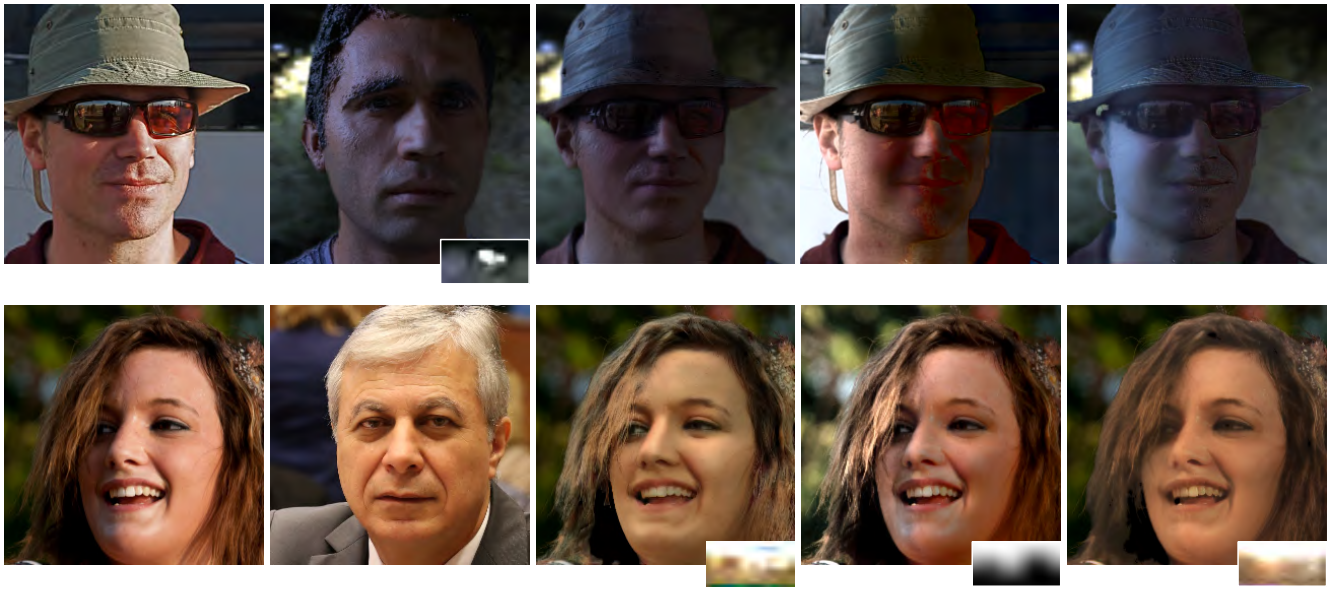
Single Image Portrait Relighting via Explicit Multiple Reflectance Channel Modeling

ZHIBO WANG, BNRist and school of software, Tsinghua University, China
XIN YU, University of Technology Sydney, Australia
MING LU, Intel Labs, China
QUAN WANG, SenseTime, China
CHEN QIAN, SenseTime, China
FENG XU, BNRist and school of software, Tsinghua University, China



(a) Input (b) Reference (c) Our method (d) [Zhou et al. 2019] (e) [Sun et al. 2019]

Fig. 1. Relit results of our system. Given a real-world portrait (a), our method generates a photorealistic relit image (c) according to a target environment map (b, top) or a reference image (b, bottom). When relighting using the environment map, we render an arbitrary face scan with the target lighting as a reference to better illustrate the lighting, not for using in our method. When relighting using the reference image, we show the estimated targeting lighting for each of the compared methods. Compared with previous methods [Sun et al. 2019; Zhou et al. 2019], our relit results are more convincing and consistent with the target lighting. The top result shows that our method can remove the original highlight, generate new highlight according to the target lighting even for the glasses and the hat. The bottom result shows that our method can remove the shadow on the right face to fit the target lighting. Images courtesy: Flickr user *Nathan Forget* ((a)-1), Flickr user *popo.uw23* ((a)-2), Flickr user *Luca Boldrini* ((b)-2).

Authors' addresses: Zhibo Wang, BNRist and school of software, Tsinghua University, 100091, Beijing, China, wzb17@mails.tsinghua.edu.cn; Xin Yu, University of Technology Sydney, Australia, xin.yu@uts.edu.au; Ming Lu, Intel Labs, China, lu199192@gmail.com; Quan Wang, SenseTime, China, wangquan@sensetime.com; Chen Qian, SenseTime, China, qianchen@sensetime.com; Feng Xu, BNRist and school of software, Tsinghua University, 100091, Beijing, China, xufeng2003@gmail.com.

0730-0301/2020/12-ART220 $15.00
https://doi.org/10.1145/3414685.3417824

Portrait relighting aims to render a face image under different lighting conditions. Existing methods do not explicitly consider some challenging lighting effects such as specular and shadow, and thus may fail in handling extreme lighting conditions. In this paper, we propose a novel framework that explicitly models multiple reflectance channels for single image portrait relighting, including the facial albedo, geometry as well as two lighting effects, *i.e.*, specular and shadow. These channels are finally composed to generate the relit results via deep neural networks. Current datasets do not support learning such multiple reflectance channel modeling. Therefore, we present a large-scale dataset with the ground-truths of the channels, enabling us to train the deep neural networks in a supervised manner. Furthermore, we develop a novel module named Lighting guided Feature Modulation (LFM). In contrast to existing methods which simply incorporate the given lighting

in the bottleneck of a network, LFM fuses the lighting by layer-wise feature modulation to deliver more convincing results. Extensive experiments demonstrate that our proposed method achieves better results and is able to generate challenging lighting effects.

CCS Concepts: • **Computing methodologies → Image-based rendering**; **Computational photography**; **Neural networks**.

Additional Key Words and Phrases: Portrait relighting, Image-based relighting, Deep neural rendering

ACM Reference Format:
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 Trans. Graph.* 39, 6, Article 220 (December 2020), 13 pages. https://doi.org/10.1145/3414685.3417824

1 INTRODUCTION

Lighting plays a critical role in portrait photography. In order to achieve a particular look, photographers often require complex equipment and sophisticated expertise to obtain a good lighting setup. This may be impractical for amateur photographers who take photos with consumer-level cameras. In this case, attaining convincing visual effects or rectifying the portrait images captured in extreme illumination conditions is highly desirable. Hence, portrait relighting techniques have become demanded and attracted significant attentions in both industry and academia.

Although portrait relighting has been extensively investigated in the past decades [Aldrian and Smith 2012; Debevec et al. 2000; Shu et al. 2017a; Sun et al. 2019], it still remains as a challenging problem. To ensure the relit results to be photorealistic, Debevec et al. [2000] exploit a sophisticated system to capture the reflectance fields of human faces and then render novel face images with different illuminations. While their system is able to manipulate lighting accurately, it relies on the specific hardware system and thus is not suitable for consumer-level usage.

To ease the specific hardware requirements of portrait relighting, many methods have been proposed to transfer the color distribution of a reference portrait image to the input portrait image. For instance, Shih et al. [2014] first densely align a reference face image to the input portrait, and then transfer the colors of the reference face to the input one. Shu et al. [2017a] further exploit 3D facial geometry to assist color transfer more authentically, and employ optimal transport [Pitie et al. 2005] to map the color distribution of a reference portrait to the input. Although these methods achieve pleasing results, they require a large amount of computation and thus are not suitable for interactive photographic applications. In addition, the generation of the lighting effects, such as specular and shadow, highly depends on the practical shading pipeline which cannot be fully modeled by just transferring color distributions.

Recently, efficient methods based on deep learning [Sun et al. 2019; Zhou et al. 2019] have been proposed for portrait relighting following practical shading pipelines, which usually employ the lighting models of Spherical Harmonics (SH) [Zhou et al. 2019] or environment maps [Sun et al. 2019]. In general, those methods simply learn an end-to-end mapping between the input headshots and relit ones. Important lighting effects, *i.e.*, specular and shadow, are not carefully treated in these end-to-end learning methods as they are not numerically significant in the integrated image loss. Also, meaningful channels like facial albedo and geometry [Sengupta et al. 2018] are also not explicitly considered. This motivates us to explicitly exploit multiple reflectance channel modeling for photorealistic portrait relighting.

In this paper, we propose a novel single image portrait relighting framework using deep neural networks. We first develop a de-lighting network to jointly recover the facial albedo and geometry, as the intrinsic channels, from an input portrait. Based on the estimated facial geometry and a given lighting, we synthesize the specular and shadow effects, as our rendering detail channels, via a Specular and Shadow estimation (SS) network. After obtaining the four channels, we present a composition network to generate the relit portrait image under the given environment map. The SS and composition networks form our relighting network. Since our framework explicitly models multiple reflectance channels, our method is able to achieve authentic lighting effects, especially for specular and shadow.

However, the key barrier to learn our networks is the lack of ground-truth supervision. Even though there are many 3D face datasets available [Baocai et al. 2009; Cao et al. 2013; Cosker et al. 2011; Savran et al. 2008; Yin et al. 2006; Zhang et al. 2013, 2014], they do not provide channels of facial albedo, geometry and lighting effects. Therefore, we present a large-scale relighting dataset. We first reconstruct the high-quality 3D faces of 438 subjects across different races and ages, and each subject performs 20 expressions. Then we randomly choose 100 subjects and render them under 391 different lighting conditions with random head poses, to generate the supervision channels for training. Once our networks are trained, our method well handles both synthetic and real-world portraits, which is demonstrated by extensive experimental results.

As for the network design, existing methods [Sun et al. 2019; Zhou et al. 2019] mainly encode the target lighting into the bottleneck of a network. We propose a Lighting guided Feature Modulation (LFM) module, which fuses the lighting into the channel features in a layer-wise fashion. Our LFM fuses lighting information into our multi-channel features effectively and forces the features to be more consistent with respect to the target lighting. Experiments show that by using LFM we are able to generate more natural relit results in accordance with the given lighting.

Overall, our contributions are summarized as follows:

- We propose a novel single image relighting framework via explicitly modeling the face intrinsic channels and the rendering detail channels, which dramatically improves the relighting performance, especially for challenging lightings with specular and shadow.
- We present the first large-scale portrait relighting dataset composed of high-quality 3D faces and multiple rendered channels for training headshot relighting networks.
- We propose a new module, namely Lighting-guided Feature Modulation (LFM), to incorporate the target lighting more effectively and authentically into relit portraits.

Table 1. Summery of popular widely-used 3D face datasets.

Dataset	Id. Num	Exp. Num	Vert. Num	Tex. Resolution	Device	Multiple channels
BU-3DFE [Yin et al. 2006]	100	25	10k-20k	1300×900	structure light	No
BU-4DFE [Zhang et al. 2013]	101	6	10k-20k	1040×1329	structure light	No
Bosphorus [Savran et al. 2008]	105	35	$\approx 35k$	1600×1200	structure light	No
FaceWarehouse [Cao et al. 2013]	150	20	$\approx 11k$	640×480	kinect	No
4DFAB [Cheng et al. 2018]	180	6	$\approx 100k$	1200×1600	kinect and 6 cameras	No
D3DFACS [Cosker et al. 2011]	10	38	$\approx 30k$	1024×1280	6 cameras	No
BP4D-Spontanous [Zhang et al. 2014]	41	27	$\approx 37k$	1024×1392	3 cameras	No
FaceScape [Yang et al. 2020]	938	20	$\approx 2m$	4096×4096	68 cameras	No
ICT-3DRFE [Stratou et al. 2011]	23	15	$\approx 1.2m$	1296×1944	structure light	No
Ours	438	20	$\approx 900k$	4096×4096	30 cameras	**Yes**

2 RELATED WORK

2.1 Portrait Relighting

Debevec et al. [2000] develop a hardware system, named Light Stage, to capture the reflectance field of a human face. Then, relighting effects are achieved by rendering 3D faces in novel illuminations and viewpoints. Although Light Stage attains promising visual results, this complicated system is not suitable for consumer-level usage. Hardware-free methods have been proposed for single image portrait relighting. Inspired by the works of color transfer, Chen et al. [2011, 2013] first decompose the input images into multiple layers and then use an edge-preserving filter to transfer the colors from a source face to a target face. Shih et al. [2014] first align a content face to a style face captured in the desired lighting condition via dense correspondence [Liu et al. 2010], and then transfer the local color distribution by the multiscale technique [Malik and Perona 1990]. Song et al. [2017] further extend the method of [Shih et al. 2014] to multiple style faces. Shu et al. [2017a] fit a 3D face to the input face image and employ optimal transport [Pitie et al. 2005] to transfer facial color distributions, thus obtaining relit results.

Instead of following the pipeline of color transfer, some works directly modify the parameters of lighting models to achieve the relighting effects. Blanz and Vetter [1999] adopt directional lighting model in their 3D morphable face model. Portrait lighting can be manipulated by changing the parameters of directional lighting model. Aldrian and Smith [2012]; Egger et al. [2018]; Wang et al. [2007, 2008] introduce Spherical Harmonics (SH) lighting model [Basri and Jacobs 2003; Ramamoorthi and Hanrahan 2001] into portrait relighting. They jointly estimate the 3D face and SH parameters to recover the facial geometry and lighting by numerical optimization. Then portrait relighting is achieved by modifying the parameters of SH lighting model and rendering the 3D face into images.

Deep learning techniques have shown their potential in photorealistic lighting effect manipulation, *e.g.* lighting normalization [Nagano et al. 2019; Zhang et al. 2020b], shadow removal [Zhang et al. 2020a]. Very recently, several methods are proposed to speed up the numerical optimization [Nestmeyer et al. 2020; Sengupta et al. 2018; Shu et al. 2017b; Sun et al. 2019; Zhou et al. 2019]. The methods [Sengupta et al. 2018; Shu et al. 2017b] relight an facial image by performing an image intrinsic decomposition via deep neural networks. As those methods use low quality face models and an SH lighting model, they only achieve coarse facial geometry and albedo. Thus, they fail to generate high quality relit results. Zhou et al. [2019] and Sun et al. [2019] use end-to-end trainable deep networks to directly obtain the desired relit face. Zhou et al. [2019] also employ the SH lighting model and train a relighting network with synthetic data. Due to the simplicity of SH rendering, the generated synthetic data are not realistic enough and thus restrict the relighting performance. Instead of using SH rendering models, Sun et al. [2019] propose to use an environment map to render realistic illumination conditions. In this work, we also adopt environment maps to build our lighting model. However, unlike previous works [Sun et al. 2019; Zhou et al. 2019] that train relighting networks in an end-to-end fashion, we found that by explicitly modeling multiple reflectance channels of facial albedo, geometry and lighting effects, we can actually better generate some challenging effects, *i.e.*, specular and shadow, given high quality facial geometry and an advanced lighting model. Recently, Nestmeyer et al. [2020] also explicitly model the shadow and specular and achieve similar conclusions. However, their method mainly focuses on image relighting under directional lightings. Since existing datasets do not provide supervision for learning such explicit modeling, we construct a large-scale relighting dataset to obtain these channels.

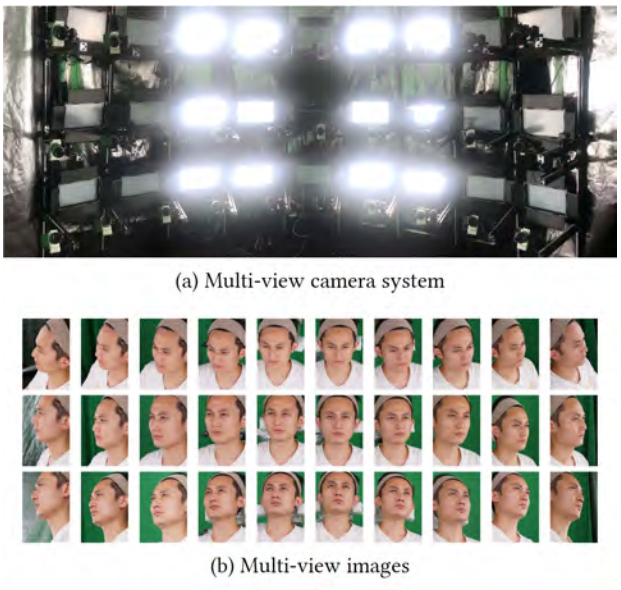
2.2 2D and 3D Face Dataset

Traditional 2D face datasets usually take various lighting conditions into consideration. For example, Gross et al. [2010] present a dataset consisting of 337 subjects with a range of facial expressions. For each subject, the face images are captured under 19 illumination conditions. Lee et al. [2005] build a dataset of 28 subjects under 9 poses and 64 shading situations. Gao et al. [2007] contribute a dataset of 1040 subjects and 15 lighting conditions. Although 2D face datasets with controlled lighting conditions are easy to build, they lack important 3D facial geometry information for photorealistic portrait relighting.

With the development of 3D sensors, many 3D face datasets become publicly available. We list several popular widely-used 3D face datasets in Table. 1. Yin et al. [2006] employ structure light to build a 3D face dataset consisting of 100 subjects and each subject has 25 expressions. Zhang et al. [2013] focus on dynamic 3D faces and capture 101 subjects with 6 expressions using structure light techniques. Cheng et al. [2018] build a similar dynamic 3D face dataset using 6 RGB cameras and a Kinect sensor. Savran et al. [2008] capture 105 subjects with 35 expressions under various facial

poses and occlusions. Cao et al. [2013] use Kinect to construct a dataset of 150 subjects with 20 expressions. Cosker et al. [2011] focus on dynamic Facial Action Coding System (FACS) data and construct a dataset with 519 sequences with Action Unit (AU) annotations. Recently, Yang et al. [2020] build a new high-quality 3D face dataset using a multi-view reconstruction technique. Based on the raw 3D faces, many works further register one template face mesh to the raw scans to deliver registered datasets [Blanz and Vetter 1999; Booth et al. 2016; Cao et al. 2013; Li et al. 2017]. All these datasets are not designed for the portrait relighting purpose and thus they do not contain multiple reflectance channel information.

Stratou et al. [2011] use structured-light stereo to reconstruct a dataset with high quality facial models with albedo and normals for both diffuse and specular rendering. However, they do not provide images rendered under various lighting conditions for the training purpose. We use a multi-view reconstruction technique to build a high-quality dataset with detailed 3D face textures. Furthermore, our high-quality dataset is rendered under 391 different lighting conditions while generating ground-truths of the multiple reflectance channels. Therefore, benefiting from our dataset, we are able to train our networks and significantly improve the relit quality in comparison to prior arts.



(a) Multi-view camera system

(b) Multi-view images

Fig. 2. Our multi-view capture system. We capture 30 images (b) simultaneously by a multi-view capture system (a). The system contains 30 Sony A5000 cameras and 27 area white lights.

3 PORTRAIT RELIGHTING DATASET

In this section, we introduce a new portrait relighting dataset. We collect high-quality face scans of 438 subjects with diverse races, different ages ranging from 17 to 69 years old and 20 facial expressions under a single uniform lighting setup using a multi-view camera system. Then, we render the face scans in different poses under various lighting conditions. The albedo I_n, normal N, specular I_{sp}, shadow I_{sh} and face parsing P channels are also provided for each image I.

3.1 3D Face Acquisition

To build our dataset, we firstly collect high fidelity face scans reconstructed from multi-view images taken by our multi-view camera system as shown in Figure 2. Our system contains 30 Sony A5000 cameras and 27 white area lights to create a uniform white lighting environment. We also ask all the subjects to carefully clean their facial skins. In this case, we believe the reconstructed texture is close to the subject's albedo. To enrich facial expressions in our dataset, subjects are asked to perform 20 predefined expressions. We use a commercial software PhotoScan to reconstruct high fidelity facial geometry and texture.

3.2 Multiple Reflectance Channel Rendering

In Figure 3, we illustrates our multiple reflectance channel rendering pipeline. We randomly sample 100 subjects out of the 438 subjects for training and 8 subjects for testing. For the lighting conditions, we randomly choose 371 environment maps for training and 20 environment maps for testing. Thus there is no overlapping between the training data and testing ones in terms of subject identities and lighting conditions. Noted that we only use 100 subjects since we found using more subjects does not bring obvious performance improvements in training. In generating training and testing images, the rendering process follows the same pipeline. We register all the face scans with a face template using the method in [Cao et al. 2013] and transform all these scans into the normalized face pose. The face scans are then placed 3.5 meters away from the camera. To increase the robustness and generalization for training our networks, the face poses represented by the Euler angles are uniformly sampled at random in the space $(\theta_x, \theta_y, \theta_z) \in [-30°, 30°] \times [-15°, 15°] \times [-15°, 15°]$. To enrich the lighting variations, we augment an environment map ℓ by rotating the environment map along its longitude (horizontal direction) at random.

For photorealistic rendering, we use the Cycles rendering engine in Blender and a Principled BSDF shader to render the face images. The reconstructed texture maps T are employed as the albedo input of the shader. To achieve visually convincing rendering effects, we empirically set the specular and the roughness to 0.6 and 0.5 respectively in rendering an image I. In addition, we set the specular coefficient to 0 to further render an image without specular I_0. By setting the shadow visibility to false in Blender, we render an image without self-occlusions I_1. By computing the differences between I and I_0 as well as the differences between I and I_1, and then converting them to the grayscale, we achieve the specular map $I_{sp} = I - I_0$ and the shadow map $I_{sh} = I_1 - I$ in accordance with the image I. We also render the albedo image I_n under the uniform lighting environment. Although multiplicative shadow maps are often used in previous relighting methods [Nestmeyer et al. 2020], we opt to employ additive shadow maps as additive shadow will ease gradient backpropagation during network optimization. Because our proposed method focuses on relighting portraits, we set the background pixels to 0 during rendering.

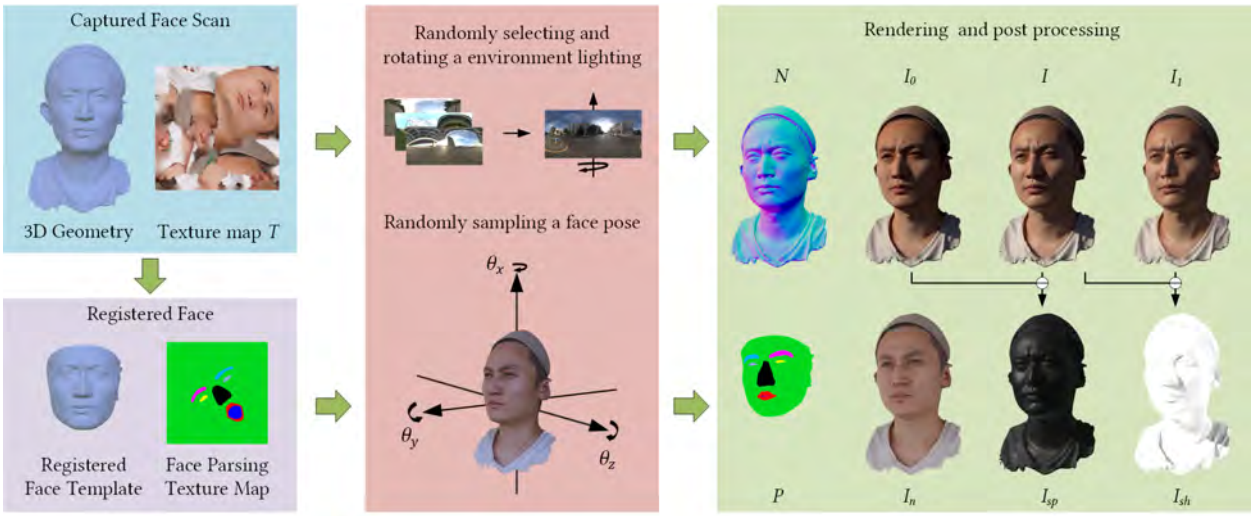


Fig. 3. Illustration of the rendering pipeline for our portrait relighting dataset generation. We first register a face template to a captured 3D scan and then align the scan to the canonical pose. An environment lighting configuration and a face pose are randomly chosen. Afterwards, we render an image I and its corresponding image without specular I_0, image without self-occlusion I_1, normal map N, albedo map I_n and parsing map P. The specular and shadow maps are obtained by computing $I - I_0$ and $I_1 - I$ in the grayscale respectively. For visualization purpose, in this paper, we apply a gamma correction to the specular image I_{sp} and an inverse operation to shadow image I_{sh}.

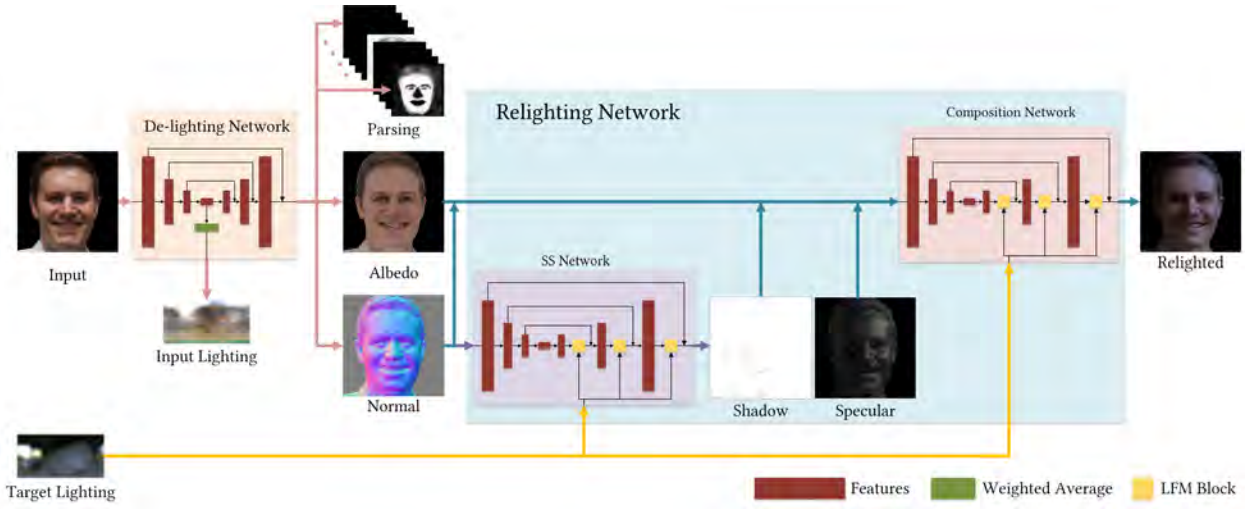


Fig. 4. Illustration of our network architecture. There are two stages in our method: the de-lighting stage and the relighting stage. The de-lighting stage aims to estimate the facial albedo, normal, and parsing maps, given an input portrait. In the relighting stage, we first employ a Specular and Shadow (SS) network to explicitly predict challenging lighting effects, *i.e.*, specular and shadow, and then we train a composition network to generate the relit results taking our estimated facial albedo, normal and lighting effects as inputs. Images courtesy: Flickr user *Peter Bright* (Input).

We represent the facial geometry using facial normal image N. To achieve the semantic face parsing maps, we manually divide the texture map of the registered face template into 8 different regions, *i.e.*, 2 eyebrows, 2 eyes, nose, inner mouth, lips and skin regions. By assigning different regions with different colors, we can easily obtain the corresponding face parsing map P as all the faces are pre-registered with the face template. An image group including $I, \ell, I_n, I_{sp}, I_{sh}, N$ and P is employed as our multi-channel images for training our networks. After removing the incomplete face scans and failure cases of rendering, we obtain 270,000 valid image groups for training and 20,000 valid image groups for testing. The resolution of the rendered images is 640×960 pixels. Since our method focuses on face regions, face bounding boxes are extracted according to the rendered parsing maps. In the training, we crop a square patch according to the bounding box with a padding value randomly sampled in the range of $[20, 40]$ and downsample it to 256×256 pixels. The rotated lighting environment map is downsampled to 16×32 using Gaussian pyramid. The environment maps are collected from

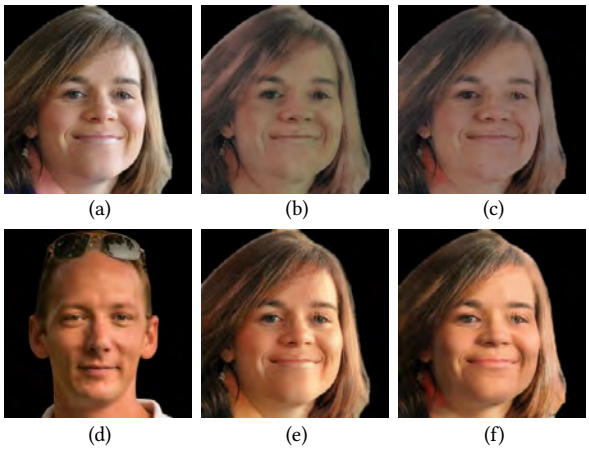


Fig. 5. Illustration of the effectiveness of the face parsing auxiliary task for training. We transfer the lighting from a reference image (d) to an input face (a) using the models trained with and without facial parsing. The model trained without parsing fails to estimate correct colors in the facial albedo image (b). Therefore, its relit image (e) still contains the specular of the input image. In contrast, the model trained with facial parsing generates a better albedo image (c) and a better relit result (f). Images courtesy: Flickr user *mcfincher* ((a)), Flickr user *Michael Erwine* ((d)).

HDRIHaven and CGTextures. The full dataset including original face scans and rendered final images will be released.

4 PROPOSED PORTRAIT RELIGHTING METHOD

We design our framework according to the following three considerations: *First*, facial albedo and normal are important factors for generating a portrait image. However, previous methods do not explicitly model them. As a consequence, we first use a de-lighting network to estimate the facial albedo and normal. Furthermore, we enhance the de-lighting network by introducing a facial parsing auxiliary task. *Second*, facial specular and shadow play an important role in generating photorealistic relit results. Therefore, we design a Specular and Shadow (SS) network to generate these effects explicitly. Our SS network estimates specular and shadow from the facial normal and the target lighting. *Third*, in order to combine the aforementioned information together, we train a composition network that takes facial albedo, normal and the estimated lighting effects as inputs to perform the final creation of the relit results.

4.1 De-lighting Stage

Given an input facial image I, our de-lighting network estimates its corresponding environment map $\hat{\ell}$, facial albedo $\hat{I_n}$ and facial normal $\hat{N}$. Here, the estimated facial normal $\hat{N}$ and albedo $\hat{I_n}$, as important facial intrinsic channels, are used for the subsequent relighting stage. To improve the performance of our de-lighting network, we introduce a facial parsing task as an auxiliary task. To be specific, this multi-task training strategy facilitates the estimation of albedo $\hat{I_n}$. Once $\hat{I_n}$ is accurately estimated, it will enable the lighting effect estimation and relit result composition in the following step, as shown in Figure 4.

Our de-lighting network is composed of an encoder-decoder architecture. We use a confidence-weighted average block [Sun et al. 2019] to estimate the lighting $\hat{\ell}$ at the bottleneck layer. The skip connection [Ronneberger et al. 2015] is employed to preserve the image details of the input. Sigmoid activation layers are used for normalizing facial albedo and normal while a Softmax layer is applied to estimate a face parsing map. In order to estimate the facial normal $\hat{N}$ and albedo $\hat{I_n}$ of an input portrait image, we employ L_1 regression losses as follows,

$$\mathcal{L}_{I_n} = \|\hat{I_n} - I_n\|_1, \tag{1}$$

$$\mathcal{L}_N = \|\hat{N} - N\|_1, \tag{2}$$

where I_n and N indicate the ground-truth albedo and normal. These two channels will be fed to our relighting stage to not only produce a relit image but also predict the specular and shadow under the target lighting.

As suggested in the work [Sun et al. 2019], estimating the source lighting would also improve the stability of the network training even though the source lighting will not be used in our relighting stage. Following the work [Weber et al. 2018], the environment lighting is estimated by a weighted-log-L_2 (wlog-L_2) distance between the ground-truth lighting ℓ and the estimated one $\hat{\ell}$, expressed as,

$$\mathcal{L}_\ell = \|\omega \odot (\log(1+\hat{\ell}) - \log(1+\ell))\|_2^2, \tag{3}$$

where ω is the solid angle of a pixel in the environment map, and $\odot$ represents the element-wise multiplication. Since an environment map is mapped to a sphere for rendering, different pixels correspond to different region sizes on a sphere. Thus, ω is used in Eqn. 3.

Due to the significantly different sizes of the semantic regions in the face parsing P maps, a multi-class focal loss [Lin et al. 2017] is employed to perform the parsing as,

$$\mathcal{L}_P = -(1-\hat{P})^\eta \odot P \odot \log(\hat{P}), \tag{4}$$

where $\hat{P}$ is the estimated face parsing map and η is a focusing parameter. The contribution of estimating this parsing map is shown in Figure 5. It does improve the results as the parsing lets the network understand the facial semantic information.

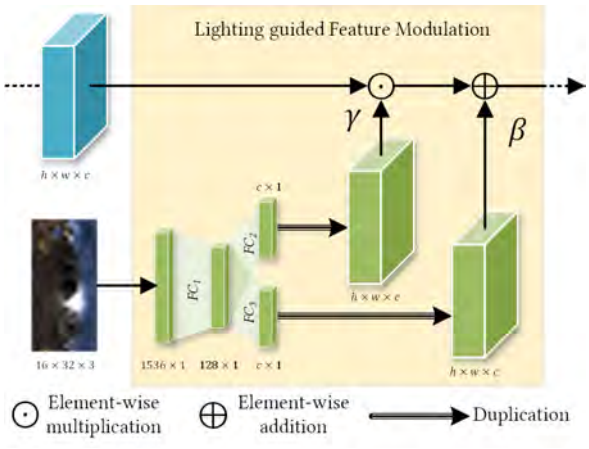
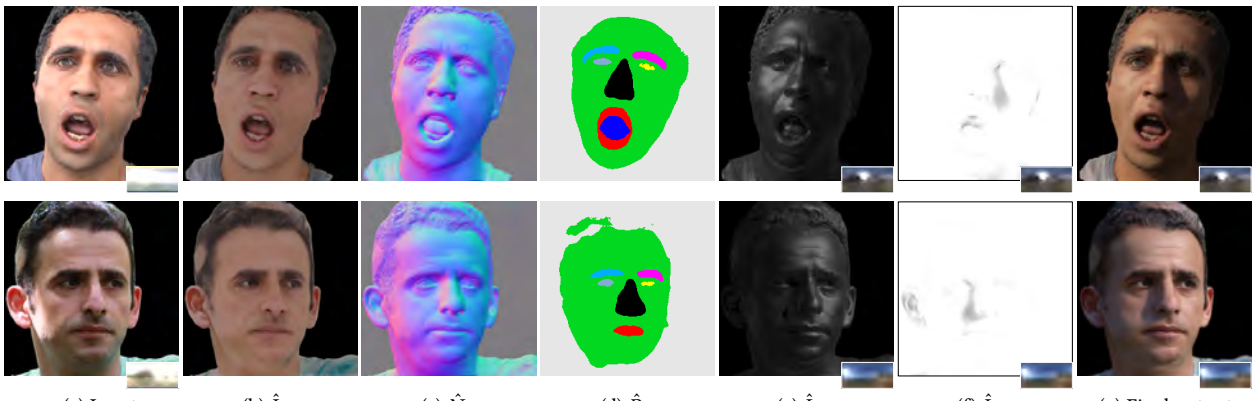


Fig. 6. Illustration of our Lighting guided Feature Modulation (LFM) module. LFM fuses the target lighting into channel features to deliver more convincing results.



(a) Input (b) $\hat{I}_n$ (c) $\hat{N}$ (d) $\hat{P}$ (e) $\hat{I}_{sp}$ (f) $\hat{I}_{sh}$ (g) Final output

Fig. 7. Illustration of the outputs of our network. The de-lighting network decomposes an input image (a) to the facial albedo (b), normal (c), parsing map (d) and environment map $\hat{\ell}$ (as seen in the bottom right of (a)). By providing an environment lighting ℓ^r (as shown in the right bottom of (e), (f) and (g)), our SS network generates the effects of specular (e) and shadow (f). Our final relit images are shown in (g). Images courtesy: Flickr user *Denis Kornetsov* ((a)-2).

Overall, we treat all the tasks equally and the total loss for training our de-lighting network is expressed as,

$$\mathcal{L}_{DL} = \mathcal{L}_{I_n} + \mathcal{L}_N + \mathcal{L}_\ell + \mathcal{L}_P. \tag{5}$$

As illustrated in Figure 7, we demonstrate the estimated results of our de-lighting network.

4.2 Relighting Stage

In our relighting stage, we first estimate the specular and shadow lighting effects with the estimated facial normal and the target lighting, and then compose the multi-channel images to achieve our relit results. Hence, there are two networks in our relighting stage: a Specular and Shadow (SS) network and a Composition network.

4.2.1 Specular and Shadow Effect Estimation. Specular and shadow are two challenging lighting effects and play an important role in photorealistic portrait relighting. Previous methods [Sun et al. 2019; Zhou et al. 2019] only constrain the relit results to be close to their ground-truths without estimating specular and shadow specifically. Compared to the diffuse component, specular and shadow components only appear in local and small regions. Therefore, without special treatment, these effects might be neglected due to the small portions in a regression loss. In contrast, we use a Specular and Shadow (SS) network to estimate them explicitly, as illustrated in Figure 4. In this manner, specular and shadow will be learned in our network.

Although facial albedo, normal and lighting all have impacts on the specular and shadow, the intensities of the specular and shadow are mainly controlled by the facial geometry and lighting. Therefore, the inputs of our SS network are the estimated facial normal and the targeting lighting. In learning specular I_{sp} and shadow I_{sh}, we force them to be similar to their ground-truths, expressed as,

$$\mathcal{L}_{SS} = \|\hat{I}_{sp} - I_{sp}\|_1 + \|\hat{I}_{sh} - I_{sh}\|_1, \tag{6}$$

where $\hat{I}_{sp}$ and $\hat{I}_{sh}$ indicate the estimated specular and shadow maps. Our SS network is also composed of an encoder-decoder structure. To reinforce the impacts of the input lighting in the SS network, we introduce a Lighting guided Feature Modulation (LFM) module. LFM will be explained in the following Section 4.2.2.

4.2.2 Lighting guided Feature Modulation. In prior works [Sun et al. 2019; Zhou et al. 2019], the target lighting is added to a network by concatenating it or its derived features with the features of an input image in the bottleneck layer. It is difficult to guarantee that the target lighting is effectively fused into the input image features. Inspired by spatial feature transform [Wang et al. 2018], adaptive instance normalization [Huang and Belongie 2017] and StyleGAN [Karras et al. 2019], we present a Lighting guided Feature Modulation (LFM) module to fuse the lighting with the features of our channel images in a layer-wise fashion. Figure 6 illustrates the design of our LFM. To be specific, we modulate the input features F_{in} using the lighting map ℓ. The lighting ℓ is first converted to a 128-dimensional feature by a fully-connected layer FC_1. Subsequently, this feature is converted to two c-dimensional features by two fully-connected layers FC_2 and FC_3: a multiplicative factor γ and an additive factor β. Then, these two factors are expanded to the $h \times w \times c$ to match the shape of F_{in}. The output features F_{out} of LFM are formulated as,

$$F_{out} = \gamma \odot F_{in} + \beta. \tag{7}$$

Note that, only the first fully-connected layer FC_1 is followed by an non-linear activation function ReLU.

4.2.3 Multi-Channel Composition Network. After obtaining our multiple reflectance channel maps (facial albedo, normal, specular and shadow), we design a network to compose them consistently in accordance with the target lighting. Specifically, we adopt an encoder-decoder architecture as well as embed our LFM for the multi-channel composition network, as illustrated in Figure 4. Here, LFM is employed after each upsampling layer to modulate the features in accordance with the target lighting. In order to train our multi-channel composition network, we take the target lighting and our estimated multi-channel images as inputs and constrain the reconstructed image $\hat{I}$ to be similar to the ground-truth I, written as,

$$\mathcal{L}_{RL} = \|\hat{I} - I\|_1. \tag{8}$$

4.3 Training Details

In order to train our de-lighting and relighting networks, we employ a two-stage training fashion. In the first stage, all the networks are trained from scratch separately, regarded as a "warm-up" phase. In training our de-lighting network, the ground-truth facial albedo, normal and parsing map are provided from one image group $\mathcal{G} = \{I, \ell, I_n, I_{sp}, I_{sh}, N, P\}$ and the objective function in Eqn. (5) is employed. In the relighting networks, we also train the SS and composition networks separately. In this way, we only need to use one image group during training. Note that, the inputs of the SS and composition networks are the ground-truth multi-channel images in the warm-up stage.

Recall that each image group is captured under one lighting. In order to fine-tune our entire network, we sample another image group capturing the same subject in the same pose but a different lighting, marked as $\mathcal{G}^r = \{I^r, \ell^r, I^r_n, I^r_{sp}, I^r_{sh}, N^r, P^r\}, I^r_n = I_n, N^r = N$ and $P^r = P$, for the relighting purpose.

Different from the "warm-up" stage, the inputs of the SS and composition networks are the predictions of our networks. Therefore, the supervision signals in Eqn. (6) and Eqn. (8) in the relighting fine-tuning stage are I^r_{sp}, I^r_{sh} and I^r, respectively. Particularly, we first fine-tune our de-lighting network by fixing the SS and multi-channel composition networks. Then, we start fine-tuning the relighting network when the loss of the de-lighting network converges. Our model is implemented in Pytorch [Paszke et al. 2019]. We set the focusing parameter η to 2 following the focal loss [Lin et al. 2017]. Adam optimizer [Kingma and Ba 2014] with its default setting is employed to train our model. The learning rate is set to 0.002 in the warm-up stage and decreased to 0.0002 in the relighting fine-tuning stage. The batch size is set to 8. In the warm-up stage, each network is trained on a NVIDIA GTX 1080Ti GPU individually for one day. During fine-tuning, we use four NVIDIA GTX 1080Ti GPUs to train the entire network for 2 days. The average running time is about 46ms for relighting an image during testing, including 11ms for the de-lighting network, 17ms for the SS network and 18ms for the composition network.

5 EXPERIMENTS

In this section, we first evaluate the key contributions of our proposed method. Then we compare our method with state-of-the-art methods quantitatively and qualitatively to show that our technique better handles challenging lightings either for the input image or for the target. Next, we present various relit results to show that our technique is robust to various cases in real portraits, including various inputs and target lightings, head poses and facial expressions, objects like hats and glasses, and so on. Finally, as we model multiple reflectance channels with semantics to perform relighting, we are able to manipulate some channels to edit the relit results. We also show results on this kind. For quantitative comparisons, we adopt RMSE, PSNR and SSIM metrics to compare the relit images and their corresponding ground-truths by our testing dataset. For qualitative comparisons, we show directly the relit results of both real images and synthetic images.

5.1 Ablation Study

To demonstrate the effectiveness of our SS network, LFM and the auxiliary task of face parsing, we conduct comprehensive ablation studies on our method quantitatively and qualitatively.

Table 2 indicates that each component of our method contributes to the final relit results measured by the metrics of RMSE, PSNR and SSIM. Figure 8 also demonstrates how each component improves relit images in terms of visual quality. The baseline is a de-lighting and relighting (DL-RL) architecture without using our SS network, LFM and face parsing.

5.1.1 SS Network. After inserting the SS network (DL-RL+SS), the RMSE is reduced by 18% compared to the baseline. The visual results start to exhibit similar lighting distributions to the ground-truths as shown in Figure 8(b) and (c). Note that, for DL-RL and DL-RL+SS, the target lighting is concatenated at the bottleneck layers of composition network and the SS network.

5.1.2 LFM. By fusing lighting using our LFM (DL-RL+SS+LFM), more precise lighting effects are generated as indicated by all the metrics in Table 2. Also in Figure 8(d), the relit portraits begin to show more lighting details on the face by employing LFM. Table 2 also indicates that LMF improves the estimation accuracy of the specular and shadow maps. Figure 9 further presents some results of the SS network with and without LFM. Here, to eliminate errors caused by other modules, we feed a ground-truth normal map and a target lighting to the SS network. Figure 9 shows that an SS network without LMF only captures the low-frequency parts of the specular and the estimated shadow is much weaker than the ground-truth. LMF helps the SS network to estimate the high-frequency details of specular and to generate more accurate shadow effects.

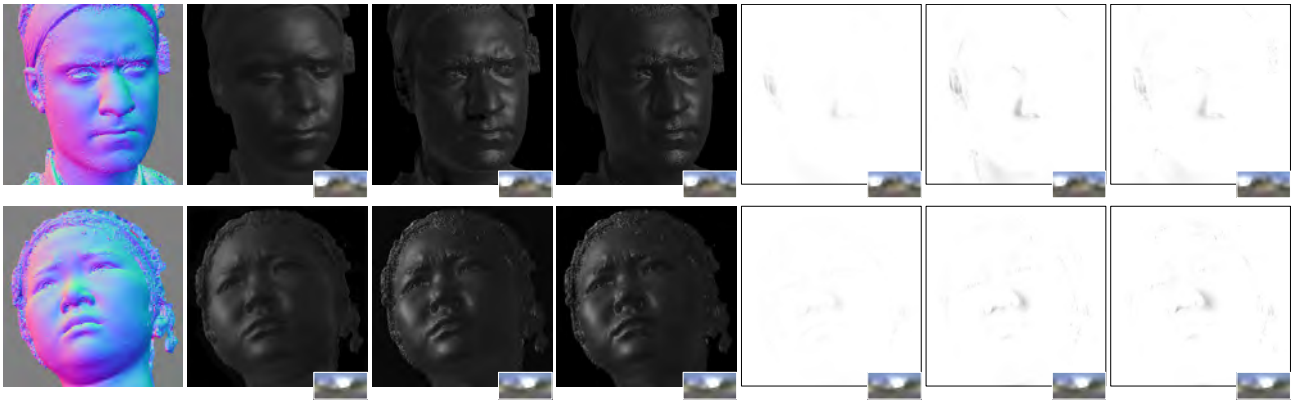
5.1.3 Parsing. As indicated in Table 2, with the help of the face parsing task, we improve the estimation of facial albedo images and thus achieve more accurate relit results (denoted by Full). Figure 8(e) also shows that the highlights and shadow on the forehead and eye regions are slightly improved by using the parsing. The reason might be that these face regions have their common lighting effects in some extreme lighting conditions which are learned by our network with the help of the parsing. Note that, Figure 5 has provided more specific evaluations on the parsing for real-world input images.

5.2 Comparisons with the State-of-the-Art

We compare our method with the state-of-the-art portrait relighting methods [Sun et al. 2019; Zhou et al. 2019] and portrait color transfer methods [Shih et al. 2014; Shu et al. 2017a] on both synthesized and real-world images. In order to compare with DPR [Zhou et al. 2019], we use the codes provided by the authors. DPR uses an SH lighting model with a white light assumption. Therefore, we convert the environment lighting to grayscale and compute the SH coefficients. To compare our method with the approach [Sun et al. 2019] quantitatively, we re-implement Sun et al.'s method and train their network on our synthesized images. Then, we evaluate their network on our synthesized data for fair comparisons. For qualitative comparisons on real images, the results are obtained from Sun et al.'s model reported in their original paper. Since color transfer methods require a reference image, we use the rendered

(a) Input (b) DL-RL (c) DL-RL+SS (d) DL-RL+SS+LFM (e) Full (f) Ground truth

Fig. 8. Qualitative evaluations on each component of our network. DL-RL indicates the baseline using our de-lighting (without face parsing) and relighting (without LFM modules) networks (b). DL-RL+SS represents the incorporation of the SS network for lighting effect estimation (c). DL-RL+SS+LFM indicates the employment of LFM in the SS and composition networks without using the face parsing task (d). Full refers to our entire network (e).



(a) Input Normal (b) W/O LFM (c) W/ LFM (d) Ground truth (e) W/O LFM (f) W/ LFM (g) Ground truth

Fig. 9. Lighting effect estimation of our SS network with and without LFM. LFM facilitates the SS network to generate specular and shadow components.

Table 2. Quantitative evaluations on each component of our portrait relighting network. DR-RL refers to the baseline de-lighting and relighting architecture. The results are evaluated on our testing dataset with 20k images.

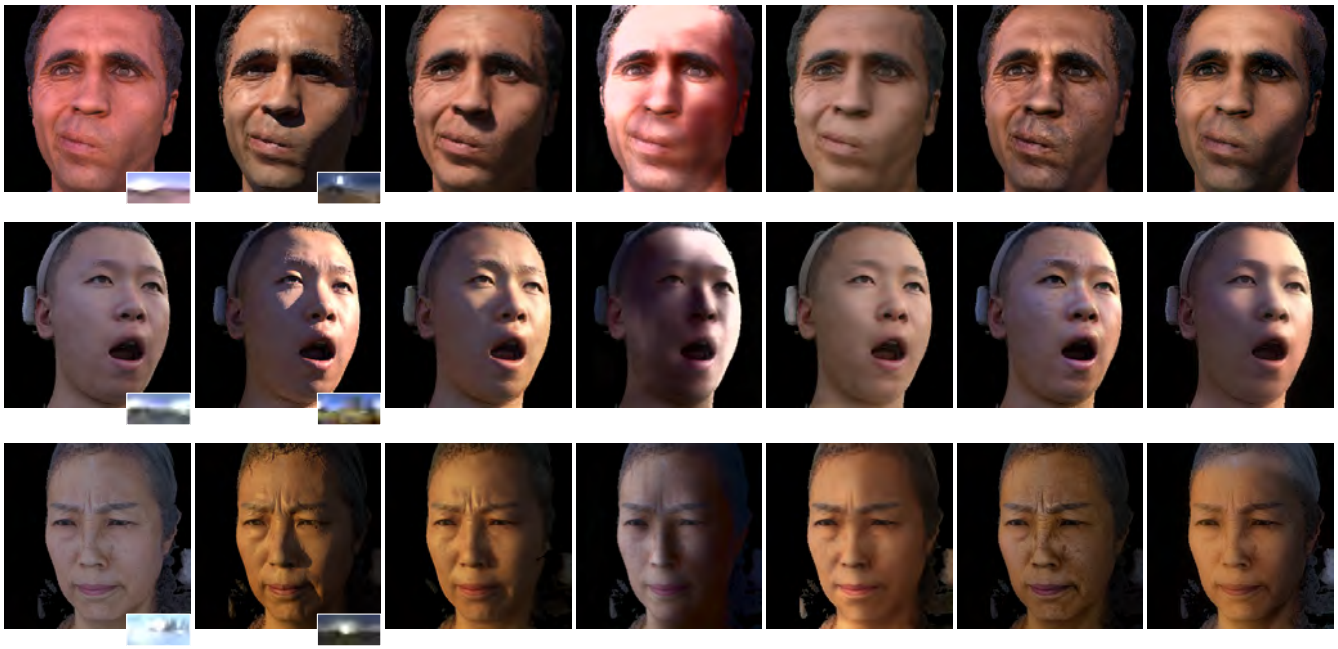
Algorithm	Relighting			Albedo	Normal	Specular	Shadow	Lighting
	RMSE	PSNR	SSIM	RMSE	RMSE	RMSE	RMSE	wlog-L_2
DL-RL	0.0511	26.4	0.870	0.0342	0.132	-	-	0.0364
DL-RL+SS	0.0415	28.2	0.882	0.0347	0.144	0.00909	0.00512	0.0356
DL-RL+SS+LFM	0.0309	31.0	0.909	0.0314	**0.131**	**0.00575**	**0.00416**	0.0301
Full	**0.0281**	**31.8**	**0.914**	**0.0271**	0.132	**0.00575**	**0.00416**	**0.0274**

ground-truth faces (*i.e.*, the same subjects under the target lighting) as the reference images for those methods [Shih et al. 2014; Shu et al. 2017a].

5.2.1 On Synthetic Data. We first compare our method against prior works on our testing dataset where we have the ground-truth images for comparisons. The inputs are a synthesized face image and an environment map to describe the target lighting.

Table 3 manifests that our method greatly outperforms the existing methods in all the metrics. Note that, since the ground-truth images are used as the input references of the color transfer methods, these methods should achieve their best performance in this case. In other words, their quantitative results in Table 3 should be higher than their real performance when the reference image is from another person's portrait. On the other hand, even though in this situation, our method still outperforms these methods.

Qualitative comparisons are shown in Figure 10. As seen in Figure 10, only our method generates realistic specular and shadow (in the eye socket and around the nose) compared with the ground-truth.



(a) Input (b) Ground truth (c) Our method (d) [Zhou et al. 2019] (e) [Sun et al. 2019] (Our implementation) (f) [Shih et al. 2014] (g) [Shu et al. 2017a]

Fig. 10. Comparisons of relit results on synthetic images. We also rendered the ground truth images under the target lighting and provided them as references for [Shih et al. 2014; Shu et al. 2017a]. Our method outperforms prior works by generating visually convincing specular and shadow. Noted that, we use our implementation of [Sun et al. 2019] trained on our dataset in this experiment.

Table 3. Quantitative comparisons of our method against prior works on our relighting testing dataset with 20k images.

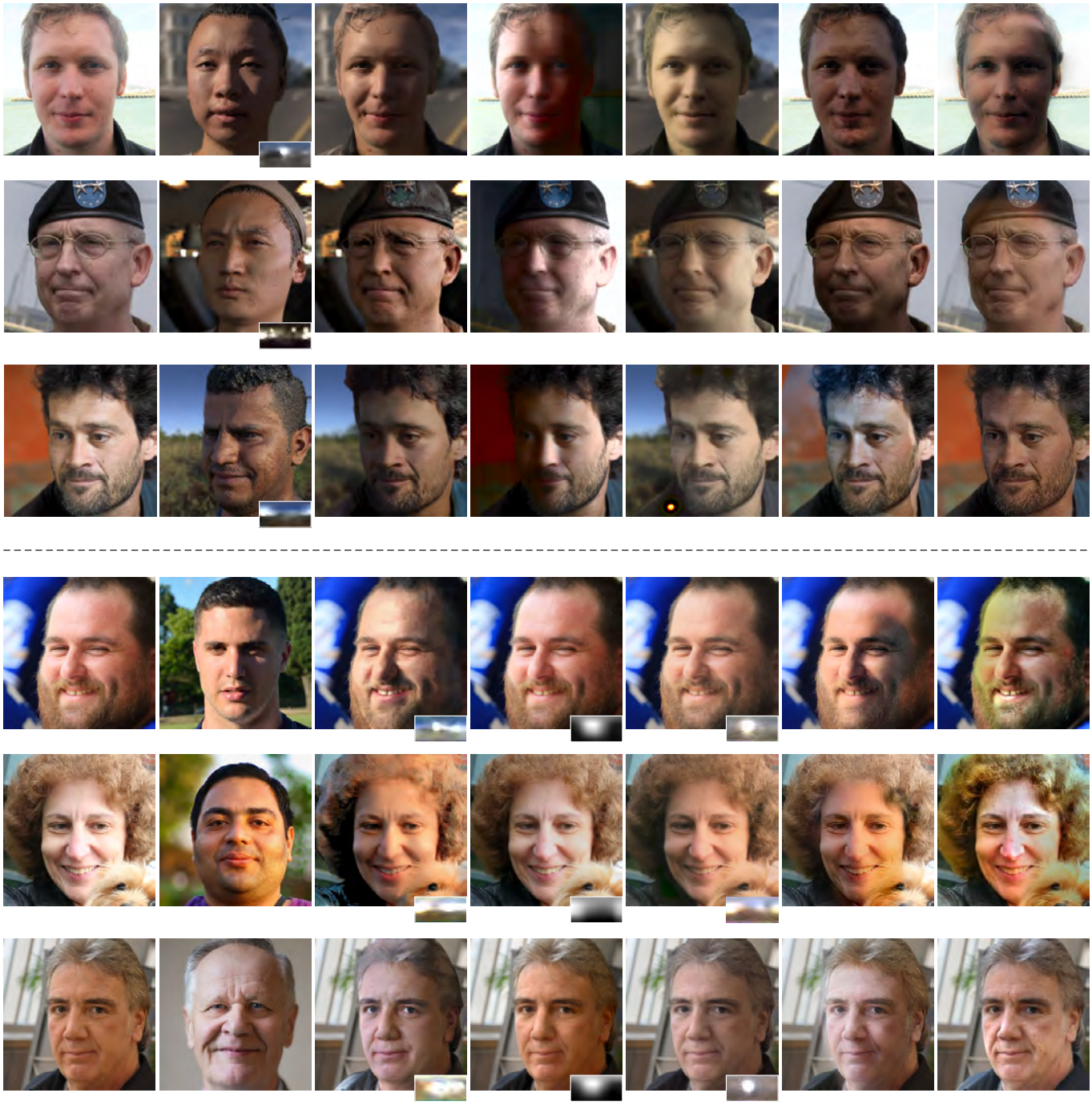
Algorithm	RMSE	PSNR	SSIM
[Shih et al. 2014]	0.0389	29.4	0.877
[Shu et al. 2017a]	0.0717	25.1	0.884
[Zhou et al. 2019]	0.1128	19.8	0.675
[Sun et al. 2019] (Our implementation)	0.0518	26.4	0.865
Our Model	**0.0281**	**31.8**	**0.914**

Zhou et al. [2019] fail to handle color changes of target lightings. Although Sun et al. [2019]'s network is trained on our synthetic images, it generates some smooth and unrealistic shadows in this case. Shu et al. [2017a]'s results look pleasing by themselves but are not consistent with the ground truth. The results of [Shih et al. 2014] are overly smooth with less facial details. Again, Shih et al. [2014]; Shu et al. [2017a] use the ground truth as their input references. Noted that, the results of [Sun et al. 2019] in Figure 11 and Table 3 are achieved using our implementation trained on our synthetic images.

5.2.2 On Real Data. Although our model is trained on synthetic images, it can be applied to real-world images and achieves photo-realistic portrait relighting results. To improve the visual effects, we follow the method [Sun et al. 2019] to change the background of an input image to the corresponding part of the environment lighting map in the results. In Figure 11, the experiments are conducted on real-world images from FFHQ [Karras et al. 2019] and the results of [Sun et al. 2019] are obtained from the model reported in their original paper.

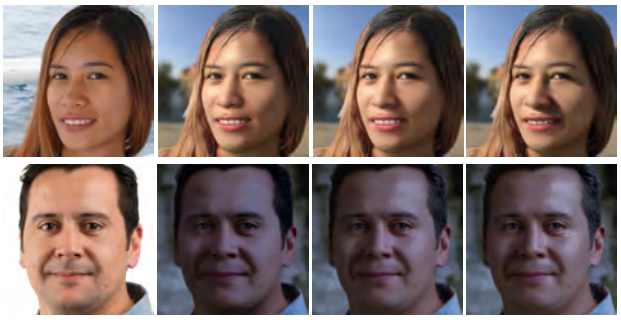
In the top part of Figure 11, we use an environment lighting map as the target lighting. An image rendered under this target lighting is also presented as the reference but not used in the experiments. Here, our method generates naturally-looking specular and shadow under extreme lighting conditions and lighting effects of the relit portraits are similar to the reference ones, still with some differences due to different face shapes. In particular, our method achieves photorealistic facial images under strong side lights. Both the generated specular and shadow are consistent with the reference, which are not achieved by the compared methods. This mainly benefits from the manner that we render the lighting effects explicitly according to the facial geometry and target lighting.

In the bottom part of Figure 11, we use a real-world portrait image as reference to provide the target lighting information. In this case, the lighting of the reference is unknown, so we apply our de-lighting network to estimate it. The state-of-the-art face relighting methods [Sun et al. 2019; Zhou et al. 2019] also have their methods to estimate the lighting from the reference images. The results again demonstrate that our method achieves the best visual quality in comparison to other competing methods. Here, besides the specular and shadow, we can further see that our method handles hair very well, which has correct color changes corresponding to the target lighting. These experiments also imply that our method is able to estimate the environment lighting accurately.



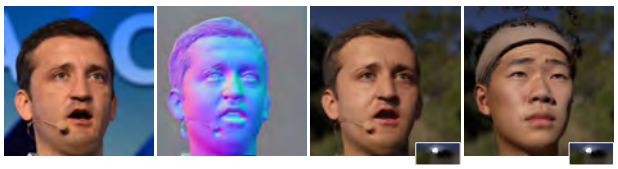
(a) Input (b) Reference (c) Our method (d) [Zhou et al. 2019] (e) [Sun et al. 2019] (f) [Shih et al. 2014] (g) [Shu et al. 2017a]

Fig. 11. Comparisons of relit results on real images. In the top part, the competing works [Sun et al. 2019; Zhou et al. 2019] and our method use target lightings to relight input portraits. The reference images are rendered under the target lightings serving as inputs of [Shih et al. 2014; Shu et al. 2017a]. In the bottom part, the works [Sun et al. 2019; Zhou et al. 2019] and our method need to estimate the target lighting from the real-world reference images. Then, the estimated lighting conditions are used to relight the input images. Images courtesy: Flickr user *Jennifer Morrow* ((a)-1), Flickr user *2nd Armored Brigade Combat Team Public Affairs* ((a)-2), Flickr user *Luca Boldrini* ((a)-3), Flickr user *alias keiner* ((a)-4), Flickr user *Edy Rung* ((b)-4), Flickr user *Siena College* ((a)-5), Flickr user *Ramakrishna Reddy Y* ((b)-5), Flickr user *Charles Vasser* ((a)-6), Flickr user *Blake Patterson* ((b)-6).



(a) (b) (c) (d)

Fig. 12. Shadow and specular components manipulation in the relit images. When we relight an input image (a), we multiply the estimated shadow (top) or specular (bottom) component of the SS network by 0, 1 and 2 and thus achieve different relighting effects (b), (c) and (d) respectively. Images courtesy: Flickr user *Harry Fozzard* ((a)-1), Flickr user *Mark Baylor* ((a)-2).



Input Normal Relit Reference

Fig. 13. Illustration of the limitations of our method. The shadow of the microphone leads to inaccurate estimation of a normal map. Thus, a strong shadow appears on the relit image. Images courtesy: Flickr user *Web Summit* (Input).

5.3 More Results and Applications

We show more results of our method and more comparisons in the accompanying video and document, where we can see that our method generates consistent videos with continuous lighting changes and is robust to various input portraits and target lightings. With the help of our SS network, our method can control the specular and shadow in the relit images. By manipulating the scale of the generated specular and shadow components before multi-channel composition, our method enables users to adjust the relighting effects as shown in Figure 12.

5.4 Limitations

Our method in general achieves promising performance on relighting portraits along with some other objects, *e.g.*, glasses or hats, as shown in Figure 1. Since it is difficult to obtain diffuse, specular of face models in our data capture system, we treat spatially variant diffuse and specular coefficients as constants and omit subsurface scattering in rendering. Hair, eyes and clothes are not specifically modeled. Thus, the rendered relighting effects are not physically accurate. However, we believe our method can achieve more visually pleasant results if those factors are taken into account in rendering our dataset. In our future work, we intend to leverage the techniques [Ghosh et al. 2011; Weyrich et al. 2006; Yamaguchi et al. 2018] to better model facial reflectance and thus improve our relighting effects.

When there are isolated objects in front of a face and cast shadow on a face, our method may generate artifacts. This is because our method only accounts for the shadow caused by self-occlusions while shadow caused by other objects is often unpredictable without fully understanding the positions and sizes of objects. Thus, that shadow will cause the inaccurate estimation of facial normal and albedo. As shown in Figure 13, the shadow of the microphone leads to incorrect estimation of facial normal in the corresponding region. In the relit image, although we use a frontal lighting, the position of the shadow has not been changed. A possible solution to handle this is to add specific samples of this kind in the training dataset following [Zhang et al. 2020a]. In this manner, the network is able to learn to handle this case. Solving these problems and combining our method with single portrait based animation methods (*e.g.*, [Averbuch-Elor et al. 2017]) will be left to our future work.

6 CONCLUSION

In this paper, we proposed a single image portrait relighting framework by explicit modeling multiple reflectance channels which embed facial albedo, geometry and the lighting effects of specular and shadow. Using our de-lighting network, we recovered the facial albedo and geometry from an input portrait. We developed an SS network to explicitly estimate the lighting effects of shadow and specular in accordance with the estimated facial geometry and the target lighting map. Our Lighting guided Feature Modulation (LFM) module blends the features of the multiple reflectance channels more consistently, thus achieving photo-realistic relit results while tackling challenging lighting effects. We presented a large-scale dataset with the ground-truth multi-channel supervision. Our dataset provides high-quality 3D faces and various useful channels, enabling us to learn the explicit modeling of the multiple reflectance channels. Extensive results demonstrate that our proposed method achieves photo-realistic relit results with challenging relighting effects, and our recovered channels can be used to manipulate the relit results with desired amount of specular and shadow.

ACKNOWLEDGMENTS

This work was supported by the National Key R&D Program of China 2018YFA0704000, the NSFC (No.61822111, 61727808, 61671268) and Beijing Natural Science Foundation (JQ19015, L182052). Feng Xu is the corresponding author. The 3rd author (Ming Lu) waives his intellectual property rights of this work. We thank Tiancheng Sun for providing the relit results of the real images.

REFERENCES

Oswald Aldrian and William AP Smith. 2012. Inverse rendering of faces with a 3D morphable model. *IEEE transactions on pattern analysis and machine intelligence* 35, 5 (2012), 1080–1093.

Hadar Averbuch-Elor, Daniel Cohen-Or, Johannes Kopf, and Michael F. Cohen. 2017. Bringing Portraits to Life. *ACM Transactions on Graphics (Proceeding of SIGGRAPH Asia 2017)* 36, 6 (2017), 196.

Yin Baocai, Sun Yanfeng, Wang Chengzhang, and Ge Yun. 2009. BJUT-3D large scale 3D face database and information processing. *Journal of Computer Research and Development* 6, 020 (2009), 4.

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.

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.

James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniah, and David Dunaway. 2016. A 3d morphable model learnt from 10,000 faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5543–5552.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2013. Facewarehouse: A 3d facial expression database for visual computing. *IEEE Transactions on Visualization and Computer Graphics* 20, 3 (2013), 413–425.

Xiaowu Chen, Mengmeng Chen, Xin Jin, and Qinping Zhao. 2011. Face illumination transfer through edge-preserving filters. In *CVPR 2011*. IEEE, 281–287.

Xiaowu Chen, Hongyu Wu, Xin Jin, and Qinping Zhao. 2013. Face illumination manipulation using a single reference image by adaptive layer decomposition. *IEEE Transactions on Image Processing* 22, 11 (2013), 4249–4259.

Shiyang Cheng, Irene Kotsia, Maja Pantic, and Stefanos Zafeiriou. 2018. 4DFAB: A large scale 4d database for facial expression analysis and biometric applications. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5117–5126.

Darren Cosker, Eva Krumhuber, and Adrian Hilton. 2011. A FACS valid 3D dynamic action unit database with applications to 3D dynamic morphable facial modeling. In *2011 International Conference on Computer Vision*. IEEE, 2296–2303.

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*. 145–156.

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.

Wen Gao, Bo Cao, Shiguang Shan, Xilin Chen, Delong Zhou, Xiaohua Zhang, and Debin Zhao. 2007. The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 38, 1 (2007), 149–161.

Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. 2011. Multiview face capture using polarized spherical gradient illumination. In *Proceedings of the 2011 SIGGRAPH Asia Conference*. 1–10.

Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. 2010. Multi-pie. *Image and Vision Computing* 28, 5 (2010), 807–813.

Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*. 1501–1510.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4401–4410.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Kuang-Chih Lee, Jeffrey Ho, and David J Kriegman. 2005. Acquiring linear subspaces for face recognition under variable lighting. *IEEE Transactions on pattern analysis and machine intelligence* 27, 5 (2005), 684–698.

Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. 2017. Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics (ToG)* 36, 6 (2017), 194.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*. 2980–2988.

Ce Liu, Jenny Yuen, and Antonio Torralba. 2010. Sift flow: Dense correspondence across scenes and its applications. *IEEE transactions on pattern analysis and machine intelligence* 33, 5 (2010), 978–994.

Jitendra Malik and Pietro Perona. 1990. Preattentive texture discrimination with early vision mechanisms. *JOSA A* 7, 5 (1990), 923–932.

Koki Nagano, Huiwen Luo, Zejian Wang, Jaewoo Seo, Jun Xing, Liwen Hu, Lingyu Wei, and Hao Li. 2019. Deep face normalization. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–16.

Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas M Lehrmann. 2020. Learning Physics-guided Face Relighting under Directional Light. In *Conference on Computer Vision and Pattern Recognition*. IEEE/CVF.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 8024–8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Francois Pitie, Anil C Kokaram, and Rozenn Dahyot. 2005. N-dimensional probability density function transfer and its application to color transfer. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, Vol. 2. IEEE, 1434–1439.

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

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*. Springer, 234–241.

Arman Savran, Neşe Alyüz, Hamdi Dibeklioğlu, Oya Çeliktutan, Berk Gökberk, Bülent Sankur, and Lale Akarun. 2008. Bosphorus database for 3D face analysis. In *European Workshop on Biometrics and Identity Management*. Springer, 47–56.

Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. 2018. SfSNet: Learning Shape, Reflectance and Illuminance of Facesin the Wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6296–6305.

YiChang Shih, Sylvain Paris, Connelly Barnes, William T Freeman, and Frédo Durand. 2014. Style transfer for headshot portraits. *ACM Transactions on Graphics (TOG)* 33, 4 (2014), 148.

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.

Yibing Song, Linchao Bao, Shengfeng He, Qingxiong Yang, and Ming-Hsuan Yang. 2017. Stylizing face images via multiple exemplars. *Computer Vision and Image Understanding* 162 (2017), 135–145.

Giota Stratou, Abhijeet Ghosh, Paul Debevec, and Louis-Philippe Morency. 2011. Effect of illumination on automatic expression recognition: a novel 3D relightable facial database. In *Face and Gesture 2011*. IEEE, 611–618.

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 (Proceedings SIGGRAPH)* (2019).

Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2018. Recovering realistic texture in image super-resolution by deep spatial feature transform. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 606–615.

Yang Wang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang, and Dimitris Samaras. 2007. Face re-lighting from a single image under harsh lighting conditions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1–8.

Yang Wang, Lei Zhang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang, and Dimitris Samaras. 2008. Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, 11 (2008), 1968–1984.

Henrique Weber, Donald Prévost, and Jean-François Lalonde. 2018. Learning to estimate indoor lighting from 3d objects. In *2018 International Conference on 3D Vision (3DV)*. IEEE, 199–207.

Tim Weyrich, Wojciech Matusik, Hanspeter Pfister, Bernd Bickel, Craig Donner, Chien Tu, Janet McAndless, Jinho Lee, Addy Ngan, Henrik Wann Jensen, et al. 2006. Analysis of human faces using a measurement-based skin reflectance model. *ACM Transactions on Graphics (TOG)* 25, 3 (2006), 1013–1024.

Shugo 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 Transactions on Graphics (TOG)* 37, 4 (2018), 1–14.

Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. 2020. FaceScape: a Large-scale High Quality 3D Face Dataset and Detailed Riggable 3D Face Prediction. *arXiv preprint arXiv:2003.13989* (2020).

Lijun Yin, Xiaozhou Wei, Yi Sun, Jun Wang, and Matthew J Rosato. 2006. A 3D facial expression database for facial behavior research. In *7th international conference on automatic face and gesture recognition (FGR06)*. IEEE, 211–216.

Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, and Peng Liu. 2013. A high-resolution spontaneous 3d dynamic facial expression database. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–6.

Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. 2014. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing* 32, 10 (2014), 692–706.

Xuaner (Cecilia) 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, Article 78 (July 2020), 14 pages.

Yang Zhang, Ivor W Tsang, Yawei Luo, Chang-Hui Hu, Xiaobo Lu, and Xin Yu. 2020b. Copy and Paste GAN: Face Hallucination from Shaded Thumbnails. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7355–7364.

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.