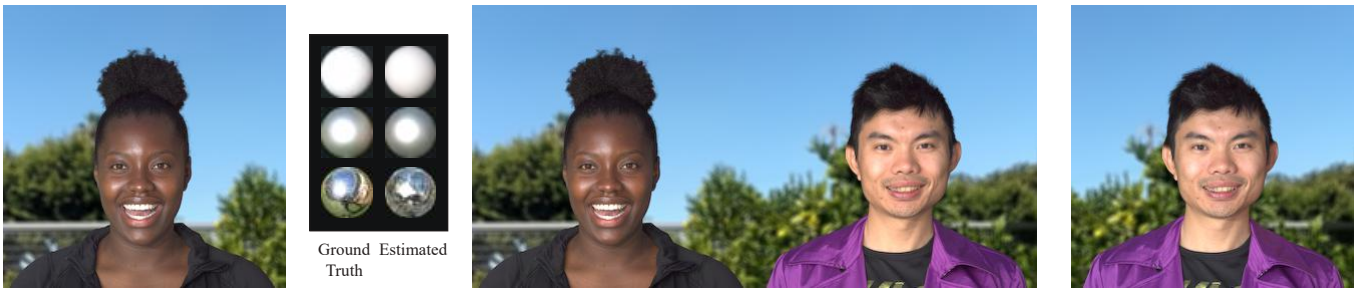
Learning Illumination from Diverse Portraits

CHLOE LEGENDRE, Google Research
WAN-CHUN MA, ROHIT PANDEY, SEAN FANELLO, CHRISTOPH RHEMANN, JASON DOURGARIAN, and JAY BUSCH, Google
PAUL DEBEVEC, Google Research



(a) Input portrait **(b)** Lighting **(c)** Original and novel subjects lit with estimated lighting **(d)** Novel subject lit by GT lighting

Fig. 1. Our network estimates HDR omnidirectional lighting from an LDR portrait image. **(a)** Input portrait image generated using a photographed reflectance basis. **(b)** Ground truth and estimated lighting, shown on diffuse, glossy, and mirror spheres. **(c)** Original and novel subjects lit consistently by the estimated lighting, using image-based relighting. **(d)** The novel subject lit with the original subject's ground truth lighting. For both subjects, the appearance under the estimated lighting closely matches the appearance under the original lighting.

We present a learning-based technique for estimating high dynamic range (HDR), omnidirectional illumination from a single low dynamic range (LDR) portrait image captured under arbitrary indoor or outdoor lighting conditions. We train our model using portrait photos paired with their ground truth environmental illumination. We generate a rich set of such photos by using a light stage to record the reflectance field and alpha matte of 70 diverse subjects in various expressions. We then relight the subjects using image-based relighting with a database of one million HDR lighting environments, compositing the relit subjects onto paired high-resolution background imagery recorded during the lighting acquisition. We train the lighting estimation model using rendering-based loss functions and add a multi-scale adversarial loss to estimate plausible high frequency lighting detail. We show that our technique outperforms the state-of-the-art technique for portrait-based lighting estimation, and we also show that our method reliably handles the inherent ambiguity between overall lighting strength and surface albedo, recovering a similar scale of illumination for subjects with diverse skin tones. We demonstrate that our method allows virtual objects and digital characters to be added to a portrait photograph with consistent illumination. Our lighting inference runs in real-time on a smartphone, enabling realistic rendering and compositing of virtual objects into live video for augmented reality applications.

1 INTRODUCTION

In both portrait photography and film production, lighting greatly influences the look and feel of a given shot. Photographers and cinematographers dramatically light their subjects to communicate a particular aesthetic sensibility and emotional tone. While films using visual effects techniques often blend recorded camera footage with computer-generated, rendered content, the realism of such composites depends on the consistency between the real-world lighting and that used to render the virtual content. Thus, visual effects practitioners work painstakingly to capture and reproduce real-world illumination inside *virtual* sets. Debevec (1998) introduced one such technique for real-world lighting capture, recording the color and intensity of omnidirectional illumination by photographing a mirror sphere using multiple exposures. This produced an HDR "image-based lighting" (IBL) environment (Debevec 2006), used for realistically rendering virtual content into real-world photographs.

Augmented reality (AR) shares with post-production visual effects the goal of realistically blending virtual content and real-world imagery. Face-based AR applications are ubiquitous, with widespread adoption in both social media and video conferencing applications. However, in real-time AR, lighting measurements from specialized capture hardware are unavailable, as acquisition is impractical for casual mobile phone or headset users. Similarly, in visual effects, on-set lighting measurements are not always available, yet lighting artists must still reason about illumination using cues in the scene. If the footage includes faces, their task is somewhat less challenging, as faces include a diversity of surface normals and reflect light somewhat predictably.

Prior work has leveraged the strong geometry and reflectance priors from faces to solve for lighting from portraits. In the years since Marschner and Greenberg (1997) introduced portrait "inverse lighting," most such techniques (Egger et al. 2018; Kemelmacher-Shlizerman and Basri 2010; Knorr and Kurz 2014; Sengupta et al. 2018; Shim 2012; Shu et al. 2017b; Tewari et al. 2018; Tewari et al. 2017; Zhou et al. 2018) have sought to recover both facial geometry and a low frequency approximation of distant scene lighting, usually represented using up to a 2^{nd} order spherical harmonic (SH) basis. The justification for this approximation is that skin reflectance is predominantly diffuse (Lambertian) and thus acts as a low-pass filter on the incident illumination. For diffuse materials, irradiance indeed lies very close to a 9D subspace well-represented by this basis (Basri and Jacobs 2003; Ramamoorthi and Hanrahan 2001a).

However, to the skilled portrait observer, the lighting at capture-time reveals itself not only through the skin's diffuse reflection, but also through the directions and extent of cast shadows and the intensity and locations of specular highlights. Inspired by these cues, we train a neural network to perform inverse lighting from portraits, estimating omnidirectional HDR illumination without assuming any specific skin reflectance model. Our technique yields higher frequency lighting that can be used to convincingly render novel subjects into real-world portraits, with applications in both visual effects and AR when off-line lighting measurements are unavailable. Furthermore, our lighting inference runs in real-time on a smartphone, enabling such applications.

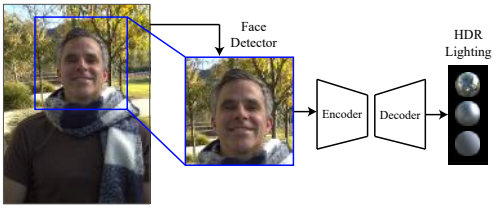


Fig. 2. We train a convolutional neural network to regress from a face-cropped input image to omnidirectional, HDR illumination.

We train our lighting estimation model in a supervised manner using a dataset of portraits and their corresponding ground truth illumination. To generate this dataset, we photograph 70 diverse subjects in a light stage system as illuminated by 331 directional light sources forming a basis on a sphere, such that the captured subject can be relit to appear as they would in any scene with image-based relighting (Debevec et al. 2000). Although a few databases of real-world lighting environments captured using traditional HDR panoramic photography techniques are publicly available, e.g. the Laval indoor and outdoor datasets with 2,000 and 12,000 scenes respectively (Gardner et al. 2017; Lalonde and Matthews 2014), we extend the LDR data collection technique of LeGendre et al. (2019) to instead capture on the order of 1 million indoor and outdoor lighting environments, promoting them to HDR via a novel non-negative least squares solver formulation before using them for relighting.

A few recent works have similarly sought to recover illumination from portraits without relying on a low-frequency lighting basis, including the deep learning methods of Sun et al. (2019) for arbitrary scenes and Calian et al. (2018) for outdoor scenes containing the sun. We show that our method out-performs both of these methods, and generalizes to arbitrary indoor or outdoor scenes.

Any attempt at lighting estimation is complicated by the inherent ambiguity between surface reflectance (albedo) and light source strength (Belhumeur et al. 1999). Stated otherwise, a pixel's shading is rendered unchanged if its albedo is halved while light source intensity doubles. Statistical priors for facial albedo have been leveraged to resolve this ambiguity (Calian et al. 2018; Egger et al. 2018; Tewari et al. 2017), but, to the best of our knowledge, we are the first to explicitly evaluate the performance of our model on a wide variety of subjects with different skin tones. In contrast, Sun et al. (2019) report lighting accuracy with a scale-invariant metric, while Calian et al. (2018) show visual results for synthetically rendered and photographed faces where the subjects are predominantly light in skin tone. We show that for a given lighting condition, our model can recover lighting at a similar scale for a variety of diverse subjects.

In summary, our contributions are the following:

- A deep learning method to estimate HDR illumination from LDR images of faces in both indoor and outdoor scenes. Our technique outperforms the previous state-of-the-art.
- A first-of-its-kind analysis that shows that our HDR lighting estimation technique reliably handles the ambiguity between light source strength and surface albedo, recovering similar illumination for subjects with diverse skin tones.

2 RELATED WORK

In this section we summarize work related to lighting capture, inverse rendering from faces, and the related topics of portrait relighting and unconstrained lighting estimation.

Lighting measurement techniques. After Debevec (1998) introduced image-based lighting from high dynamic range panoramas, subsequent work proposed more general acquisition techniques including recording the extreme dynamic range of sunny daylight with a fisheye lens (Stumpfel et al. 2004) and recording HDR video with a mirror sphere (Unger et al. 2006; Waese and Debevec 2002). Debevec et al. (2012) and Reinhard et al. (2010) presented more practical techniques to recover the full dynamic range of daylight by augmenting the typical mirror sphere capture with simultaneous photography of a diffuse, gray sphere that allowed for saturated light source intensity recovery. We extend these techniques to promote one million real-world, clipped panoramas to HDR.

Inverse Rendering. The joint recovery of scene geometry, material reflectance, and illumination given only an image, thereby inverting the image formation or rendering process, is a long-studied problem in computer vision (Lombardi and Nishino 2016; Ramamoorthi and Hanrahan 2001b; Yu et al. 1999). Similarly, the topic of "intrinsic image" decomposition has received considerable attention, recovering shading and reflectance, rather than geometry and illumination (Barrow et al. 1978; Land and McCann 1971). "Shape from Shading" methods aim to recover geometry under known illumination (Horn 1970), while another variant jointly recovers "Shape, Illumination, and Reflectance from Shading" (Barron and Malik 2014).

Recently, significant progress has been made in the domain of inverse rendering from portrait images or videos, with the goal of recovering a 3D face model with illumination and/or reflectance (Egger et al. 2018; Kemelmacher-Shlizerman and Basri 2010; Sengupta et al. 2018; Tewari et al. 2018; Tewari et al. 2017; Tran et al. 2019; Tran and Liu 2019; Yamaguchi et al. 2018). Many of these techniques rely on geometry estimation via fitting or learning a 3D morphable model (Blanz and Vetter 1999), and they model skin reflectance as Lambertian and scene illumination using a low-frequency 2nd order SH basis. In contrast, our goal is to recover higher frequency illumination useful for rendering virtual objects with diverse reflectance characteristics beyond Lambertian.

Inverse Lighting from Faces. Marschner and Greenberg (1997) introduced the problem of "inverse lighting," estimating the directional distribution and intensity of incident illumination falling on a rigid object with measured geometry and reflectance, demonstrating lighting estimation from portraits as one such example. With the appropriate lighting basis and reflectance assumption, the problem was reduced to inverting a linear system of equations. The linearity of light transport was similarly leveraged in follow-up work to estimate lighting from faces (Shahlaei and Blanz 2015; Shim 2012), including for real-time AR (Knorr and Kurz 2014), but these approaches estimated either a small number of point light sources or again used a low frequency 2^{nd} order SH lighting basis. Specular reflections from the eyes of portrait subjects have been leveraged to estimate higher frequency illumination, but as the reflections of bright light sources are likely to be clipped, the recovery of the full dynamic range of natural illumination is challenging to recover from a single exposure image (Nishino and Nayar 2004).

Several new deep learning techniques for inverse lighting from faces have been proposed. Zhou et al. (2018) estimated 2^{nd} order SH illumination from portraits. For higher frequency lighting estimates, Yi et al. (2018) recovered illumination by first estimating specular highlights and ray-tracing them into a panorama of lighting directions. However, this model produced HDR IBL maps that are mostly empty (black), with only dominant light source colors and intensities represented. In contrast, we estimate plausible omnidirectional illumination. Calian et al. (2018) trained an autoencoder on a large database of outdoor panoramas to estimate lighting from LDR portraits captured outdoors, combining classical inverse lighting and deep learning. While this method produced impressive results for outdoor scenes with natural illumination, it is not applicable to indoor scenes or outdoor scenes containing other sources of illumination. Our model, in contrast, generalizes to arbitrary settings. Critically, neither Yi et al. (2018) nor Calian et al. (2018) evaluated model performance on subjects with diverse skin tones, which we feel is an important variation axis for lighting estimation error analysis. Both works presented qualitative results only for photographed subjects and rendered computer-generated models with fair skin.

Portrait Relighting. Marchner and Greenberg (1997) also proposed *portrait relighting* and *portrait lighting transfer*, showing that the lighting from one portrait subject could be used to approximately relight another subject, such that the two could be convincingly composited together into one photograph. Recent works solved this problem either with a mass transport (Shu et al. 2017a) or deep learning (Sun et al. 2019; Zhou et al. 2019) approach. Sun et al. (2019) estimated illumination while training a portrait relighting network. Lighting estimates from this technique proved superior compared with two other state-of-the-art methods (Barron and Malik 2014; Sengupta et al. 2018). Similarly to Sun et al. (2019), we generate photo-realistic, synthetic training data using a set of reflectance basis images captured in an omnidirectional lighting system, or light stage, relying on the technique of *image-based relighting* (Debevec et al. 2000; Nimeroff et al. 1995) to synthesize portraits lit by novel sources of illumination. However, in contrast to Sun et al. (2019), we extend a recent environmental lighting capture technique (LeGendre et al. 2019) to expand the number of lighting environments used for training data, employ a set of loss functions designed specifically for lighting estimation, and use a lightweight model to achieve lighting inference at real-time frame rates on a mobile device. Even when trained on the same dataset, we show that our lighting estimation model outperforms that of Sun et al. (2019), the previous state-of-the-art for lighting estimation from portraits.

Lighting Estimation. Given the prominence of virtual object compositing in both visual effects and AR, it is unsurprising that lighting estimation from arbitrary scenes (not from portraits) is also an active research area. Several works have sought to recover outdoor, natural illumination from an unconstrained input image (Hold-Geoffroy et al. 2019, 2017; Lalonde et al. 2009; Lalonde and Matthews 2014; Zhang et al. 2019). Several deep learning based methods have recently tackled indoor lighting estimation from unconstrained images (Gardner et al. 2017; Garon et al. 2019; Song and Funkhouser 2019). Cheng et al. (2018) estimated lighting using a deep learning technique given two opposing views of a panorama. For AR applications, LeGendre at al. (2019) captured millions of LDR images of three diffuse, glossy, and mirror reference spheres as they appeared in arbitrary indoor and outdoor scenes, using this dataset to train a model to regress to omnidirectional HDR lighting from an unconstrained image. We leverage this lighting data collection technique but extend it to explicitly promote the captured data to HDR so that it can be used for image-based relighting, required for generating our synthetic portraits. LeGendre et al. (2019) trained their model using a combination of rendering-based and adversarial losses, which we extend to the multi-scale domain for superior performance.

3 METHOD

3.1 Training Data Acquisition and Processing

To train a model to estimate lighting from portrait photographs in a supervised manner, we require many portraits labeled with ground truth illumination. Since no such real-world, dataset exists, we synthesize portraits using the data-driven technique of image-based relighting, shown by Debevec et al. (2000) to produce photo-realistic relighting results for human faces, appropriately capturing complex light transport phenomena for human skin and hair e.g. sub-surface and asperity scattering and Fresnel reflections. Noting the difficulty of generating labeled imagery for the problem of inverse lighting from faces, many prior works have instead relied on renderings of 3D models of faces (Calian et al. 2018; Yi et al. 2018; Zhou et al. 2018), which often fail to represent these complex phenomena.

Reflectance Field Capture. Debevec et al. (2000) introduced the 4D *reflectance field* $R(\theta, \phi, x, y)$ to denote a subject lit from any lighting direction (θ, ϕ) for image pixels (x, y) and showed that taking the dot product of this reflectance field with an HDR lighting environment similarly parameterized by (θ, ϕ) relights the subject to appear as they would in that scene. To photograph a subject's reflectance field, we use a computer-controllable sphere of 331 white LED light sources, similar to that of Sun et al. (2019), with lights spaced 12° apart at the equator. The reflectance field is formed from a set of reflectance basis images (see Fig. 3), photographing the subject as

each of the directional LED light sources is individually turned on one-at-a-time within the spherical rig. We capture these "One-Light-At-a-Time" (OLAT) images for multiple camera viewpoints, shown in Fig. 4. In total we capture 331 OLAT images for each subject using six color Ximea machine vision cameras with 12 megapixel resolution, placed 1.7 meters from the subject. The cameras are positioned roughly in front of the subject, with five cameras with 35 mm lenses capturing the upper body of the subject from different angles, and one additional camera with a 50 mm lens capturing a close-up image of the face with tighter framing.

Fig. 3. "One-Light-at-a-Time" images: 24 of the 331 lighting directions.

Fig. 4. Our six camera viewpoints for an example lighting direction.

We capture reflectance fields for 70 diverse subjects, each performing nine different facial expressions and wearing different accessories, yielding about 630 sets of OLAT sequences from six different camera viewpoints, for a total of 3780 unique OLAT sequences. In addition to age and gender diversity, we were careful to photograph subjects spanning a wide range of skin tones, as seen in Fig. 5.

Fig. 5. Representative portraits of the 70 recorded subjects.

Since acquiring a full OLAT sequence for a subject takes six seconds, there can be subject motion over the sequence. We therefore employ an optical flow technique (Anderson et al. 2016) to align the images, interspersing at every 11^{th} OLAT frame one extra "tracking" frame with even, consistent illumination to ensure the brightness constancy constraint for optical flow is met, as in Wenger et al. (2005). This step preserves the sharpness of image features when performing the relighting operation, which linearly combines aligned OLAT images.

Alpha Matte Acquisition. For the two frontal camera views, we also acquire images to compute an alpha matte for each subject, so we can composite them over novel backgrounds. We acquire a first image where the subject is unlit and a grey background material placed behind the subject is lit relatively evenly by six LED sources. We also photograph a clean plate of the background under the same lighting condition without the subject in the scene, such that the alpha matte can be computed by dividing the first image by the clean plate, as in Debevec et al. (2002). Although the work of Sun et al. (2019) uses a human segmentation algorithm to remove sections of the image corresponding to background elements (e.g. the light stage rig and lights visible within the camera view), we use our more accurate alpha matte for our frontal views (Fig. 8b). For the remaining views, we compute an approximate segmentation using a method designed to handle the challenging task of segmenting hair from background imagery (Tkachenka et al. 2019).

HDR Lighting Environment Capture. To relight our subjects with photographed reflectance fields, we require a large database of HDR lighting environments, where no light sources are clipped. While there are a few such datasets containing on the order of thousands of indoor panoramas (Gardner et al. 2017) or the upper hemisphere of outdoor panoramas (Lalonde and Matthews 2014), deep learning models are typically enhanced with a greater volume of training data. Thus, we extend the video-rate capture technique of LeGendre et al. (2019) to collect on the order of 1 million indoor and outdoor lighting environments. This work captured background images augmented by a set of diffuse, matte silver, and mirrored reference spheres held in the lower part of the frame as in Fig. 6. These three spheres reveal different cues about the scene illumination. The mirror ball reflects omnidirectional high frequency lighting, but bright light sources will be clipped, altering both their intensity and color. The near-Lambertian BRDF of the diffuse ball, in contrast, acts as a low-pass filter on the incident illumination, capturing a blurred but relatively complete record of total scene irradiance. Without explicitly promoting these LDR sphere appearances to a record of HDR environmental illumination, LeGendre et al. (2019) regressed from the unconstrained background images to HDR lighting using an in-network, differentiable rendering step, predicting illumination to match the clipped, LDR ground truth sphere appearances. In contrast, we require a true HDR record of the scene illumination to use for relighting our subjects, so, unlike LeGendre et al. (2019), we must explicitly promote the three sphere appearances into an estimate of their corresponding HDR lighting environment.

Promoting LDR Sphere Images to HDR Lighting. Given captured images of the three reflective spheres, perhaps with clipped pixels, we

Fig. 6. Background images with ground truth lighting recorded by diffuse, matte silver, and mirrored spheres as in LeGendre et al. (2019).

wish to solve for HDR lighting that could have plausibly produced these three sphere appearances. We first record the reflectance field for the diffuse and matte silver spheres, again using the light stage system. We convert their reflectance basis images into the same relative radiometric space, normalizing based on the incident light source color. We then project the reflectance basis images into the mirror ball mapping (Reinhard et al. 2010) (Lambert azimuthal equal-area projection), accumulating energy from the input images for each new lighting direction (θ, ϕ) on a 32×32 image of a mirror sphere as in LeGendre et al. (2019), forming the reflectance field $R(\theta, \phi, x, y)$, or, sliced into individual pixels, $R_{x,y}(\theta, \phi)$.

For lighting directions (θ, ϕ) in the captured mirror ball image *without* clipping for color channel c, we recover the scene lighting $L_c(\theta, \phi)$ by simply scaling the mirror ball image pixel values by the inverse of the measured mirror ball reflectivity (82.7%). For lighting directions (θ, ϕ) *with* clipped pixels in the original mirror ball image, we set the pixel values to 1.0, scale this by the inverse of the measured reflectivity forming $L_c(\theta, \phi)$, and then subsequently solve for a residual missing lighting intensity $U_c(\theta, \phi)$ using a non-negative least squares solver formulation. Given an original image pixel value $p_{x,y,c,k}$ for BRDF index k (e.g. diffuse or matte silver), and color channel c, and the measured reflectance field $R_{x,y,c,k}(\theta, \phi)$, due to the superposition principle of light, we can write:

$$p_{x,y,c,k} = \sum_{\theta,\phi} R_{x,y,c,k}(\theta, \phi)[L_c(\theta, \phi) + U_c(\theta, \phi)] \tag{1}$$

This generates a set of m linear equations for each BRDF k and color channel c, equal to the number of sphere pixels in the reflectance basis images, with n unknown residual light intensities. For lighting directions without clipping, we know that $U_c(\theta, \phi) = 0$. For each color channel, with $km > n$, we can solve for the unknown $U_c(\theta, \phi)$ values using non-negative least squares, ensuring light is only added, not removed. In practice, we exclude clipped pixels $p_{x,y,c,k}$ from the solve. Prior methods have recovered clipped light source intensities by comparing the pixel values from a photographed diffuse sphere with the diffuse convolution of a clipped panorama (Debevec et al. 2012; Reinhard et al. 2010), but, to the best of our knowledge, we are the first to use photographed reflectance bases and multiple BRDFs. In Fig. 7 upper rows, we show input sphere images extracted from LDR imagery ("ground truth"), and in lower rows, we show the three spheres rendered using Eqn. 1, lit with the HDR illumination recovered from the solver.

We observed when solving for $U_c(\theta, \phi)$ treating each color channel independently, brightly-hued red, green, and blue light sources were produced, often at geometrically-nearby lighting directions,

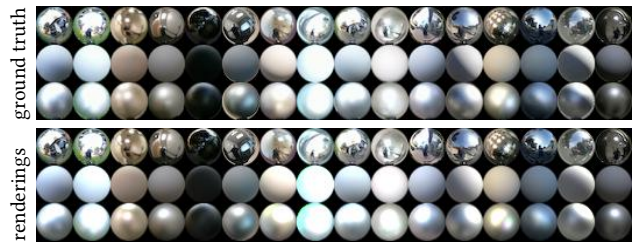


Fig. 7. Upper: ground truth LDR sphere images (inputs to the LDR to HDR linear solver). Lower: spheres rendered using the recovered HDR illumination, using image-based relighting and the captured reflectance basis.

rather than a single light source with greater intensity in all three colors channels. To recover results with more plausible, neutrally-colored light sources, we add a cross color channel regularization based on the insight that the color of the photographed diffuse grey ball reveals the average color balance $(R_{avg}, G_{avg}, B_{avg})$ of the bright light sources in the scene. We thus add to our system of equations a new set of linear equations with weight $\lambda = 0.5$:

$$\frac{[L_{c=R}(\theta, \phi) + U_{c=R}(\theta, \phi)]}{[L_{c=G}(\theta, \phi) + U_{c=G}(\theta, \phi)]} = \frac{R_{avg}}{G_{avg}} \tag{2}$$

$$\frac{[L_{c=R}(\theta, \phi) + U_{c=R}(\theta, \phi)]}{[L_{c=B}(\theta, \phi) + U_{c=B}(\theta, \phi)]} = \frac{R_{avg}}{B_{avg}} \tag{3}$$

These regularization terms penalize the recovery of strongly-hued light sources of a different color balance than the target diffuse ball. Debevec et al. (2012) noted that a regularization term could be added to encourage similar intensities for geometrically-nearby lighting directions, but this would not necessarily prevent the recovery of strongly-hued lights. We recover $U_c(\theta, \phi)$ using the Ceres solver (Agarwal et al. [n. d.]), promoting our 1 million captured sphere appearances to HDR illumination. As the LDR images from this video-rate data collection method are 8-bit and encoded as sRGB, possibly with local tone-mapping, we first linearize the sphere images assuming $\gamma = 2.2$, as required for our linear system formulation.

Portrait Synthesis. Using our photographed reflectance fields for each subject and our HDR-promoted lighting, we generate relit portraits with ground truth illumination to serve as training data. We again convert the reflectance basis images into the same relative radiometric space, calibrating based on the incident light source color. As our lighting environments are represented as 32×32 mirror ball images, we project the reflectance fields onto this basis, again accumulating energy from the input images for each new lighting direction (θ, ϕ) as in LeGendre et al. (2019). Each new basis image is a linear combination of the original 331 OLAT images.

The lighting capture technique also yields a high-resolution background image corresponding to the three sphere appearances. Since such images on their own contain useful cues for extracting lighting estimates (Gardner et al. 2017; Hold-Geoffroy et al. 2017), we composite our relit subjects onto the these backgrounds rather than onto a black frame as in Sun et al. (2019), as shown in Fig. 8, producing images which mostly appear to be natural photographs taken out in the wild. Since the background images are 8-bit sRGB, we clip and

apply this transfer function to the relit subject images prior to compositing. As in-the-wild portraits are likely to contain clipped pixels (especially for 8-bit live video for mobile AR), we discard HDR data for our relit subjects to match the expected inference-time inputs.

(a) background **(b)** alpha matte **(c)** relit subject **(d)** composited

Fig. 8. **(a)** A background with paired HDR illumination, shown via the inset spheres (upper right). **(b)** Alpha matte from our system. **(c)** Subject relit with the illumination from **a**. **(d)** Subject relit and composited into **a.**

Face Localization. Although background imagery may provide contextual cues that aid in lighting estimation, we do not wish to waste our network's capacity learning a face detector. Instead, we compute a face bounding box for each input, and during training and inference we crop each image, expanding the bounding box by 25%. During training we add slight crop region variations, randomly changing their position and extent. In our implementation, we use the BlazeFace detector of Bazarevsky et al. (2019), but any could be used. In Fig. 9 we show example cropped inputs to our model.

Fig. 9. Example synthetic training portraits, cropped to the bounding box of the detected face. Upper right corners: ground truth HDR illumination for each training example (not included as input during training).

3.2 Network Architecture

The input to our model is an sRGB encoded LDR image, with the crop of the detected face region of each image resized to an input resolution of 256×256 and normalized to the range of $[-0.5, 0.5]$. We use an encoder-decoder architecture with a latent vector of size 1024 at the bottleneck, representing log-space HDR illumination, as the sun can be several orders of magnitude brighter than the sky (Stumpfel et al. 2004). The encoder consists of five 3×3 convolutions each followed by a blur-pooling operation (Zhang 2019), with successive filter depths of 16, 32, 64, 128, and 256, followed by one last convolution with a filter size of 8×8 and depth 256, and finally a fully-connected layer. The decoder consists of three sets of 3×3 convolutions of filter depths 64, 32, and 16, each followed by a bilinear-upsampling operation. The final output of the network is a 32×32 HDR image of a mirror ball representing log-space omnidirectional illumination.

We also use an auxiliary discriminator architecture to add an adversarial loss term, enforcing estimation of plausible high frequency illumination (see Sec. 3.3). This network takes as input clipped images of ground truth and predicted illumination from the main model, and tries to discriminate between the real and generated examples. The discriminator encoder consists of three 3×3 convolutions each followed by a max-pooling operation, with successive filter depths of 64, 128, and 256, followed by a fully connected layer of size 1024 before the final output layer. As our main network's decoder includes several upsampling operations, our network is implicitly learning information at multiple scales. We leverage this multi-scale output to provide inputs to the discriminator not just of the full-resolution 32×32 clipped lighting image, but also of a lighting image at each scale: 4×4, 8×8, and 16×16, using the multi-scale gradient technique of MSG-GAN (Karnewar and Wang 2020). As the lower-resolution feature maps produced by our generator network have more than 3 channels, we add a convolution operation at each scale as extra branches of the network, producing multiple scales of 3-channel lighting images to supply to the discriminator.

3.3 Loss Function

Multi-scale Image-Based Relighting Rendering Loss. LeGendre et al. (2019) describe a differentiable image-based relighting rendering loss, used for training a network to estimate HDR lighting $\hat{L}$ from an unconstrained image. This approach minimizes the reconstruction loss between the ground truth sphere images I for multiple BRDFs and the corresponding network-rendered spheres $\hat{I}$, lit with the predicted illumination. We use this technique to train our model for inverse lighting from portraits, relying on these sphere renderings to learn illumination useful for rendering virtual objects of a variety of BRDFs. We produce sphere renderings $\hat{I}$ in-network using image-based relighting and photographed reflectance fields for each sphere of BRDF index k (mirror, matte silver, or diffuse), and color channel c, with $\hat{L_c}(\theta, \phi)$ as the intensity of light for the direction (θ, ϕ):

$$\hat{I}_{x,y,k,c} = \sum_{\theta,\phi} R_{x,y,k,c}(\theta,\phi)\hat{L_c}(\theta,\phi). \tag{4}$$

As in LeGendre et al. (2019), our network similarly outputs a log space image Q of HDR illumination, with pixel values $Q_c(\theta, \phi)$, so sphere images are rendered as:

$$\hat{I}_{x,y,k,c} = \sum_{\theta,\phi} R_{x,y,k,c}(\theta,\phi) e^{Q_c(\theta,\phi)}. \quad (5)$$

With binary mask $\hat{M}$ to mask out the corners of each sphere, $\gamma = 2.2$ for gamma-encoding, λ_k as an optional weight for each BRDF, and a differentiable soft-clipping function Λ as in LeGendre et al. (2019), the final LDR image reconstruction loss L_{rec} comparing ground truth images I_k and network-rendered images $\hat{I}_k$ is:

$$L_{\text{rec}} = \sum_{k=0}^{2} \lambda_k \left\| \hat{M} \odot (\Lambda(\hat{I}_k)^{\frac{1}{\gamma}} - \Lambda(I_k)) \right\|_1. \quad (6)$$

Rather than use the LDR sphere images captured in the video-rate data collection as the reference images I_k, we instead render the spheres with the HDR lighting recovered from the linear solver of Sec. 3.1, gamma-encoding the renderings with $\gamma = 2.2$. This ensures that the same lighting is used to render the "ground truth" spheres as the input portraits, preventing the propagation of residual error from the HDR lighting recovery to our model training phase.

We finally add extra convolution branches to convert the multi-scale feature maps of the decoder into 3-channel images representing log-space HDR lighting at successive scales. We then extend the rendering loss function of LeGendre et al. (2019) (Eqn. 6) to the multi-scale domain, rendering mirror, matte silver, and diffuse spheres during training in sizes 4×4, 8×8, 16×16, and 32×32. With scale index represented by s, and an optional weight for each as λ_s, our multi-scale image reconstruction loss is written as:

$$L_{\text{ms-rec}} = \sum_{s=0}^{3} \sum_{k=0}^{2} \lambda_s \lambda_k \left\| \hat{M} \odot (\Lambda(\hat{I}_k)^{\frac{1}{\gamma}} - \Lambda(I_k)) \right\|_1. \quad (7)$$

Adversarial Loss. Recent work in unconstrained lighting estimation has shown that adversarial loss terms improve the recovery of high-frequency information compared with using only image reconstruction losses (LeGendre et al. 2019; Song and Funkhouser 2019). Thus, we add an adversarial loss term with weight λ_{adv} as in LeGendre et al. (2019). However, in contrast to this technique, we use a multi-scale GAN architecture that flows gradients from the discriminator to the generator network at multiple scales (Karnewar and Wang 2020), providing the discriminator with different sizes of both real and generated clipped mirror ball images.

3.4 Implementation Details

We use Tensorflow and the ADAM (Kinga and Ba 2015) optimizer with $\beta_1 = 0.9, \beta_2 = 0.999$, a learning rate of 0.00015 for the generator network, and, as is common, one 100× lower for the discriminator network, alternating between training the generator and discriminator. We set $\lambda_k = 0.2, 0.6, 0.2$ for the mirror, diffuse, and matte silver BRDFs respectively, set $\lambda_s = 1$ to weight all image scales equally, set $\lambda_{adv} = 0.004$, and use a batch size of 32. As the number of lighting environments is orders of magnitude larger than the number of subjects, we found that early stopping at 1.2 epochs appears to prevent over-fitting to subjects in the training set. We use the ReLU activation function for the generator network and the ELU activation function (Clevert et al. 2016) for the discriminator. To augment our dataset, we flip both the input images and lighting environments across the vertical axis.

Datasets. We split our 70 subjects into two groups: 63 for training and 7 for evaluation, ensuring that all expressions and camera views for a given subject belong to the same subset. We manually select the 7 subjects to include both skin tone and gender diversity. In total, for each of our 1 million lighting environments, we randomly select 8 OLAT sequences to relight from the training set (across subjects, facial expressions, and camera views), generating a training dataset of 8 million portraits with ground truth illumination (examples in Fig. 9). Using the same method, we capture lighting environments in both indoor and outdoor locations unseen in training to use for our evaluation, pairing these only with the evaluation subjects.

4 EVALUATION

In this section, we compare against prior techniques, perform an ablation study to investigate the performance gains for various sub-components, and measure performance across our diverse evaluation subjects. We also use our lighting estimates to render and composite virtual objects into real-world imagery.

4.1 Comparisons

Accurately estimated lighting should correctly render objects with arbitrary reflectance properties, so we test our model's performance using L_{rec}. This metric compares the appearance of three spheres (diffuse, matte silver, and mirror) as rendered with the ground truth versus estimated illumination. In Table 1, we compare our model against Sun et al. (2019), Calian et al. (2018), and a 2nd order SH decomposition of the ground truth lighting. We use our own implementation for Sun et al. (2019), training the model on our data for a fair comparison. As in the original implementation, we train the model with random crops from portraits composited over black backgrounds (not real-world imagery). As the method includes loss terms on both relit portraits and lighting, we generate 4 million portrait pairs from our original images and train the joint portrait relighting / lighting estimation model. To compare with Calian et al. (2018), the authors generously computed outdoor lighting for a set of portraits. However, the scale of their lighting depends on an albedo prior fit to a different dataset. So, for a best case comparison, we re-scale the author-provided illumination such that the total scene radiance matches that of the ground truth. Finally, we compare against the 2nd order SH decomposition, as this represents the best case scenario for any monocular face reconstruction technique that models illumination with this low frequency basis.

For the LDR image reconstruction losses, our model out-performs Sun et al. (2019) and Calian et al. (2018) for the diffuse and matte silver spheres. However, Sun et al. (2019) out-performs ours for the mirror sphere, as it its log-space loss on lighting is similar to L_{rec} for the mirror ball (but in HDR). As expected, the 2nd order SH approximation of the ground truth illumination out-performs our model for L_{rec} for the diffuse ball, since a low frequency representation of illumination suffices for rendering Lambertian materials. However, our model out-performs the 2nd order SH decomposition for L_{rec} for both the matte silver and mirror balls, with non-Lambertian BRDFs. This suggests that lighting produced by our model is better suited for rendering diverse materials.

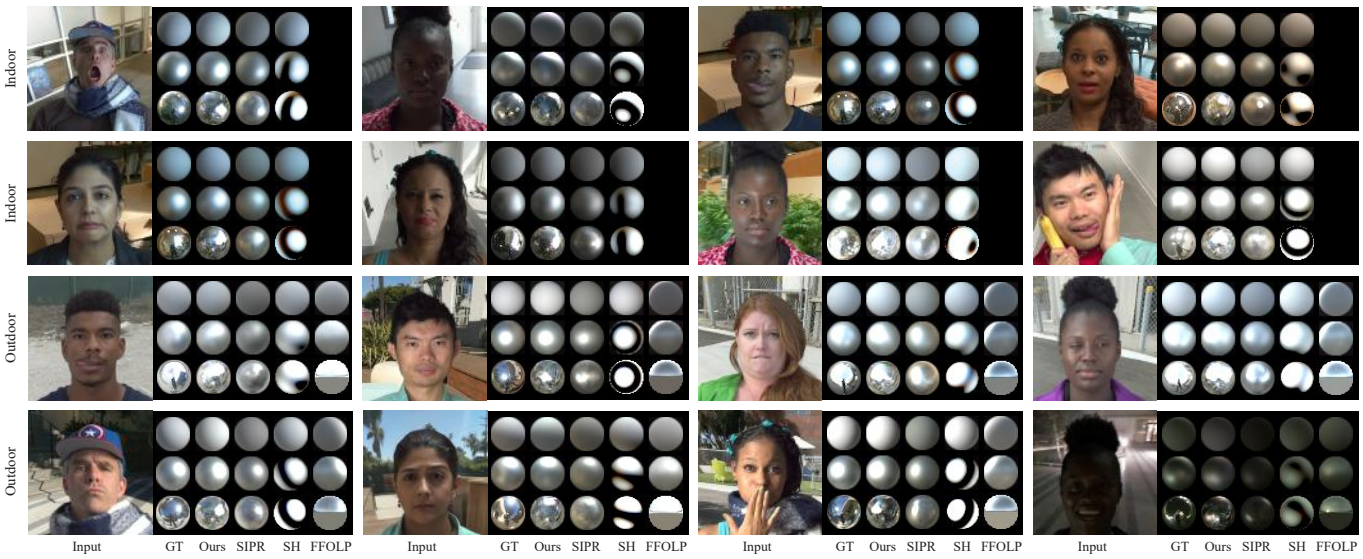


Fig. 10. Comparison sphere renderings (diffuse, matte silver, and mirror) for evaluation subjects and indoor and outdoor lighting environments. We compare our method against "Single Image Portrait Relighting" (SIPR) (Sun et al. 2019), the second order SH decomposition of the ground truth illumination, and for outdoor scenes, a radiance-scaled version of "From Faces to Outdoor Light Probes" (FFOLP) (Calian et al. 2018). Our model more faithfully recovers the total scene radiance compared with SIPR, and, unlike the SH decomposition, is useful for rendering materials with BRDFs beyond Lambertian.

Table 1. Comparison among methods: Average L_1 loss by BRDF [diffuse (d), mirror (m), and matte silver (s) spheres (in columns)], for evaluation portraits. We compare ground truth sphere images with those *rendered* using the HDR lighting inference, for *unseen* indoor (UI) and outdoor (UO) locations. ($^*n = 237$ for Calian et al. (2018) due to face tracking failures.)

	$L_{1(d)}$		$L_{1(s)}$		$L_{1(m)}$	
$n = 270^*$	UI	UO	UI	UO	UI	UO
Our model	**0.069**	**0.056**	**0.087**	**0.072**	0.181	0.157
2nd order SH of GT	**0.016**	**0.015**	0.120	0.109	0.306	0.247
Sun et al. (2019)	0.145	0.120	0.113	0.100	**0.154**	**0.139**
Calian et al. (2018)	–	0.158	–	0.163	–	0.215

In Table 2, we compare the relative radiance for each color channel for our model and that of Sun et al. (2019), computed as the sum of the pixels of the predicted illumination subtracted from the ground truth illumination, divided by the sum of the ground truth. We show that on average, the illumination recovered by the method of Sun et al. (2019) is missing 41% of the scene radiance. In contrast, for this randomly selected evaluation subset, our method adds on average 9% to the total scene radiance. As our rendering-based loss terms include matching the appearance of a diffuse ball, which is similar to a diffuse convolution of the HDR lighting environment, our method is able to more faithfully recover the total scene radiance.

Table 2. Average relative radiance difference [(GT - Pred) / GT] for estimated lighting, comparing our method and Sun et al. (2019).

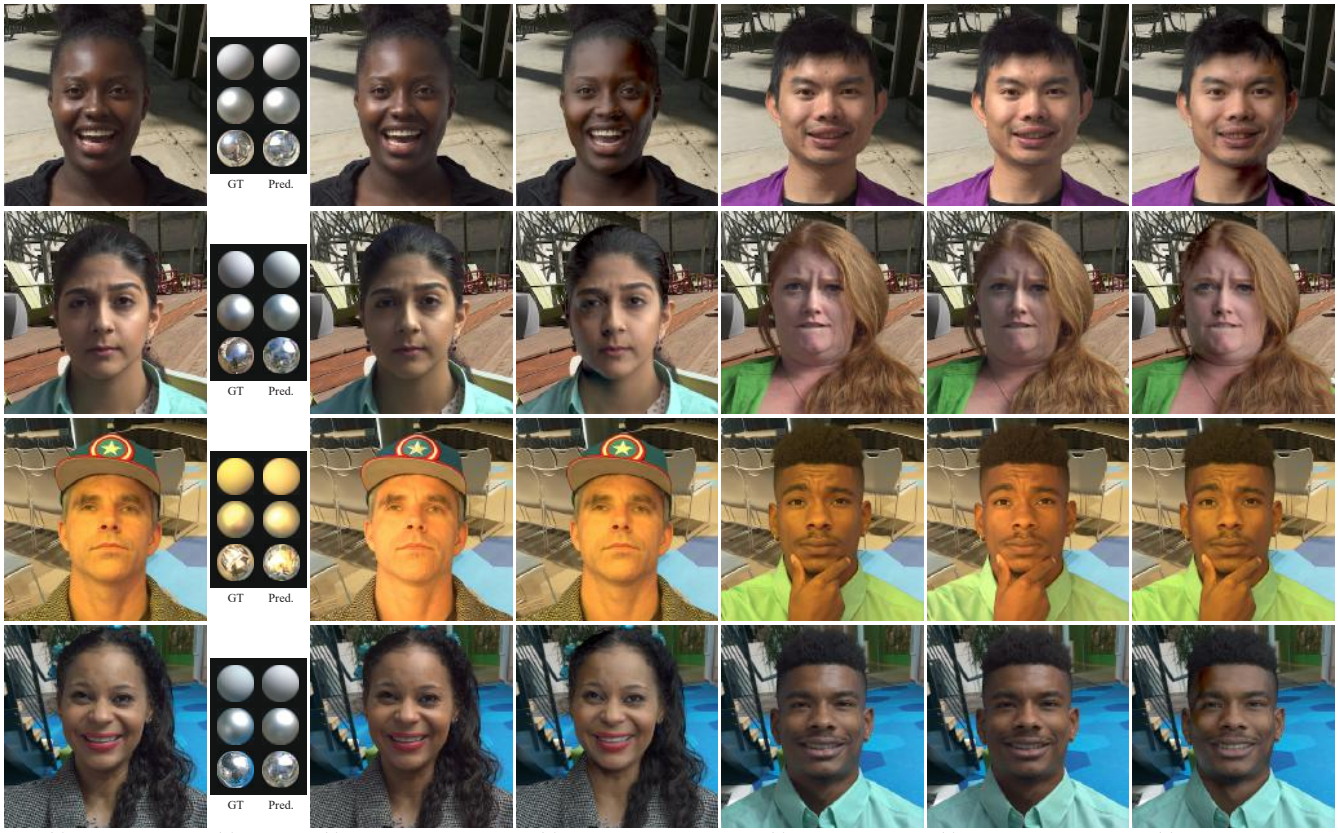
	Red Channel		Green Channel		Blue Channel	
$n = 270$	UI	UO	UI	UO	UI	UO
Our model	**-9.04%**	**-6.22%**	**-6.22%**	**-6.10%**	**-7.66%**	**-17.88%**
Sun et al. (2019)	34.53%	41.79%	38.31%	44.55%	39.73%	48.19%

In Fig. 10 we show qualitative results, rendering the three spheres using illumination produced using our method, that of Sun et al. (2019) labeled as "SIPR", that of a 2nd order SH decomposition, and that of Calian et al. (2018) for outdoor scenes, labeled as "FFOLP". The missing scene radiance from the method of Sun et al. (2019) is apparent looking at the diffuse sphere renderings, which are considerably darker than ground truth for this method. While the 2nd order SH approximation of the ground truth lighting produces diffuse sphere renderings nearly identical to the ground truth, Fig. 10 again shows how this approximation is ill-suited to rendering non-Lambertian materials. For the method of Calian et al. (2018), the sun direction is misrepresented as our evaluation lighting environments include a diversity of camera elevations, with the horizon line not exclusively along the equator of the mirror sphere.

In Fig. 11, we show an example where the illumination is estimated from a synthetic LDR portrait of a given subject (Fig. 11**a**), with the estimated and ground truth illumination in Fig. 11**b**. We then use both the estimated illumination from our model and the 2nd order SH approximation of the ground truth to light the same subject, shown in Fig. 11**c** and **d** respectively. For lighting environments with high frequency information (rows 1, 2, and 4 in Fig. 11), our lighting estimates produce portraits that more faithfully match the input images. These results highlight the limitation inherent in the Lambertian skin reflectance assumption.

4.2 Ablation Study

In Table 3, we report L_{rec} for each BRDF when evaluating each component of our system. We compare a baseline model using the single-scale losses (LeGendre et al. 2019) to our proposed model trained with multi-scale losses ($L_{\text{ms-rec}}$ and MSG-GAN). The multi-scale loss modestly decreases L_{rec} for both the diffuse and matte silver spheres, while increasing that of the mirror sphere. This increase is expected, as the adversarial loss for the mirror ball pulls the estimate away from an overly-blurred image that minimizes



(a) input image **(b)** lighting **(c)** lit with pred. from **a** **(d)** lit with SH of GT **(e)** lit with GT **(f)** lit with pred. from **a** **(g)** lit with SH of GT

Fig. 11. **(a)** Inputs to our model, generated using image-based relighting and a photographed reflectance basis for each evaluation subject. **(b)** Left: ground truth (GT) lighting used to generate **a**; Right: lighting estimated from **a** using our method. **(c)** The same subject lit with the predicted lighting. **(d)** The same subject lit with the 2nd order SH decomposition of the GT lighting. **(e)** A new subject lit with the GT lighting. **(f)** The new subject lit with the illumination estimated from **a** using our method. **(g)** The new subject lit with the 2nd order SH decomposition of the GT lighting. Our method produces lighting environments that can be used to realistically render virtual subjects into existing scenes, while the 2nd order SH lighting leads to an overly diffuse skin appearance.

L_{rec}. In Fig. 12, we show the visual impact of the multi-scale loss term, which synthesizes more high frequency details.

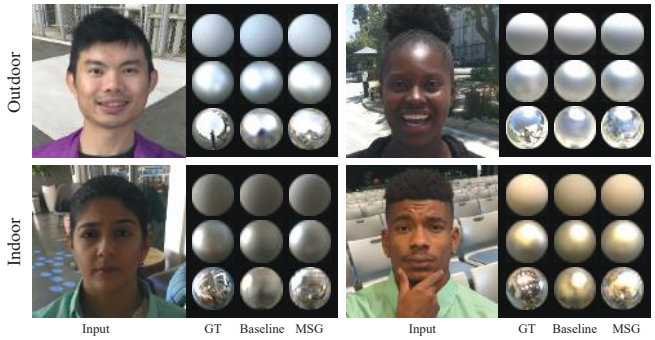


Fig. 12. Our multi-scale losses increase the sharpness of features in the recovered illumination, as shown in the mirror ball images (bottom rows), compared with baseline. Upper-right grid shown at +1 stop for display.

In Table 3, we also compare our baseline model, trained on images cropped using a face detector, to a model trained on random crops as in Sun et al. (2019), labeled "No Face Detector." The face detector

Table 3. Average L_1 loss by BRDF: diffuse (d), mirror (m), and matte silver (s) spheres (in columns), for lighting estimated from portraits of our evaluation subjects, using our technique with and without different features.

	$L_{1(d)}$		$L_{1(s)}$		$L_{1(m)}$	
Model, n = 3968	UI	UO	UI	UO	UI	UO
Proposed (no Multi-scale Losses)	0.054	0.050	0.076	0.069	**0.144**	**0.128**
No Face Detector	0.055	0.051	0.080	0.075	0.151	0.136
No Background Imagery	0.057	0.053	0.078	0.072	0.147	0.133
Proposed with Multi-scale Losses	**0.050**	**0.047**	**0.072**	**0.067**	0.156	0.141
log-L_2 Loss (as in Sun et al. (2019))	0.151	0.133	0.114	0.103	0.152	0.132
No Face (LeGendre et al. (2019))	0.136	0.135	0.144	0.137	0.174	0.166

imparts some modest improvement. Additionally, we compare our baseline model, trained on portraits composited onto real-world background imagery matching the ground truth illumination, to one trained without backgrounds, with subjects composited instead over black as in Sun et al. (2019). (The evaluation in this case is also performed on subjects against black backgrounds). The backgrounds also impart some modest improvement. We further show that our baseline model outperforms a model trained using the log-L_2 loss on HDR lighting of Sun et al. (2019). As this loss function does

not include a rendering step, this is somewhat expected. Finally, we compare against a model trained using *only* random crops of the background imagery, without portraits, using the single-scale loss terms. This table entry, labeled as "No Face," is equivalent to LeGendre et al. (2019), but trained on our background images and with our network architecture. As expected, the presence of faces in the input images significantly improves model performance.

4.3 Lighting Consistency for Diverse Skin Tones

In Table 4, we report L_{rec} for each of the three spheres individually, for 496 test examples in unseen indoor and outdoor lighting environments for each evaluation subject. Each example set includes diverse camera viewpoints, facial expressions, and hats/accessories. In Fig. 13, we plot the data of Table 4 to visualize that while there are some slight variations in L_{rec} across subjects, the model's performance appears similar across diverse skin tones.

Table 4. Average L_1 loss by BRDF: diffuse (d), mirror (m), and matte silver (s) spheres (in columns), for lighting estimated from portraits of our evaluation subjects, numbered 1-7 (see Fig. 13). This table corresponds with Fig. 13.

	$L_{1(d)}$		$L_{1(s)}$		$L_{1(m)}$	
n = 496	UI	UO	UI	UO	UI	UO
Subject 1	0.050	0.052	0.074	0.071	0.161	0.154
Subject 2	0.063	0.065	0.084	0.081	0.169	0.162
Subject 3	0.049	0.051	0.073	0.072	0.160	0.154
Subject 4	0.048	0.049	0.073	0.072	0.155	0.149
Subject 5	0.040	0.041	0.066	0.066	0.152	0.147
Subject 6	0.042	0.043	0.063	0.063	0.148	0.142
Subject 7	0.051	0.050	0.071	0.070	0.153	0.146

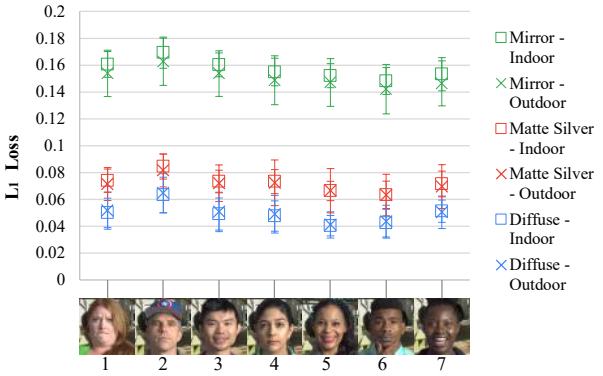


Fig. 13. Average L_{rec} for individual evaluation subjects, with n = 496 each for unseen indoor and outdoor scenes. This plot corresponds with Table. 4. Our model's performance is similar for subjects of diverse skin tones.

While L_{rec} is a useful metric, its absolute value operation masks the sign of residual error. To see whether radiance is missing or added to the predicted lighting for each subject, we also show the total relative radiance difference [(GT-Pred.)/GT] for each color channel for each subject in Fig. 14. The trend lines in Fig. 14 show that for evaluation subjects with smaller albedo values (measured as an average of each subject's forehead region), some energy in the estimated lighting is missing relative to the ground truth, with the inverse true for subjects with larger albedo values. For both indoor and outdoor scenes, this relative radiance difference is on average ±20% for evaluation subjects with very dark or very light skin tones, respectively, and smaller for subjects with medium skin tones. Nonetheless, as our evaluation subject with the lightest skin tone has an albedo value almost 3.5× that of our evaluation subject with the darkest skin tone, the network has mostly learned the correct scale of illumination across diverse subjects. In Fig. 15, we show examples where our model recovers similar lighting for different LDR input portraits of our evaluation subjects, where each is lit with the same ground truth illumination. In Fig. 11, we show that for a given input portrait (Fig. 11**a**), and lighting estimated from this portrait using our method Fig. 11**b**), we can accurately light a subject of a different skin tone (Fig. 11**f**) *without* adjusting the scale of the illumination and composite them into the original image, closely matching that subject's ground truth appearance (Fig. 11**e**). An additional such example is shown in Fig. 1.

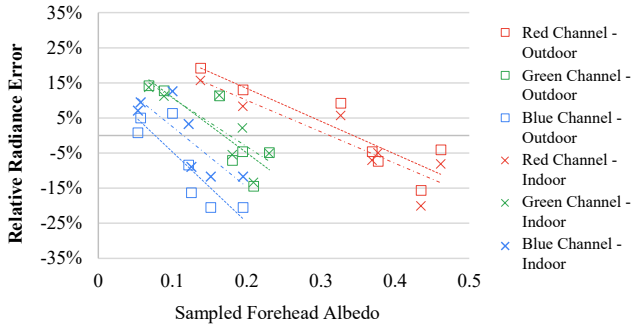


Fig. 14. y axis: Average relative error in total radiance [(GT-Pred.)/GT] for each color channel for each of our evaluation subjects (n = 496 each for unseen indoor and outdoor scenes). x axis: Each subject's average RGB albedo, sampled from the forehead under a unit sphere of illumination.

4.4 Lighting Consistency across Head Poses

We did not observe any marked differences in the lighting estimated for a given subject for different head poses or facial expressions. In Fig. 16, we show that similar illumination is recovered for different camera views and expressions for one of the evaluation subjects.

4.5 Real-World Results

In Fig. 20 we show lighting estimation from real-world portraits in-the-wild, for a diverse set of subjects, including one wearing a costume with face-paint. While ground truth illumination is not available, the sphere renderings produced using our lighting inference look qualitatively plausible. These results suggest that our model has generalized well to arbitrary portraits.

5 APPLICATIONS

Mobile Augmented Reality. Our lighting inference runs in real-time on a mobile device (CPU: 27.5 fps, GPU: 94.3 fps on a Google Pixel 4 smartphone), enabling real-time rendering and compositing of virtual objects for smartphone AR applications. We show our inference running in real-time in our supplemental video.

Digital Double Actor Replacement. In Fig. 17, we estimate lighting from in-the-wild portraits (**a**), and then light a virtual character to composite into the original scene with consistent illumination.

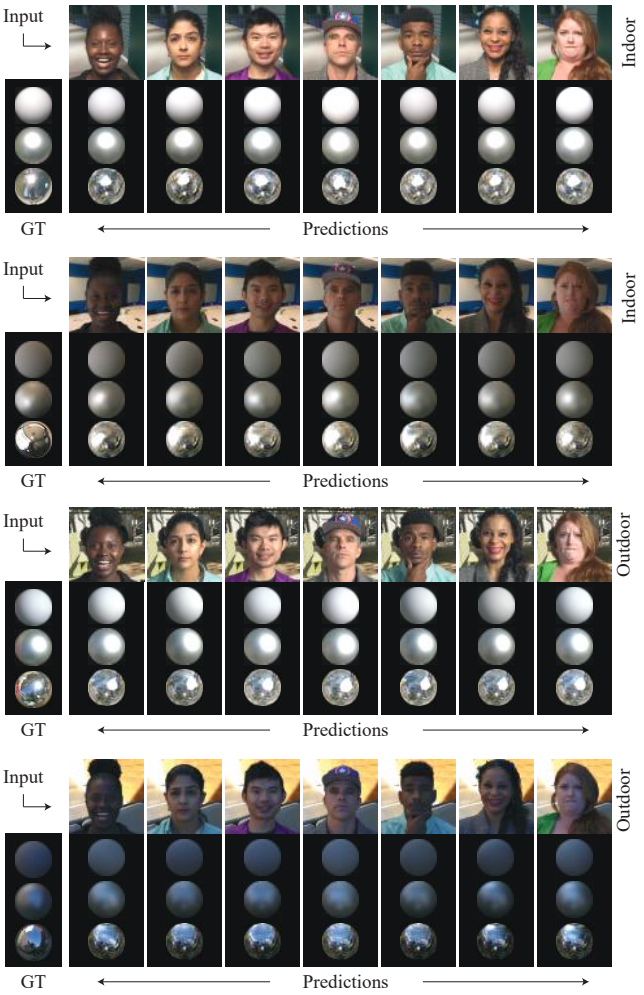


Fig. 15. Left: spheres rendered with the ground truth illumination. Remaining columns: spheres rendered with the illumination produced using our technique, for input portraits of different subjects all lit with the same ground truth illumination. Our model recovers lighting at a similar scale for LDR input portraits of subjects with a variety of skin tones.

These examples suggest that our method could be used for digital double actor replacement, without on-set lighting measurements.

Post-Production Virtual Object Compositing. In Fig. 18 we render and composite a set of shiny virtual balloons into a "selfie" portrait, using lighting estimates produced by our method. We show a version with motion in our supplemental video.

6 LIMITATIONS

As our method relies on a face detector, it fails if no face is detected. Fig. 19 shows two other failure modes: an example where a saturated pink hat not observed in training leads to an erroneous lighting estimate, and an example where the illumination color is incorrectly estimated for an input environment with unnatural color balance. This input example was generated by scaling the red channel of the ground truth illumination by a factor of 3. Future work could address the first limitation with additional training data spanning a broader range of accessories, while the second limitation could be addressed with data augmentation via adjusting the white balance of

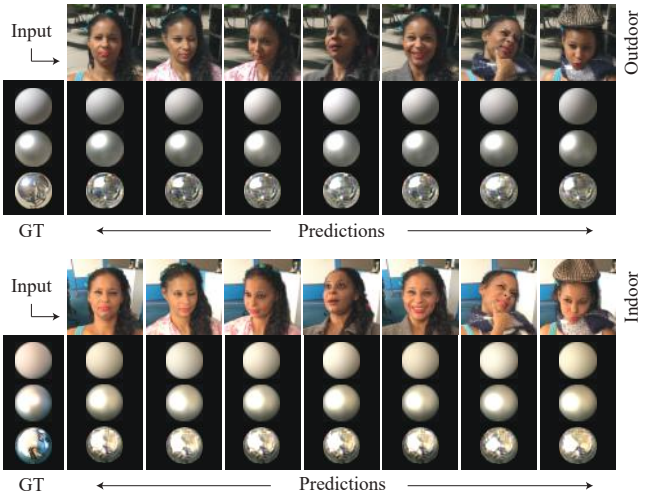


Fig. 16. Left: spheres rendered with the ground truth illumination. Remaining columns: spheres rendered with the illumination produced using our technique, for input portraits of the same subject with different head poses and expressions, lit with the same illumination. Our method recovers similar lighting across facial expressions and head poses.

(a) input in-the-wild portrait **(b)** lighting **(c)** digital character lit with **b**

Fig. 17. **(a)** In-the-wild input portraits. **(b)** Lighting estimated by our technique. **(c)** A digital human character rendered with the predicted illumination, composited into the original scene. Digital character model by Ian Spriggs, rendered in V-Ray with the VRayAISurfaceSkin shader.

Fig. 18. Virtual balloons composited into "selfie" portraits using lighting estimated by our technique.

the ground truth illumination. Finally, our lighting model assumes distant illumination, so our method is not able to recover complex local lighting effects.

Fig. 19. Two example failure cases. Left: A brightly-hued hat not observed in training. Right: Input lighting environment with non-natural color balance.

7 CONCLUSION

We have presented a learning-based technique for estimating omnidirectional HDR illumination from a single LDR portrait image. Our model was trained using a photo-realistic, synthetically-rendered dataset of portraits with ground truth illumination generated using reflectance fields captured in a light stage, along with more than one million lighting environments captured using an LDR video-rate technique, which we promoted to HDR using a novel linear solver formulation. We showed that our method out-performs both the previous state-of-the-art in portrait-based lighting estimation, and, for non-Lambertian materials, a low-frequency, second order spherical harmonics decomposition of the ground truth illumination. We are also, to the best of our knowledge, the first to explicitly evaluate our lighting estimation technique for subjects of diverse skin tones, while demonstrating recovery of a similar scale of illumination for different subjects. Our technique runs in real-time on a mobile device, suggesting its usefulness for improving the photo-realism of face-based augmented reality applications. We further demonstrated our method's utility for post-production visual effects, showing that digital characters can be composited into real-world photographs with consistent illumination learned by our model.

REFERENCES

Sameer Agarwal, Keir Mierle, and Others. [n. d.]. Ceres Solver. http://ceres-solver.org. ([n. d.]).

Robert Anderson, David Gallup, Jonathan T Barron, Janne Kontkanen, Noah Snavely, Carlos Hernández, Sameer Agarwal, and Steven M Seitz. 2016. Jump: virtual reality video. *ACM Transactions on Graphics (TOG)* 35, 6 (2016), 1–13.

Jonathan T Barron and Jitendra Malik. 2014. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence* 37, 8 (2014), 1670–1687.

Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. 1978. Recovering intrinsic scene characteristics. *Comput. Vis. Syst* 2 (1978), 3–26.

Ronen Basri and David W Jacobs. 2003. Lambertian reflectance and linear subspaces. *IEEE transactions on pattern analysis and machine intelligence* 25, 2 (2003), 218–233.

Valentin Bazarevsky, Yury Kartynnik, Andrey Vakunov, Karthik Raveendran, and Matthias Grundmann. 2019. Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047* (2019).

Peter N Belhumeur, David J Kriegman, and Alan L Yuille. 1999. The bas-relief ambiguity. *International journal of computer vision* 35, 1 (1999), 33–44.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. 187–194.

Dan A Calian, Jean-François Lalonde, Paulo Gotardo, Tomas Simon, Iain Matthews, and Kenny Mitchell. 2018. From Faces to Outdoor Light Probes. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 51–61.

Dachuan Cheng, Jian Shi, Yanyun Chen, Xiaoming Deng, and Xiaopeng Zhang. 2018. Learning Scene Illumination by Pairwise Photos from Rear and Front Mobile Cameras. In *Computer Graphics Forum*, Vol. 37. Wiley Online Library, 213–221.

Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. 2016. Fast and Accurate Deep Network Learning by Exponential Linear Units (ELUs). In *International Conference on Learning Representations (ICLR)*.

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 and interactive techniques*. ACM, 189–198.

Paul Debevec. 2006. Image-based lighting. In *ACM SIGGRAPH 2006 Courses*.

Paul Debevec, Paul Graham, Jay Busch, and Mark Bolas. 2012. A single-shot light probe. In *ACM SIGGRAPH 2012 Talks*. ACM, 10.

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 the 27th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 145–156.

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.

Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. 2018. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *International Journal of Computer Vision* 126, 12 (2018), 1269–1287.

Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. 2017. Learning to Predict Indoor Illumination from a Single Image. *ACM Trans. Graph.* 36, 6, Article 176 (Nov. 2017), 14 pages. https://doi.org/10.1145/3130800.3130891

Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. 2019. Fast Spatially-Varying Indoor Lighting Estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6908–6917.

Yannick Hold-Geoffroy, Akshaya Athawale, and Jean-François Lalonde. 2019. Deep sky modeling for single image outdoor lighting estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6927–6935.

Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. 2017. Deep outdoor illumination estimation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, Vol. 2.

Berthold KP Horn. 1970. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. (1970).

Animesh Karnewar and Oliver Wang. 2020. Msg-gan: Multi-scale gradients for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7799–7808.

Ira Kemelmacher-Shlizerman and Ronen Basri. 2010. 3D face reconstruction from a single image using a single reference face shape. *IEEE transactions on pattern analysis and machine intelligence* 33, 2 (2010), 394–405.

D Kinga and J Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, Vol. 5.

Sebastian B Knorr and Daniel Kurz. 2014. Real-time illumination estimation from faces for coherent rendering. In *2014 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE, 113–122.

Jean-François Lalonde, Alexei A Efros, and Srinivasa G Narasimhan. 2009. Estimating natural illumination from a single outdoor image. In *Computer Vision, 2009 IEEE*

Fig. 20. Diffuse, matte silver, and mirror spheres rendered using illumination estimated using our technique from the input portraits in-the-wild.

12th International Conference on. IEEE, 183–190.

Jean-François Lalonde and Iain Matthews. 2014. Lighting estimation in outdoor image collections. In *3D Vision (3DV), 2014 2nd International Conference on*, Vol. 1. IEEE, 131–138.

Edwin H Land and John J McCann. 1971. Lightness and retinex theory. *Josa* 61, 1 (1971), 1–11.

Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. 2019. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5918–5928.

Stephen Lombardi and Ko Nishino. 2016. Reflectance and illumination recovery in the wild. *IEEE transactions on pattern analysis and machine intelligence* 38, 1 (2016), 129–141.

Stephen R Marschner and Donald P Greenberg. 1997. Inverse lighting for photography. In *Color and Imaging Conference*, Vol. 1997. Society for Imaging Science and Technology, 262–265.

Jeffry S Nimeroff, Eero Simoncelli, and Julie Dorsey. 1995. Efficient re-rendering of naturally illuminated environments. In *Photorealistic Rendering Techniques*. Springer, 373–388.

Ko Nishino and Shree K Nayar. 2004. Eyes for relighting. In *ACM Transactions on Graphics (TOG)*, Vol. 23. ACM, 704–711.

Ravi Ramamoorthi and Pat Hanrahan. 2001a. On the relationship between radiance and irradiance: determining the illumination from images of a convex Lambertian object. *JOSA A* 18, 10 (2001), 2448–2459.

Ravi Ramamoorthi and Pat Hanrahan. 2001b. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 117–128.

Erik Reinhard, Wolfgang Heidrich, Paul Debevec, Sumanta Pattanaik, Greg Ward, and Karol Myszkowski. 2010. *High dynamic range imaging: acquisition, display, and image-based lighting*. Morgan Kaufmann.

Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. 2018. SfSNet: learning shape, reflectance and illuminance of faces in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6296–6305.

Davoud Shahlaei and Volker Blanz. 2015. Realistic inverse lighting from a single 2d image of a face, taken under unknown and complex lighting. In *2015 11th IEEE international conference and workshops on automatic face and gesture recognition (FG)*, Vol. 1. IEEE, 1–8.

Hyunjung Shim. 2012. Faces as light probes for relighting. *Optical Engineering* 51, 7 (2012), 077002.

Zhixin Shu, Sunil Hadap, Eli Shechtman, Kalyan Sunkavalli, Sylvain Paris, and Dimitris Samaras. 2017a. Portrait lighting transfer using a mass transport approach. *ACM Transactions on Graphics (TOG)* 36, 4 (2017), 1.

Zhixin Shu, Ersin Yumer, Sunil Hadap, Kalyan Sunkavalli, Eli Shechtman, and Dimitris Samaras. 2017b. Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5541–5550.

Shuran Song and Thomas Funkhouser. 2019. Neural illumination: Lighting prediction for indoor environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6918–6926.

Jessi Stumpfel, Chris Tchou, Andrew Jones, Tim Hawkins, Andreas Wenger, and Paul Debevec. 2004. Direct HDR capture of the sun and sky. In *Proceedings of the 3rd international conference on Computer graphics, virtual reality, visualisation and interaction in Africa*. ACM, 145–149.

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 Trans. Graph.* 38, 4, Article Article 79 (July 2019), 12 pages. https://doi.org/10.1145/3306346.3323008

A. Tewari, M. Zollhöfer, P. Garrido, F. Bernard, H. Kim, P. Pérez, and C. Theobalt. 2018. Self-supervised Multi-level Face Model Learning for Monocular Reconstruction at over 250 Hz. In *Proceedings of Computer Vision and Pattern Recognition (CVPR 2018)*.

Ayush Tewari, Michael Zollhöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Pérez, and Christian Theobalt. 2017. Mofa: Model-based deep convolutional face autoencoder for unsupervised monocular reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, Vol. 2. 5.

Andrei Tkachenka, Gregory Karpiak, Andrey Vakunov, Yury Kartynnik, Artsiom Ablavatski, Valentin Bazarevsky, and Siargey Pisarchyk. 2019. Real-time Hair Segmentation and Recoloring on Mobile GPUs. *arXiv preprint arXiv:1907.06740* (2019).

Luan Tran, Feng Liu, and Xiaoming Liu. 2019. Towards high-fidelity nonlinear 3D face morphable model. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1126–1135.

Luan Tran and Xiaoming Liu. 2019. On learning 3d face morphable model from in-the-wild images. *IEEE transactions on pattern analysis and machine intelligence* (2019).

Jonas Unger, Stefan Gustavson, and Anders Ynnerman. 2006. Densely Sampled Light Probe Sequences for Spatially Variant Image Based Lighting. In *Proceedings of*

the 4th international conference on Computer graphics and interactive techniques in Australasia and Southeast Asia (GRAPHITE) (GRAPHITE 06). ACM, 341–347.

Jamie Waese and Paul Debevec. 2002. P.: A real-time high dynamic range light probe. In *In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques: Conference Abstracts and Applications, p. 247. ACM Press/Addison-Wesley Publishing Co.*

Andreas Wenger, Andrew Gardner, Chris Tchou, Jonas Unger, Tim Hawkins, and Paul Debevec. 2005. Performance relighting and reflectance transformation with time-multiplexed illumination. *ACM Transactions on Graphics (TOG)* 24, 3 (2005), 756–764.

Shuco Yamaguchi, Shunsuke Saito, Koki Nagano, Yajie Zhao, Weikai Chen, Kyle Olszewski, Shigeo Morishima, and Hao Li. 2018. High-fidelity Facial Reflectance and Geometry Inference from an Unconstrained Image. *ACM TOG* (2018).

Renjiao Yi, Chenyang Zhu, Ping Tan, and Stephen Lin. 2018. Faces as lighting probes via unsupervised deep highlight extraction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 317–333.

Yizhou Yu, Paul Debevec, Jitendra Malik, and Tim Hawkins. 1999. Inverse global illumination: Recovering reflectance models of real scenes from photographs. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*. ACM Press/Addison-Wesley Publishing Co., 215–224.

Jinsong Zhang, Kalyan Sunkavalli, Yannick Hold-Geoffroy, Sunil Hadap, Jonathan Eisenman, and Jean-François Lalonde. 2019. All-weather deep outdoor lighting estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10158–10166.

Richard Zhang. 2019. Making convolutional networks shift-invariant again. *arXiv preprint arXiv:1904.11486* (2019).

Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W Jacobs. 2019. Deep single-image portrait relighting. In *Proceedings of the IEEE International Conference on Computer Vision*. 7194–7202.

Hao Zhou, Jin Sun, Yaser Yacoob, and David W Jacobs. 2018. Label Denoising Adversarial Network (LDAN) for Inverse Lighting of Faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6238–6247.