. For these experiments, we computed multiple metrics to assess the quality of image rendering. We compute the mean absolute error (MAE), defined as the ℓ_1 distance between the predicted relit image and the ground truth relit image, the mean squared error (MSE), the structural similarity index measure (SSIM) [Wang et al. 2004], and finally a perceptual loss (LPIPS, the Learned Perceptual Image Patch Similarity metric [Zhang et al. 2018]). To limit the comparison to relighting quality only, all errors are computed only on the foreground and ground truth masks which are used for all the methods for a fair comparison. The proposed approach outperforms the previous state-of-the-art-techniques on every metric for the portrait relighting task.

We also compare across methods for portraits photographed in-the-wild under arbitrary illumination conditions, with qualitative results shown in Fig. 11. As with the evaluation dataset, our approach is able to relight subjects of diverse skin tones, while the approach of Wang et al. [2020] generates image artifacts for skin tones not observed in training. Once again, the method of Sun et al. [2019] is often unable to remove existing or add novel high-frequency detail like specular highlights to the portrait. Our method is particularly effective at removing harsh specular highlights from the input image (Figure 11, first column) and generalizes well to in-the-wild images.

Table 1. Quantitative evaluations on test images with ground truth. We compare our method and state-of-the-art methods for portrait relighting.

	Proposed	Wang et al. [2020]	Sun et al. [2019]
MAE ↓	**0.0309**	0.0907	0.0691
MSE ↓	**0.0028**	0.0122	0.0104
SSIM ↑	**0.9394**	0.5639	0.8708
LPIPS ↓	**0.0686**	0.1814	0.1111

6.2 Comparisons with State-of-the-Art: Matting

To validate the need for our custom portrait matting module, we compare our approach with the best-performing state-of-art approaches that are also available as pre-trained models. In particular, we selected the methods of Li and Lu [2020] and Xu et al. [2017], which both rank among the top-performing methods in the popular "alphamatting.com" benchmark [Rhemann et al. 2009]. Quantitative results for our portrait dataset with ground truth are reported in Table 2. We provided the same trimap for each method, computed following the procedure of Appendix A.1. Our proposed method specifically trained on portraits out-performs both pre-trained approaches. We demonstrate qualitative results for this comparison in Figure 13, where our method is able to recover sharper boundaries and fine details, and therefore more accurate alpha mattes. We also observed that existing pre-trained models often generated artifacts around the silhouette boundary due to color similarities between the foreground and background regions.

6.2.1 Relighting for Alpha Matting. As a byproduct of our proposed approach, we analyze the importance of data augmentation when training an alpha matting estimation model. In particular, we assess the effect of using accurate relighting when generating a large dataset of realistic composited images. Previous related works rely on composited datasets that do not model the illumination of the target scene, instead using plain matting into random backgrounds, e.g. [Xu et al. 2019]. We show that a more realistic training dataset including relighting improves matting estimation. To do so, we trained our matting network on the generated dataset (Section 5.4) and compared with a network trained on the same data *without* relighting the subjects to match the target scenes, instead simply using a "fully lit" image for the foreground. Additionally, we trained another version of this model, using a popular color transfer technique [Reinhard et al. 2001] to harmonize foreground colors to the target background before the final composition. A similar approach has been recently successfully applied for data augmentation when training deep learning-based stereo matching algorithms [Song et al. 2020; Tankovich et al. 2021]. We compare matting results for these three models in Fig. 12, with quantitative results in Table 2. To evaluate performance, we report the Mean Squared Error, SAD,

and Gradient [Rhemann et al. 2009] on the unknown pixels of the evaluation images (ground truth trimaps). The SAD and gradient losses are scaled by 1000 due to the large resolution of the images used in our experiment. As shown in Table 2 and Fig. 12, applying a color transfer technique on the fully lit foregrounds improves the performance of the matting model, but it still produces worse results compared to our model trained on a more realistic dataset where the foreground illumination is consistent with the background imagery.

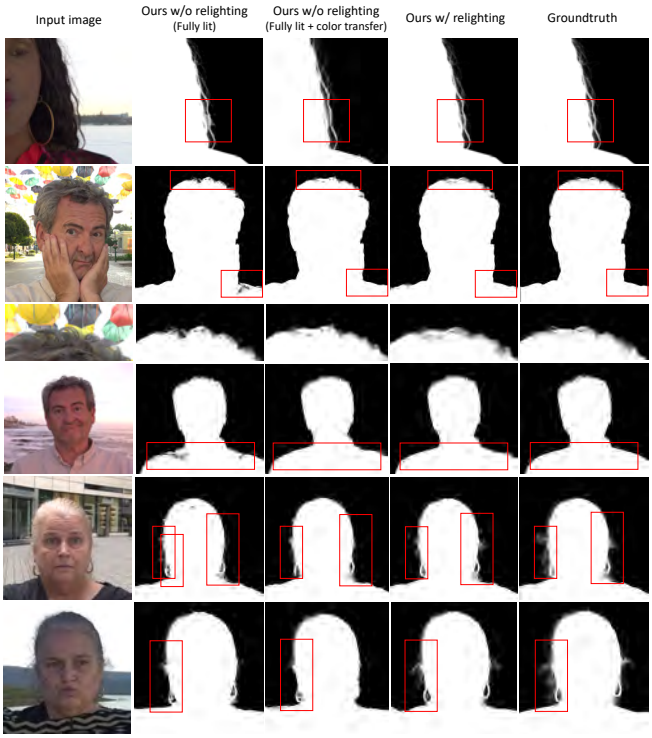

Fig. 12. Relighting for alpha matting estimation, qualitative comparison. Training the matting estimation model on our realistic ground truth composites (relighting dataset, Section 5.4) improves model performance on evaluation images with arbitrary lighting conditions.

These results suggest that: ① training a matting model using imagery with illumination harmonized to the target background improves matting model performance and generalization to arbitrary illumination conditions at test-time, especially for images with illumination substantially different from a simple, flatly-lit portrait. ② Leveraging off-the-shelf, pre-trained matting modules for background replacement in portraits would lead to sub-optimal results, since most of these models are trained on only a few natural images which are composited on random backgrounds (minor foreground color variability and often captured using flat lighting).

To further validate this, we evaluated the impact of using an off-the-shelf state-of-art matting module on the overall relighting and background replacement task. In particular, we selected Xu et al. [2017] as the top performing method on portraits based on our quantitative evaluations and combined with our relighting module to generate composited images. In Figure 14 we show the comparison with our full system. Note how this off-the-shelf model often suffers from artifacts around the subject, breaking the overall realism.

Table 2. Ablation Study and Comparisons: Relighting for Alpha Matting Estimation. We report the Sum of Absolute Differences (SAD), Mean Squared Error (MSE) and the Gradient error (Grad) [Rhemann et al. 2009] on our evaluation split. The SAD and gradient losses are scaled by 1000 due to the large resolution of the images used in this experiment (2048 × 1504).

Matting Ablation Study and Comparisons			
	SAD ↓	MSE ↓	Grad ↓
Closed-Form Matting. Levin et al. [2007]	23.3697	0.1241	4.0672
Pre-trained GCA Matting [Li and Lu 2020]	22.0385	0.1213	3.3602
Pre-trained Deep Image Matting [Xu et al. 2017]	21.4093	0.1171	3.3338
Ours (trained w/o relighting dataset, fully lit)	18.3489	0.0856	2.9378
Ours (trained with color transfer [Reinhard et al. 2001])	17.9895	0.0817	2.8514
Ours (trained w/ relighting dataset)	**15.5190**	**0.0764**	**2.4292**

6.3 Ablation Study

In this section we analyze the individual components of the proposed framework to justify our design decisions.

6.3.1 Light Representation. To prove the effectiveness of our pixel-aligned lighting representation as a standalone feature that improves the relighting capability of our neural renderer, we fix our network architecture, losses, and the training set and employ different alternatives only for the lighting representation. In particular, we compare with the approach of Sun et al. [2019], where a low resolution HDR lighting panorama is injected into the bottleneck layer of the relighting network, and with the Light guided Feature Modulation (*LFM*) approach of Wang et al. [2020], which extracts scales and offsets from the HDR map using a fully connected network for modulating the decoder features of the relighting network. The selected competing light representations are both re-trained using our training set. We show qualitative results in Figure 15. Even when trained on the same dataset, our lighting representation allows for more accurately rendered specular highlights, while also preserving sharp details. This experiment also underlines that relying solely on high-quality training data is not sufficient to obtain state-of-art results and that the light representation is a key component of our framework. Quantitative results for this ablation study are shown in Table 3.

6.3.2 Use of Specular Light Maps. To isolate the effect of our specular light maps, we trained another model without providing any such maps to the network. Rendering non-Lambertian effects, then, would thus be framed as a residual learning task as in Nestmeyer et al. [2020]. We also evaluated a variant where we removed *Specular Net*, instead supplying the network with a single specular light map S with specular exponent $n = 16$. We show qualitative results in Fig. 16. The blended specular light map helps guide the network towards generating non-Lambertian effects, while using a single specular light map, or using no specular lights maps at all, leads to overly-smooth results with missing high-frequency specularities, as shown Fig. 16. In this easy example, where the target illumination is not all that dissimilar from that of the input image, the models without our full approach cannot synthesize specularities, such as on the subject's nose. Quantitative results for this ablation are also

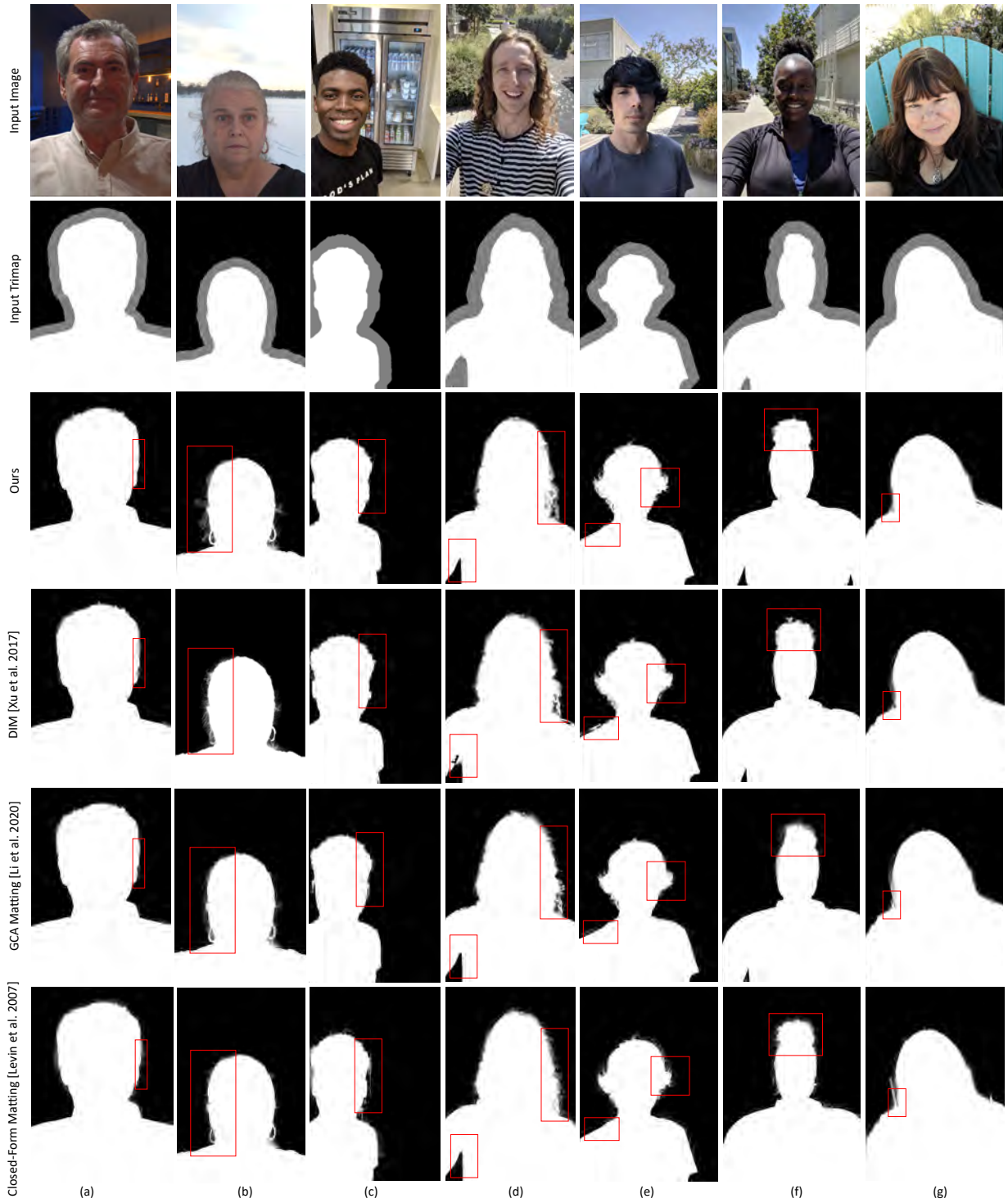

Fig. 13. (a,b): Estimated alpha mattes for evaluation images made by compositing light stage images. (c-g): Estimated alpha mattes for in-the-wild images. Note how the proposed matting module better generalizes on in-the-wild-images when compared with state-of-art pre-trained models using Adobe-1k and alphamatting.com datasets.

Fig. 14. Comparison with an off-the-shelf state-of-art matting module of Xu et al. [2017]. This experiment shows the need of a custom matting network trained specifically on portraits to achieve sufficient realism for the background replacement task.

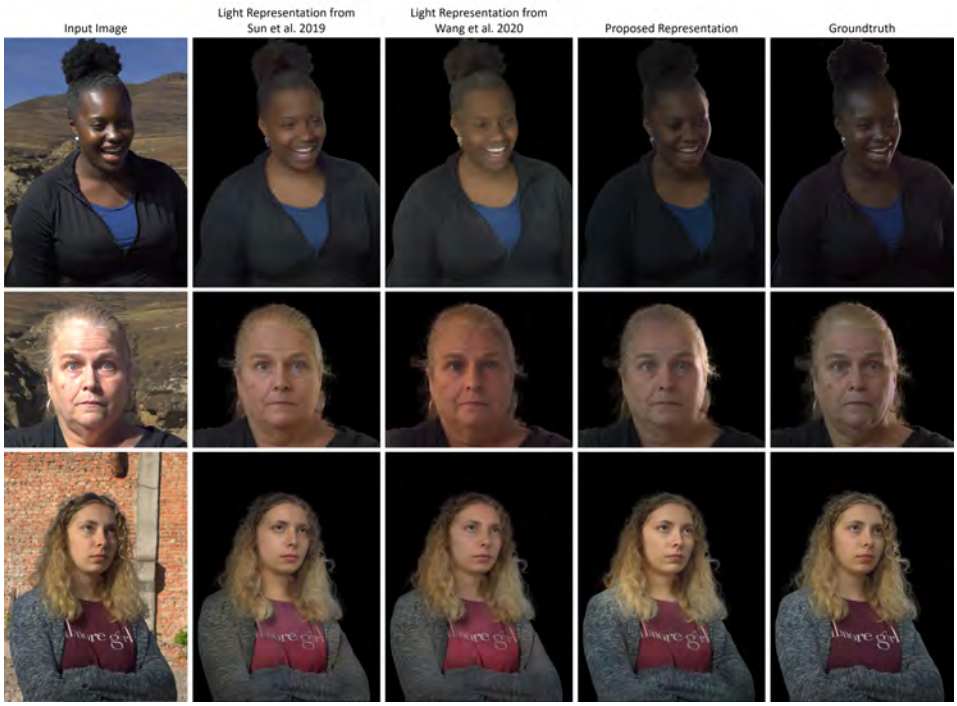

Fig. 15. Relighting results with the proposed light representation and comparisons with those of Sun et al. [2019] and Wang et al. [2020]. Our approach better captures specular highlights and self-occlusions.

shown in Table 3, where our main network slightly outperforms these other baselines.

6.3.3 Adversarial Loss. The adversarial loss on the face region allows the network to learn to generate plausible, more photorealistic

Fig. 16. Specular Light Maps Ablation. The proposed model preserves specular highlights when compared with a network trained to predict the residual from just the diffuse component ("No Specular Light Maps") or the use of a single specular light map.

Table 3. Ablation Study: Quantitative Evaluation for Relighting Module. We compare relighting results when using different lighting representations and when various components of our system are removed.

Ablation Study				
	MAE ↓	MSE ↓	LPIPS ↓	SSIM ↑
Proposed	**0.0176**	**0.0013**	**0.0802**	**0.9601**
Light rep. of Sun et al. [2019]	0.0347	0.0049	0.0979	0.9355
Light rep. of Wang et al. [2020]	0.0293	0.0034	0.0961	0.9416
No Adversarial Face Loss	0.0289	0.0033	0.1069	0.9371
Single Specular Light Map	0.0184	0.0014	0.0811	0.9598
No Specular Light Maps	0.0225	0.0020	0.0868	0.9502

Fig. 17. The adversarial loss is an important component to preserve photo-realism in the synthesized images.

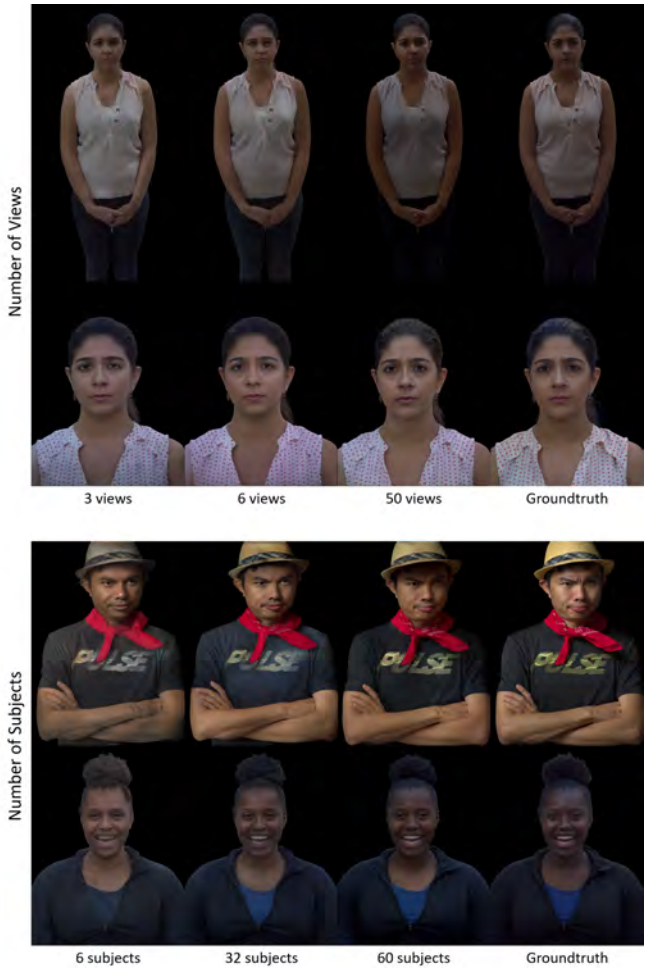

Fig. 18. Relighting results for models trained with a different number of viewpoints (top) and subjects (bottom). Increasing the number of camera viewpoints slightly improves the overall quality, especially for full body portraits. A large and diverse training dataset with many photographed individuals improves model generalization and relighting results.

facial imagery. In Fig. 17, we show relighting results obtained using our full model (bottom left) and another model trained without the adversarial loss (bottom right). For challenging input images as in Fig. 17 (top left), the network trained without the adversarial loss struggles to remove bright specular highlights and hard shadows. This performance is also demonstrated through our quantitative evaluation in Table 3.

6.3.4 Dataset Size. Here we assess the effect of the dataset size. In particular we evaluate the quality improvements with respect to the number of viewpoints (i.e. cameras) and the number of photographed subjects. Qualitative results are shown in Fig. 18, where we trained the system with 3, 6 or 50 cameras and 6, 32 or 60 subjects. The proposed algorithm appears to be moderately robust to the number of views and, although the results improve when more viewpoints are added, the quality seems acceptable even when trained on just 3 close-up cameras ($\sim$ 30 degrees apart from each other). On the other hand, decreasing the number of photographed subjects used to train the model degrades the quality of the final relit images. A large and diverse set of individuals photographed with different apparel seems to be necessary for generalization, as also discussed in the limitation section (Sec. 8).

6.4 Additional Qualitative Evaluation

We conclude the evaluation section by showing additional capabilities of the proposed system, such as intermediate outputs and relighting using directional light sources.

6.4.1 Intermediate Outputs. Whereas our final goal is image relighting and compositing, our network predicts multiple intermediate outputs as shown in Fig. 19 on a diverse set of in-the-wild portraits. Despite the very challenging input lighting conditions, our approach estimates robust alpha mattes, albedo images, and surface normals. See also the supplementary video for additional results on live action sequences.

6.4.2 Directional Light Prediction. We also render one of the evaluation subjects as illuminated by directional light sources, generating HDR panoramic lighting environments to approximately match the positioning of the lights within our light stage. Essentially, here we are using our framework to synthesize OLAT images. Single light sources can be used to emphasize complex light transport effects such as specular highlights and subsurface scattering, which are crucial to achieving true photorealism. We show qualitative results in Figure 20, comparing the predicted images with the ground truth OLAT images acquired in the light stage. Our approach synthesizes both diffuse and specular components and learns self-shadowing directly from the data. The model, however, does not accurately reproduce specular highlights in the eyes (see limitations, Section 8).

7 APPLICATIONS

Computational Photography. The most general application of our technique is to perform relighting and background replacement for portraits captured in-the-wild. We show several examples in Fig. 21, where we selected three subjects (top row) and then applied our technique to composite the subjects into multiple scenes. The method can also handle dynamic lighting environments, i.e. where the lighting environment rotates around the subject, yielding consistency and stability across frames. This is demonstrated in Fig. 22 and Fig. 1 (bottom row). The results show that the relighting network produces realistic and plausible diffuse and specular reflections, can simulate plausible rim lighting along the edge of the face, and can reconstruct diffuse skin tones obscured by specular reflections in the source images. The network also simulates a version of the veiling glare one might expect to see in a backlit photo (see the middle rendering in Fig. 22), since the light stage training data includes natural glare from lights in the back of the stage. The supplementary video contains additional results for a diverse set of subjects.

Live-Action Compositing. Although the full approach is designed to operate on still images, we can also apply our technique to videos, as demonstrated in our supplementary video. Despite the per-frame computation, with no explicitly-modeled temporal consistency, our approach produces overall accurate compositing of the moving subject, with occasional temporal inconsistencies in the predicted foreground. Temporal considerations are discussed further in Section 8, along with potential mitigation strategies.

Any Image Can Be Your Background. Our approach assumes that an HDR lighting environment corresponding to the desired background imagery is available. We achieved this in practice by generating background plates via perspective projection of high-resolution, HDR panoramas. However, this assumption somewhat limits the applicability of our approach, because it cannot be used in conjunction with in-the-wild backgrounds, where illumination is typically unknown. To relax this requirement, we combine our method with LeGendre et al. [2019], which estimates the illumination from any arbitrary image with a field-of-view similar to smartphone video. In Fig. 23, we show how even with the approximated lighting, we can produce compelling composites.

Portrait Lighting Transfer. In another application, we transfer the lighting from one portrait to another. We use the method of LeGendre et al. [2020] to estimate the illumination from a first portrait, and then apply the illumination to a new portrait, with believably consistent results. Example results are shown in Fig. 24.

Material Editing. The proposed per-pixel lighting representation offers some control over the material properties of the subject during neural rendering. To demonstrate this effect, we artificially adjust the Phong exponents used to prefilter the HDR lighting environments at inference time (not during training). An example of this is demonstrated in Fig. 25; an application of this technique is portrait shine removal.

8 LIMITATIONS

Although our proposed relighting model consistently outperforms previous works in relighting, it has limitations, as illustrated in Fig. 26. First, the albedo inference can be inaccurate for clothing, possibly due to the lack of enough diversity of apparel in the training set (Fig. 26, first row). This rarely happens for skin regions, however, as skin albedo colors form a strong prior, belonging to a relatively small subset of possible pixel values. In contrast, the color space of clothing is largely unconstrained. One potential mitigation strategy could be to leverage a semantic mask to guide the network to distinguish clothing from skin. We note, however, that our approach is the first to attempt to relight portrait regions beyond the face and hair.

Additionally, although our relighting model generates compelling non-Lambertian effects such as specularities on the face, it is not able to synthesize specular reflections from the eyes of the subject (Fig.

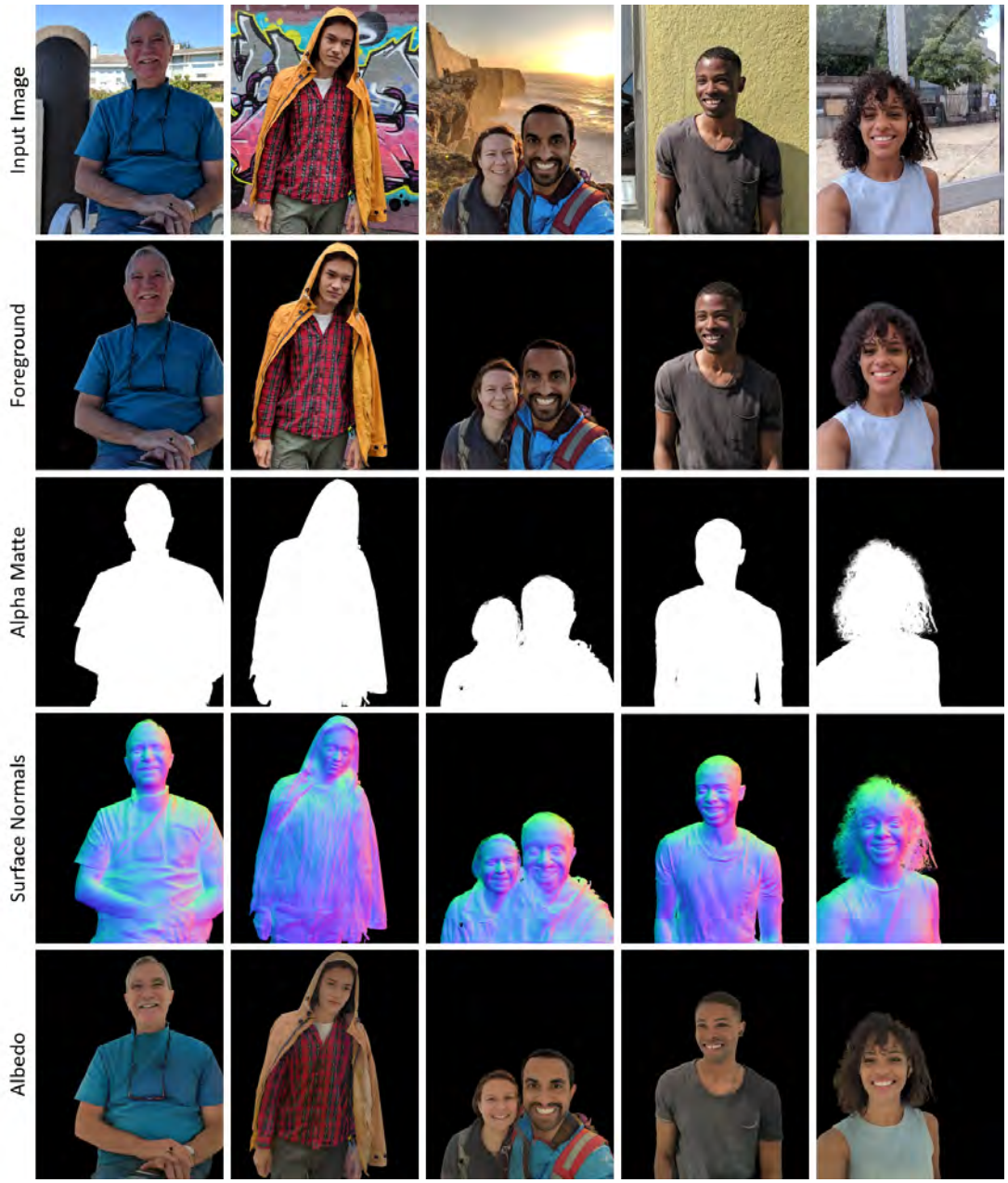

Fig. 19. Intermediate Outputs for in-the-wild photographs. While the final goal is background replacement with convincing relighting, our framework also infers intermediate components such as foreground, alpha matte, surface normals and albedo.

26, middle row). This is likely due to the lack of explicit supervision on the eyes, which contribute to a small portion of the loss functions. Although the angular spacing between the lights for the light stage system that we use aliases the BRDF for highly reflective surfaces such as eyes, prior work has shown that neural light source interpolation is possible [Sun et al. 2020]. Future work could explicitly supervise this region of interest, perhaps leveraging ground truth specular/diffuse separation from polarization difference imaging [Ghosh et al. 2010; Ma et al. 2007].

Furthermore, our approach operates per-frame, which can leads to temporal instabilities, especially in alpha matte prediction. In Fig. 26, bottom row, we show three consecutive frames with different errors in the predicted foreground. Improving temporal consistency might be as simply as giving an additional input to the matting network, such as the predicted matte from the previous frame or multiple

Fig. 20. Single directional light prediction. Our method learns to model self-occlusions, shadows, and spatially-varying specular reflections.

video frames. Our model will also require considerable optimization to perform efficiently enough for real-time applications.

Finally, our approach relies on an input HDR lighting environment. Although we have demonstrated results for a background image *without* paired illumination leveraging existing models for unconstrained lighting estimation [LeGendre et al. 2019, 2020], we expect that addressing the full problem in an end-to-end manner would provide the best possible results.

9 CONCLUSION

We have proposed a complete system – from data capture to model training – used to perform portrait relighting and background replacement given only a single in-the-wild RGB portrait and a new target HDR lighting environment as input. Our matting approach maintains high-frequency details around foreground/background boundaries and our relighting model accurately models the subject's appearance as they are composited in the new scene, with consistent illumination. To form convincingly relit composites, we introduced a novel, physically-based and pixel-aligned lighting representation used for training our relighting module. Experiments on in-the-wild images demonstrate that our relighting model can convincingly render non-Lambertian effects including subsurface scattering and specular reflections, outperforming state-of-the-art techniques for portrait relighting.

Fig. 21. Application: Relighting and compositing for casual photography. Given in-the-wild input portraits (row 1), we relight and composite the subjects into novel backgrounds with consistent illumination (rows 2-4).

REFERENCES

Jonathan T. Barron and Jitendra Malik. 2015. Shape, Illumination, and Reflectance from Shading. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 8 (2015).

Richard Bluff, Landis Fields, Abby Keller, Hayden Jones, and Rachel Rose. 2020. ILM Presents "This is the Way" — the Making of Mandalorian. In *ACM SIGGRAPH 2020 Production Sessions (SIGGRAPH 2020)*. Association for Computing Machinery, New York, NY, USA, Article 6, 1 pages. https://doi.org/10.1145/3368850.3383439

Brian Cabral, Marc Olano, and Philip Nemec. 1999. Reflection Space Image Based Rendering. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '99)*. ACM Press/Addison-Wesley Publishing Co., USA, 165–170. https://doi.org/10.1145/311535.311553

Shaofan Cai, Xiaoshuai Zhang, Haoqiang Fan, Haibin Huang, Jiangyu Liu, Jiaming Liu, Jiaying Liu, Jue Wang, and Jian Sun. 2019. Disentangled Image Matting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Guanying Chen, Kai Han, and Kwan-Yee K. Wong. 2018a. TOM-Net: Learning Transparent Object Matting From a Single Image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. 2018b. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*. 801–818.

Paul Debevec. 1998. Rendering Synthetic Objects into Real Scenes: Bridging Traditional and Image-Based Graphics with Global Illumination and High Dynamic Range Photography. In *Proceedings of the 25th Annual Conference on Computer Graphics*

Fig. 22. Application: Relighting and compositing for casually-shot photos. Our technique enables photorealistic rendering of the subjects in the first column into novel scenes. By simply *rotating* the target HDR lighting environment, a photographer can create a suite of compelling new portraits, using light to contour and highlight the subject's face. The network successfully removes a bright spot of light on the cheek in the top row, simulates plausible rim lighting effects with apparent Fresnel gain in the middle row, and reconstructs diffuse skin tone under the broad specular reflection on the forehead in the bottom row.

and Interactive Techniques (SIGGRAPH '98). Association for Computing Machinery, New York, NY, USA, 189–198. https://doi.org/10.1145/280814.280864

Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the Reflectance Field of a Human Face. In *Proceedings of SIGGRAPH 2000 (SIGGRAPH '00)*.

Paul Debevec, Andreas Wenger, Chris Tchou, Andrew Gardner, Jamie Waese, and Tim Hawkins. 2002. A lighting reproduction approach to live-action compositing. *ACM Transactions on Graphics (TOG)* 21, 3 (2002), 547–556.

Julie Dorsey, James Arvo, and Donald Greenberg. 1995. Interactive Design of Complex Time-Dependent Lighting. *IEEE Comput. Graph. Appl.* 15, 2 (March 1995), 26–36. https://doi.org/10.1109/38.365003

Per Einarsson, Charles-Felix Chabert, Andrew Jones, Wan-Chun Ma, Bruce Lamond, Tim Hawkins, Mark Bolas, Sebastian Sylwan, and Paul Debevec. 2006. Relighting Human Locomotion with Flowed Reflectance Fields. In *Proceedings of the 17th Eurographics Conference on Rendering Techniques (EGSR)*.

Marco Forte and François Pitié. 2020. *F*, *B*, Alpha Matting. arXiv:cs.CV/2003.07711

Abhijeet Ghosh, Tongbo Chen, Pieter Peers, Cyrus A Wilson, and Paul Debevec. 2010. Circularly polarized spherical illumination reflectometry. In *ACM SIGGRAPH Asia 2010 papers*. 1–12.

Ned Greene. 1986. Environment mapping and other applications of world projections. *IEEE Computer Graphics and Applications* 6, 11 (1986), 21–29.

Kaiwen Guo, Peter Lincoln, Philip Davidson, Jay Busch, Xueming Yu, Matt Whalen, Geoff Harvey, Sergio Orts-Escolano, Rohit Pandey, Jason Dourgarian, Danhang Tang, Anastasia Tkach, Adarsh Kowdle, Emily Cooper, Mingsong Dou, Sean Fanello, Graham Fyffe, Christoph Rhemann, Jonathan Taylor, Paul Debevec, and Shahram Izadi. 2019. The Relightables: Volumetric Performance Capture of Humans with Realistic Relighting. In *ACM TOG*.

Pierre-Loïc Hamon, James Harmer, Stuart Penn, and Nicolas Scapel. 2014. Gravity: Motion Control and Face Integration. In *ACM SIGGRAPH 2014 Talks (SIGGRAPH '14)*. Association for Computing Machinery, New York, NY, USA, Article 35, 1 pages. https://doi.org/10.1145/2614106.2614193

Qiqi Hou and Feng Liu. 2019. Context-Aware Image Matting for Simultaneous Foreground and Alpha Estimation. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*. IEEE, 4129–4138. https://doi.org/10.1109/ICCV.2019.00423

Tingbo Hou and Tyler Mullen. 2020. *Background Features in Google Meet, Powered by Web ML*. https://ai.googleblog.com/2020/10/background-features-in-google-meet.html

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1125–1134.

Yoshihiro Kanamori and Yuki Endo. 2018. Relighting Humans: Occlusion-aware Inverse Rendering for Full-body Human Images. In *SIGGRAPH Asia*. ACM.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).

Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul E. Debevec. 2019. DeepLight: Learning Illumination for Unconstrained Mobile Mixed Reality. *CVPR* (2019).

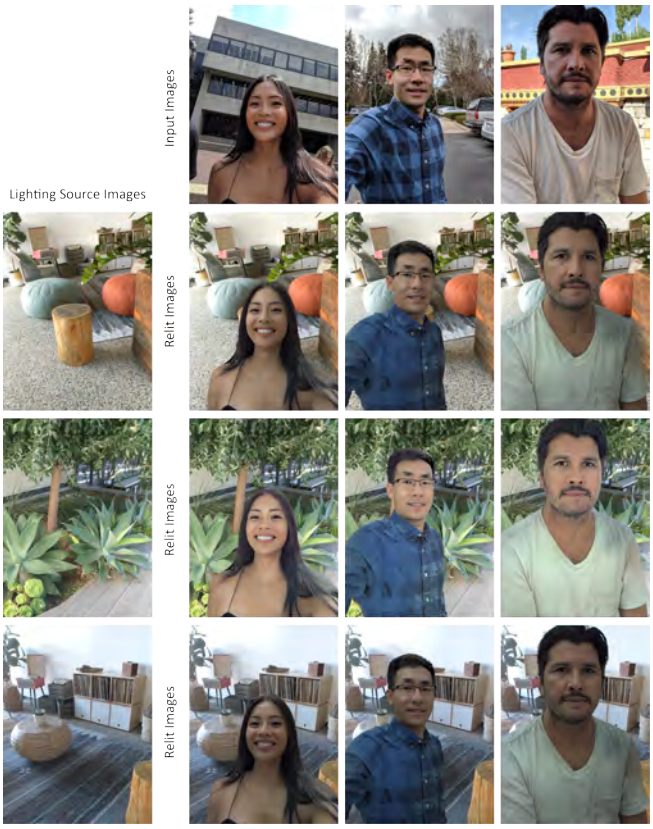

Fig. 23. Application: Any Image Can Be Your Background. We estimate illumination from the input background images in the left column using LeGendre et al. [2019] and then relight the input foreground images in the top row using our approach, compositing the subjects into the new backgrounds with plausibly consistent illumination.

Fig. 24. Application: Portrait lighting transfer. We estimate illumination from portraits in the far left column using LeGendre et al. [2020] and then relight the input images in the top row using our approach, compositing the subjects into the original scenes with consistent illumination.

Fig. 25. Application: Material editing. We can adjust the specular exponent of the Phong model used to prefilter the target HDR lighting environment during inference, which allows for changing the breadth of specular highlights in the relit portraits. This can be seen on the cheek in the top row and nose in the bottom row.

Chloe LeGendre, Wan-Chun Ma, Rohit Pandey, Sean Fanello, Christoph Rhemann, Jason Dourgarian, Jay Busch, and Paul Debevec. 2020. Learning Illumination from Diverse Portraits. In *SIGGRAPH Asia 2020 Technical Communications*.

Anat Levin, Dani Lischinski, and Yair Weiss. 2007. A closed-form solution to natural image matting. *IEEE transactions on pattern analysis and machine intelligence* 30, 2 (2007), 228–242.

Yaoyi Li and Hongtao Lu. 2020. Natural Image Matting via Guided Contextual Attention. arXiv:cs.CV/2001.04069

Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. 2020. Real-Time High-Resolution Background Matting. *arXiv* (2020), arXiv–2012.

Sebastian Lutz, Konstantinos Amplianitis, and Aljosa Smolic. 2018. AlphaGAN: Generative adversarial networks for natural image matting. In *British Machine Vision Conference 2018, BMVC 2018, Newcastle, UK, September 3-6, 2018*. BMVA Press, 259. http://bmvc2018.org/contents/papers/0915.pdf

Wan-Chun Ma, Tim Hawkins, Pieter Peers, Charles-Felix Chabert, Malte Weiss, and Paul Debevec. 2007. Rapid Acquisition of Specular and Diffuse Normal Maps from Polarized Spherical Gradient Illumination. In *Proceedings of the Eurographics Conference on Rendering Techniques (EGSR'07)*.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. 2017. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2794–2802.

Ricardo Martin-Brualla, Rohit Pandey, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Julien Valentin, Sameh Khamis, Philip Davidson, Anastasia Tkach, Peter Lincoln, Adarsh Kowdle, Christoph Rhemann, Dan B Goldman, Cem Keskin, Steve Seitz, Shahram Izadi, and Sean Fanello. 2018. LookinGood: Enhancing Performance Capture with Real-time NeuralRe-Rendering. In *SIGGRAPH Asia*.

Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhoefer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, Peter Denny, Sofien Bouaziz, Peter Lincoln, Matt Whalen, Geoff Harvey, Jonathan Taylor, Shahram Izadi, Andrea Tagliasacchi, Paul Debevec, Christian Theobalt, Julien Valentin, and Christoph Rhemann. 2019. Deep Reflectance Fields - High-Quality Facial Reflectance Field Inference From Color Gradient Illumination. *ACM Transactions on Graphics (Proceedings SIGGRAPH)*.

Abhimitra Meka, Maxim Maximov, Michael Zollhoefer, Avishek Chatterjee, Hans-Peter Seidel, Christian Richardt, and Christian Theobalt. 2018. LIME: Live Intrinsic Material Estimation. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. 11.

Abhimitra Meka, Rohit Pandey, Christian Häne, Sergio Orts-Escolano, Peter Barnum, Philip David-Son, Daniel Erickson, Yinda Zhang, Jonathan Taylor, Sofien Bouaziz, Chloe LeGendre, Wan-Chun Ma, Ryan Overbeck, Thabo Beeler, Paul Debevec, Shahram Izadi, Christian Theobalt, Christoph Rhemann, and Sean Fanello. 2020. Deep Relightable Textures: Volumetric Performance Capture with Neural Rendering.

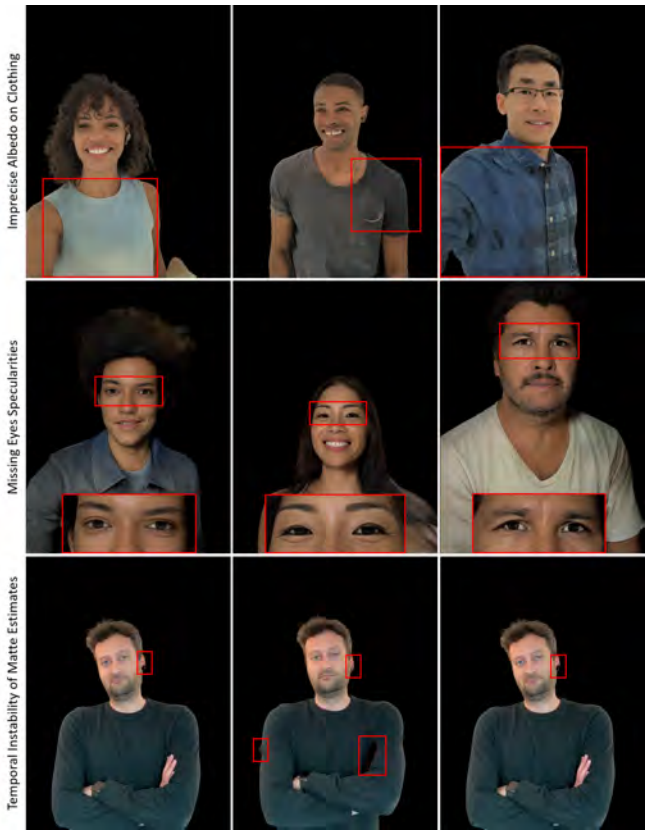

Fig. 26. Limitations. Clothing that is substantially different from what is seen in training may cause poor estimates of albedo color (first row). Although we reproduce plausible non-Lambertian reflections on skin, our methods misses the specular highlights in the eyes (middle row). Finally, when applied to video sequence, the approach may exhibit temporal inconsistency in the alpha matte prediction.

ACM Transactions on Graphics (2020).
Gene S Miller and CD Hoffman. 1984. Illumination and reflection maps. *Course Notes for Advanced Computer Graphics Animation, SIGGRAPH 84* (1984).
Oliver Nalbach, Elena Arabadzhiyska, Dushyant Mehta, Hans-Peter Seidel, and Tobias Ritschel. 2017. Deep Shading: Convolutional Neural Networks for Screen-Space Shading. 36, 4 (2017).
Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas M. Lehrmann. 2020. Learning Physics-guided Face Relighting under Directional Light. In *CVPR*.
Rohit Pandey, Anastasia Tkach, Shuoran Yang, Pavel Pidlypenskyi, Jonathan Taylor, Ricardo Martin-Brualla, Andrea Tagliasacchi, George Papandreou, Philip Davidson, Cem Keskin, Shahram Izadi, and Sean Fanello. 2019. Volumetric Capture of Humans with a Single RGBD Camera via Semi-Parametric Learning. In *CVPR*.
Bui Tuong Phong. 1975. Illumination for computer generated pictures. *Commun. ACM* 18, 6 (1975), 311–317.
Thomas Porter and Tom Duff. 1984. Compositing digital images. In *Proceedings of the 11th annual conference on Computer graphics and interactive techniques*. 253–259.
Ravi Ramamoorthi and Pat Hanrahan. 2001. An efficient representation for irradiance environment maps. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. 497–500.
Erik Reinhard, Michael Ashikhmin, Bruce Gooch, and Peter Shirley. 2001. Color Transfer between Images. *IEEE Computer Graphics and Applications* 21, 5 (2001), 34–41. https://doi.org/10.1109/38.946629
Peiran Ren, Yue Dong, Stephen Lin, Xin Tong, and Baining Guo. 2015. Image Based Relighting Using Neural Networks. *ACM Transactions on Graphics* 34, 4 (July 2015).
Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. 2009. A perceptually motivated online benchmark for image matting. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 1826–1833. https://doi.org/10.1109/CVPR.2009.5206503
Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI* (2015).
Mark Sagar. 2005. Reflectance field rendering of human faces for" spider-man 2". In *ACM SIGGRAPH 2005 Courses*. 14–es.
Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. 2020. PIFuHD: Multi-Level Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
Shen Sang and M. Chandraker. 2020. Single-Shot Neural Relighting and SVBRDF Estimation. In *ECCV*.
Soumyadip Sengupta, Vivek Jayaram, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. 2020. Background Matting: The World is Your Green Screen. In *Computer Vision and Pattern Recognition (CVPR)*.
Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. 2017. Portrait lighting transfer using a mass transport approach. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1.
Xiao Song, Guorun Yang, Xinge Zhu, Hui Zhou, Zhe Wang, and Jianping Shi. 2020. AdaStereo: A Simple and Efficient Approach for Adaptive Stereo Matching. *CoRR* abs/2004.04627 (2020).
Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. 2019. Single image portrait relighting. *ACM Transactions on Graphics (TOG)* 38, 4 (2019), 79.
Tiancheng Sun, Zexiang Xu, Xiuming Zhang, Sean Fanello, Christoph Rhemann, Paul Debevec, Yun-Ta Tsai, Jonathan T Barron, and Ravi Ramamoorthi. 2020. Light stage super-resolution: continuous high-frequency relighting. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–12.
Vladimir Tankovich, Christian Häne, Yinda Zhang, Adarsh Kowdle, Sean Fanello, and Sofien Bouaziz. 2021. HITNet: Hierarchical Iterative Tile Refinement Network for Real-time Stereo Matching. *CVPR* (2021).
Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. 2017. Deep Image Harmonization. *CVPR* (2017).
Yun-Ta Tsai and Rohit Pandey. 2020. *Portrait Light: Enhancing Portrait Lighting with Machine Learning*. https://ai.googleblog.com/2020/12/portrait-light-enhancing-portrait.html
Jue Wang and Michael F. Cohen. 2006. Simultaneous Matting and Compositing. In *ACM SIGGRAPH*.
Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.
Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. 2020. Single Image Portrait Relighting via Explicit Multiple Reflectance Channel Modeling. *ACM SIGGRAPH Asia and Transactions on Graphics* (2020).
Andreas Wenger, Andrew Gardner, Chris Tchou, Jonas Unger, Tim Hawkins, and Paul Debevec. 2005. Performance Relighting and Reflectance Transformation with Time-Multiplexed Illumination. In *SIGGRAPH*.
Robert J. Woodham. 1989. *Photometric Method for Determining Surface Orientation from Multiple Images*. MIT Press, Cambridge, MA, USA.
Steve Wright. 2013. *Digital compositing for film and video*. Taylor & Francis.
Ning Xu, Brian Price, Scott Cohen, and Thomas Huang. 2017. Deep image matting. In *CVPR 2017 (Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017)*. United States, 311–320. https://doi.org/10.1109/CVPR.2017.41
Zexiang Xu, Sai Bi, Kalyan Sunkavalli, Sunil Hadap, Hao Su, and Ravi Ramamoorthi. 2019. Deep View Synthesis from Sparse Photometric Images. *SIGGRAPH* (2019).
Zexiang Xu, Kalyan Sunkavalli, Sunil Hadap, and Ravi Ramamoorthi. 2018. Deep image-based relighting from optimal sparse samples. *ACM Trans. on Graphics* (2018).
Greg Zaal, Sergej Majboroda, and Andreas Mischok. 2020. HDRI Haven. https://www.hdrihaven.com/. Accessed: 2021-01-23.
He Zhang, Jianming Zhang, Federico Perazzi, Zhe Lin, and Vishal M Patel. 2020b. Deep Image Compositing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 365–374.
Richard Zhang. 2019. Making convolutional networks shift-invariant again. In *International Conference on Machine Learning*. PMLR, 7324–7334.
Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. *IEEE conference on computer vision and pattern recognition (CVPR)* (2018).
Xuaner Zhang, Jonathan T. Barron, Yun-Ta Tsai, Rohit Pandey, Xiuming Zhang, Ren Ng, and David E. Jacobs. 2020a. Portrait Shadow Manipulation. *ACM Transactions on Graphics (TOG)* 39, 4.
Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David Jacobs. 2019. Deep Single Image Portrait Relighting. In *ICCV*.
Douglas E Zongker, Dawn M Werner, Brian Curless, and David H Salesin. 1999. Environment matting and compositing. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 205–214.

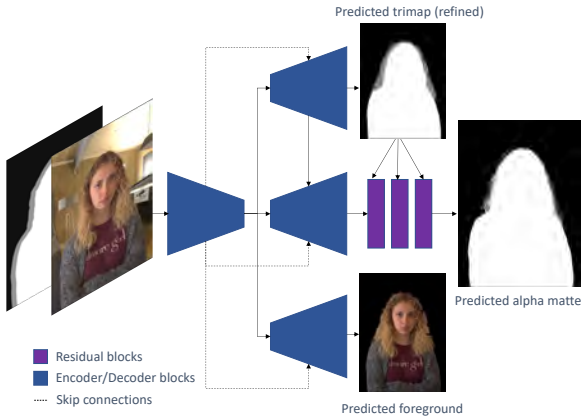

Fig. 27. Matting and Foreground Estimation Network: Our multi-task model predicts a refined trimap, foreground colors, and an alpha matte.

A MATTING MODULE DETAILS

Alpha matting estimation refers to the process of extracting an alpha matte, and depending on the scenario, the colors of the foreground and background objects in an image. Every pixel C in the original image is thus represented as a linear combination of a foreground pixel color F and a background pixel color B [Porter and Duff 1984]:

$$C = \alpha * F + (1 - \alpha) * B \tag{2}$$

Given an image C, simultaneously solving for F, B, and α is under-constrained. For RGB images, there are seven unknowns (RGB values of F and B, and α) and just three equations, one per channel of C. This under-determined system motivates additional constraints leveraged in classical matting estimation methods such as color sampling, using a known background color, and/or using a *trimap*.

We propose to solve this problem using a deep learning approach. Our matting network takes as input an RGB image (C) and a coarse trimap of the foreground (T). A multi-task encoder-decoder architecture (U-Net [Ronneberger et al. 2015]) is then used to predict a refined trimap $\hat{T}$, the alpha channel α, and the foreground F (see Fig. 27). Our multi-task model is inspired by [Cai et al. 2019]; however, we add a decoder branch that also predicts foreground colors. As demonstrated in [Chen et al. 2018a; Forte and Pitié 2020; Hou and Liu 2019], this type of deep learning architecture can handle the simultaneous estimation of alpha matte and foreground colors if supervised with appropriate ground truth data.

A.1 Input Trimap

An initial coarse segmentation is computed using an off-the-shelf foreground segmentation network [Chen et al. 2018b] trained to segment people in images. Then, the input trimap T is generated by thresholding the foreground probabilities from the segmentation mask and applying *erode* and *dilate* morphological operations to define the unknown region.

A.2 Feature Extractor

The feature extractor takes as input the RGB image C and the trimap T and performs a series of convolutions with kernel size 3×3 with ReLU activations and $32, 64, 128, 256, 512$ channels for each level respectively. Its output is passed through three different decoder branches that predict the refined trimap $\hat{T}$, the alpha matte α, and the foreground F (Fig. 27, right).

A.3 Trimap Refinement

A first decoder branch predicts the refined trimap $\hat{T}$ using convolutions with skip connections from the encoder. The decoder consists of 4 layers with $256, 128, 64, 32$ filters, extracted with 3×3 convolutions followed by Leaky ReLU activations. (Fig. 27, top branch).

A.4 Alpha Matte Prediction

A second decoder with the same architecture as the trimap refinement predicts an alpha matte α. Its output is passed through a series of 3 residual blocks with 7×7, 5×5 and 3×3 convolutions (with ReLU activations) together with the input RGB image to predict final *refined* alpha mask α. This refinement step (residual learning) has proven to be effective improving fine details of the final estimated alpha matte [Cai et al. 2019; Chen et al. 2018a; Xu et al. 2019].

A.5 Foreground Prediction

Finally, a third branch takes as input the encoded features to predict the foreground F. The decoder architecture matches that of the trimap refinement branch.

A.6 Matting Module Loss Functions

The matting module relies on the following loss terms:

Trimap Loss $\mathcal{L}_T$: $E(T_{\text{gt}}, \hat{T})$. This term computes the sparse cross entropy loss between the refined trimap $\hat{T}$ and the ground truth trimap T_{gt}.

Alpha Loss $\mathcal{L}_\alpha$: $\ell_1(\alpha_{\text{gt}}, \alpha)$. To infer the alpha matte, we simply compute an ℓ_1 norm between the ground truth matte α_{gt} and the inferred matte α, calculated on the unknown regions of T_{gt}.

Pyramid Laplacian Loss $\mathcal{L}_{\text{Lap}}$: $\sum_{i=1}^{5} = 2^{i-1} * \ell_1(\text{Lap}^i(\alpha_{\text{gt}}), \text{Lap}^i(\alpha))$. This multi-scale loss on the predicted alpha matte takes the difference between two Laplacian pyramid representations, accounting from local and global differences. Contributions from deeper levels are scaled according to their spatial support. As discussed in previous works [Forte and Pitié 2020; Hou and Liu 2019], this loss often improves quantitative results.

Foreground Loss $\mathcal{L}_F$: $\ell_1(\sum F_{\text{gt}}, F)$. Finally, an ℓ_1 loss between the predicted foreground F and the ground truth foreground F_{gt} is minimized. This loss is only computed for pixels where the foreground is visible, $\alpha_{\text{gt}} > 0$.

Compositional Loss: $\mathcal{L}_C$. This term computes the ℓ_1 norm between the ground truth input RGB colors and a composited image using the predicted foreground RGB colors, ground truth background, and the predicted alpha matte. This constrains the network to follow the alpha matting equation, improving predictions [Xu et al. 2017].

The matting model was trained end-to-end using a weighted sum of previous losses:

$$\mathcal{L}_{\text{matting}} = \lambda_T * \mathcal{L}_T + \lambda_\alpha * \mathcal{L}_\alpha + \lambda_{\text{Lap}} * \mathcal{L}_{\text{Lap}} + \lambda_F * \mathcal{L}_F + \lambda_C * \mathcal{L}_C \tag{3}$$

We empirically determined values for these hyperparameters and set $\lambda_T = \lambda_F = \lambda_C = 1$, $\lambda_{\text{Lap}} = 4$ and $\lambda_\alpha = 2$.