Faces as Lighting Probes via Unsupervised Deep Highlight Extraction

Renjiao Yi[1,2], Chenyang Zhu[1,2], Ping Tan[1], Stephen Lin[3]

[1] Simon Fraser University, Burnaby, Canada
{renjiaoy, cza68, pingtan}@sfu.ca
[2] National University of Defense Technology, Changsha, China
[3] Microsoft Research, Beijing, China
stevelin@microsoft.com

Abstract. We present a method for estimating detailed scene illumination using human faces in a single image. In contrast to previous works that estimate lighting in terms of low-order basis functions or distant point lights, our technique estimates illumination at a higher precision in the form of a non-parametric environment map. Based on the observation that faces can exhibit strong highlight reflections from a broad range of lighting directions, we propose a deep neural network for extracting highlights from faces, and then trace these reflections back to the scene to acquire the environment map. Since real training data for highlight extraction is very limited, we introduce an unsupervised scheme for finetuning the network on real images, based on the consistent diffuse chromaticity of a given face seen in multiple real images. In tracing the estimated highlights to the environment, we reduce the blurring effect of skin reflectance on reflected light through a deconvolution determined by prior knowledge on face material properties. Comparisons to previous techniques for highlight extraction and illumination estimation show the state-of-the-art performance of this approach on a variety of indoor and outdoor scenes.

Keywords: Illumination estimation · unsupervised learning

1 Introduction

Spicing up selfies by inserting virtual hats, sunglasses or toys has become easy to do with mobile augmented reality (AR) apps like *Snapchat* [43]. But while the entertainment value of mobile AR is evident, it is just as clear to see that the generated results are usually far from realistic. A major reason is that virtual objects are typically not rendered under the same illumination conditions as in the imaged scene, which leads to inconsistency in appearance between the object and its background. For high photorealism in AR, it is thus necessary to estimate the illumination in the image, and then use this estimate to render the inserted object compatibly with its surroundings.

Illumination estimation from a single image is a challenging problem because lighting is intertwined with geometry and reflectance in the appearance of a

scene. To make this problem more manageable, most methods assume the geometry and/or reflectance to be known [18,19,27,30,32,36,37,48]. Such knowledge is generally unavailable in practice; however, there exist priors about the geometry and reflectance properties of human faces that have been exploited for illumination estimation [10,12,17,34]. Faces are a common occurrence in photographs and are the focus of many mobile AR applications. The previous works on face-based illumination estimation consider reflections to be diffuse and estimate only the low-frequency component of the environment lighting, as diffuse reflectance acts as a low-pass filter on the reflected illumination [32]. However, a low-frequency lighting estimate often does not provide the level of detail needed to accurately depict virtual objects, especially those with shiny surfaces.

In addressing this problem, we consider the parallels between human faces and mirrored spheres, which are conventionally used as lighting probes for acquiring ground truth illumination. What makes a mirrored sphere ideal for illumination recovery is its perfectly sharp specular reflections over a full range of known surface normals. Rays can be traced from the camera's sensor to the sphere and then to the surrounding environment to obtain a complete environment map that includes lighting from all directions and over all frequencies, subject to camera resolution. We observe that faces share these favorable properties to a large degree. They produce fairly sharp specular reflections (highlights) over its surface because of the oil content in skin. Moreover, faces cover a broad range of surface normals, and there exist various methods for recovering face geometry from a single image [2,10,34,38,51]. Unlike mirrored spheres, the specular reflections of faces are not perfectly sharp and are mixed with diffuse reflection. In this paper, we propose a method for dealing with these differences to facilitate the use of faces as lighting probes.

We first present a deep neural network for separating specular highlights from diffuse reflections in face images. The main challenge in this task is the lack of ground truth separation data on real face images for use in network training. Although ground truth separations can be generated synthetically using graphics models [41], it has become known that the mismatch between real and synthetic data can lead to significant reductions in performance [42]. We deal with this issue by pretraining our network with synthetic images and then finetuning the network using an unsupervised strategy with real photos. Since there is little real image data on ground truth separations, we instead take advantage of the property that the diffuse chromaticity values over a given person's face are relatively unchanged from image to image, aside from a global color rescaling due to different illumination colors and sensor attributes. From this property, we show that the diffuse chromaticity of multiple aligned images of the same face should form a low-rank matrix. We utilize this low-rank feature in place of ground truth separations to finetune the network using multiple real images of the same face, downloaded from the MS-celeb-1M database [7]. This unsupervised finetuning is shown to significantly improve highlight separation over the use of supervised learning on synthetic images alone.

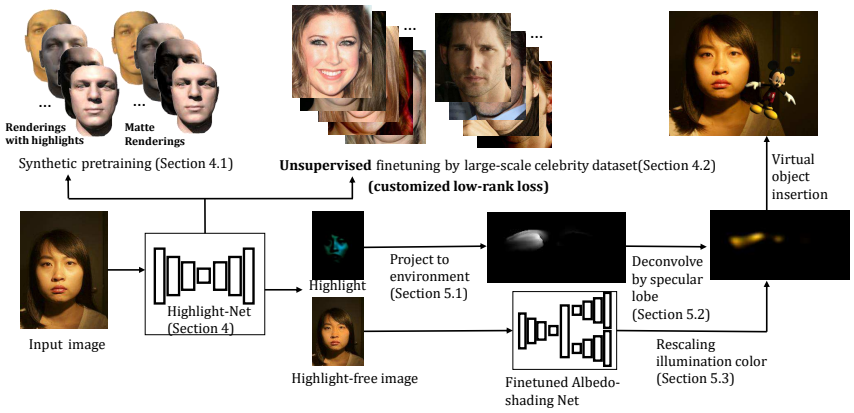


Fig. 1. Overview of our method. An input image is first separated into its highlight and diffuse layers. We trace the highlight reflections back to the scene according to facial geometry to recover a non-parametric environment map. A diffuse layer obtained through intrinsic component separation [24] is used to determine illumination color. With the estimated environment map, virtual objects can be inserted into the input image with consistent lighting.

With the extracted specular highlights, we then recover the environment illumination. This recovery is inspired by the frequency domain analysis of reflectance in [32], which concludes that reflected light is a convolved version of the environment map. Thus, we estimate illumination through a deconvolution of the specular reflection, in which the deconvolution kernel is determined from prior knowledge of face material properties. This approach enables recovery of higher-frequency details in the environment lighting.

This method is validated through experimental comparisons to previous techniques for highlight extraction and illumination estimation. On highlight extraction, our method is shown to produce results that more closely match the ground truth acquired by cross-polarization. For illumination estimation, greater precision is obtained over a variety of both indoor and outdoor scenes. We additionally show that the 3D positions of local point lights can be estimated using this method, by triangulating the light source positions from the environment maps of multiple faces in an image. With this 3D lighting information, the spatially variant illumination throughout a scene can be obtained. Recovering the detailed illumination in a scene not only benefits AR applications but also can promote scene understanding in general.

2 Related work

Highlight extraction involves separating the diffuse and specular reflection components in an image. This problem is most commonly addressed by removing highlights with the help of chromatic [11, 46, 47, 52] as well as spatial [22, 44, 45]

information from neighboring image areas, and then subtracting the resulting diffuse image from the original input to obtain the highlight component. These techniques are limited in the types of surface textures that can be handled, and they assume that the illumination color is uniform or known.

In recent work [16], these restrictions are avoided for the case of human faces by utilizing additional constraints derived from physical and statistical face priors. Our work also focuses on human faces but employs a deep learning approach instead of a physics-based solution for highlight extraction. While methods developed from physical models have a tangible basis, they might not account for all factors that influence image appearance, and analytical models often provide only a simplified approximation of natural mechanisms. In this work, we show that directly learning from real image data can lead to improved results that additionally surpass deep learning on synthetic training data [41].

Illumination estimation is often performed from a single image, as this is the only input available in many applications. The majority of single-image methods assume known geometry in the scene and estimate illumination from shading [18,30,32,48] and shadows [18,26,27,36,37]. Some methods do not require geometry to be known in advance, but instead they infer this information from the image by employing priors on object geometry [1, 20, 25, 33] or by fitting shape models for faces [6, 10, 12, 17, 34]. Our work also makes use of statistical face models to obtain geometric information for illumination estimation.

An illumination environment can be arbitrarily complex, and nearly all previous works employ a simplified parametric representation as a practical approximation. Earlier techniques mainly estimate lighting as a small set of distant point light sources [18, 27, 36, 37, 48]. More recently, denser representations in the form of low-order spherical harmonics [1, 6, 10, 12, 17, 32, 34] and Haar wavelets [26] have been recovered. The relatively small number of parameters in these models simplifies optimization but provides limited precision in the estimated lighting. A more detailed lighting representation may nevertheless be infeasible to recover from shading and shadows because of the lowpass filtering effect of diffuse reflectance [32] and the decreased visibility of shadow variations under extended lighting.

Greater precision has been obtained by utilizing lighting models specific to a certain type of scene. For outdoor environments, sky and sun models have been used for accurate recovery of illumination [3, 9, 13, 14]. In research concurrent to ours, indoor illumination is predicted using a convolutional neural network trained on data from indoor environment maps [5]. Similar to our work, it estimates a non-parametric representation of the lighting environment with the help of deep learning. Our approach differs in that it uses human faces to determine the environment map, and employs deep learning to recover an intermediate quantity, namely highlight reflections, from which the lighting can be analytically solved. Though our method has the added requirement of having a face in the image, it is not limited to indoor scenes and it takes advantage of more direct evidence about the lighting environment. We later show that this more direct evidence can lead to higher precision in environment map estimates.

Highlight reflections have been used together with diffuse shading to jointly estimate non-parametric lighting and an object's reflectance distribution function [19]. In that work, priors on real-world reflectance and illumination are utilized as constraints to improve inference in an optimization-based approach. The method employs an object with known geometry, uniform color, and a shiny surface as a probe for the illumination. By contrast, our work uses arbitrary faces, which are a common occurrence in natural scenes. As shown later, the optimization-based approach can be sensitive to the complications presented by faces, such as surface texture, inexact geometry estimation, and spatially-variant reflectance. Our method reliably extracts a key component of illumination estimation – highlight reflections – despite these obstacles by using a proposed deep learning scheme.

3 Overview

As shown in Figure 1, we train a deep neural network called *Highlight-Net* to extract the highlight component from a face image. This network is trained in two phases. First, pretraining is performed with synthetic data (Section 4.1). Subsequently, the network is finetuned in an unsupervised manner with real images from a celebrity dataset (Section 4.2).

For testing, the network takes an input image and estimates its highlight layer. Together with reconstructed facial geometry, the extracted highlights are used to obtain an initial environment map, by tracing the highlight reflections back towards the scene. This initial map is blurred due to the band-limiting effects of surface reflectance [32]. To mitigate this blur, our method performs deconvolution on the environment map using kernels determined from facial reflectance statistics (Section 5).

4 Face highlight removal

4.1 Pretraining with synthetic data

For Highlight-Net, we adopt a network structure used previously for intrinsic image decomposition [24], a related image separation task. To pretrain this network, we render synthetic data using generic face models [29] and real indoor and outdoor HDR environment maps collected from the Internet. Details on data preparation are presented in Section 6.1. With synthetic ground truth specular images, we minimize the L2 loss between the predicted and ground truth highlights for pretraining.

4.2 Unsupervised finetuning on real images

With only pretraining on synthetic data, Highlight-Net performs inadequately on real images. This may be attributed to the limited variation of face shapes, textures, and environment maps in the synthetic data, as well as the gap in

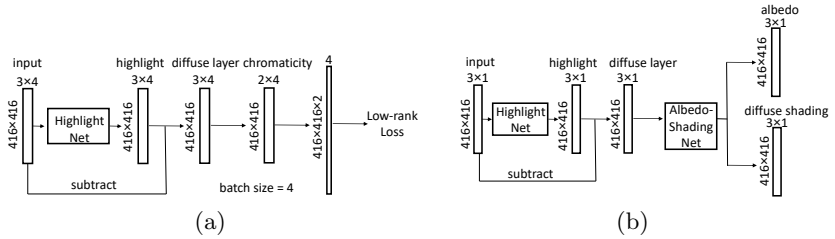


Fig. 2. (a) Network structure for finetuning Highlight-Net; (b) Testing network structure for separating an input face image into three layers: highlight, diffuse shading, and albedo.

appearance between synthetic and real face images. Since producing a large-scale collection of real ground-truth highlight separation data is impractical, we present an unsupervised strategy for finetuning Highlight-Net that only requires real images of faces under varying illumination environments.

This strategy is based on the observation that the diffuse chromaticity over a given person's face should be consistent in different images, regardless of illumination changes, because a person's facial surface features should remain the same. Among images of the same face, the diffuse chromaticity map should differ only by global scaling factors determined by illumination color and sensor attributes, which we correct in a preprocessing step. Thus, a matrix constructed by stacking the aligned diffuse chromaticity maps of a person should be of low rank. In place of ground-truth highlight layers of real face images, we use this low-rank property of ground-truth diffuse layers to finetune our Highlight-Net.

This finetuning is implemented using the network structure shown in Figure 2 (a), where Highlight-Net is augmented with a low-rank loss. The images for training are taken from the MS-celeb-1M database [7], which contains 100 images for each of 100,000 celebrities. After some preprocessing described in Section 6.1, we have a set of aligned frontal face images under a consistent illumination color for each celebrity.

From this dataset, four face images of the same celebrity are randomly selected for each batch. A batch is fed into Highlight-Net to produce the estimated highlight layers for the four images. These highlight layers are subtracted from the original images to obtain the corresponding diffuse layers. For a diffuse layer I_d, its diffuse chromaticity map is computed per-pixel as

$$chrom(I_d) = \frac{1}{(I_d(r) + I_d(g) + I_d(b))} \left(I_d(r), I_d(g)\right) \tag{1}$$

where r, g, and b denote the color channels. Each diffuse chromaticity map is then reshaped into a vector I^{dc}, and the vectors of the four images are stacked into a matrix $D = \left[I_1^{dc}, I_2^{dc}, I_3^{dc}, I_4^{dc}\right]^T$. With a low-rank loss enforced on D, Highlight-Net is finetuned through backpropagation.

Since the diffuse chromaticity of a face should be consistent among images, the rank of matrix D should ideally be one. So we define the low-rank loss as its

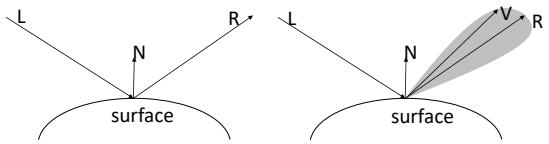


Fig. 3. Left: Mirror reflection. Right: Specular reflection of a rough surface.

second singular value, during backpropagation the partial derivative of σ_2 with respect to each matrix element is evaluated according to [28]:

$$\begin{aligned} &D = U\Sigma V^T, \qquad \Sigma = diag(\sigma_1, \sigma_2, \sigma_3, \sigma_4), \\ &loss_{lowrank} = \sigma_2, \qquad \frac{\partial \sigma_2}{\partial D_{i,j}} = U_{i,2} \times V_{j,2}. \end{aligned} \tag{2}$$

5 Illumination estimation

5.1 Environment map initialization

The specular reflections of a mirror are ideal for illumination estimation, because the observed highlights can be exactly traced back to the environment map when surface normals are known. This exact tracing is possible because a highlight reflection is directed along a single reflection direction R that mirrors the incident lighting direction L about the surface normal N, as shown on the left side of Figure 3. This raytracing approach is widely used to capture environment maps with mirrored spheres in computer graphics applications.

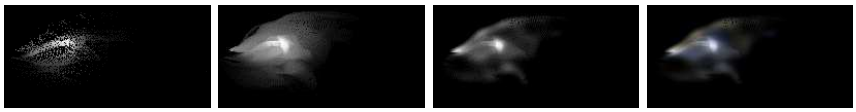
For the specular reflections of a rough surface like human skin, the light energy is instead tightly distributed around the mirror reflection direction, as illustrated on the right side of Figure 3. This specular lobe can be approximated by the specular term of the Phong model [31] as

$$I_s = k_s(R \cdot V)^\alpha, \qquad R = 2(L \cdot N)N - L \tag{3}$$

where k_s denotes the specular albedo, V is the viewing direction, and α represents the surface roughness. We specifically choose to use the Phong model to take advantage of statistics that have been compiled for it, as described later.

As rigorously derived in [32], reflection can be expressed as the environment map convolved with the surface BRDF (bidirectional reflectance distribution function), e.g., the model in Equation 3. Therefore, if we trace the highlight component of a face back toward the scene, we obtain a convolved version of the environment map, where the convolution kernel is determined by the specular reflectance lobe. With surface normals computed using a single-image face reconstruction algorithm [51], our method performs this tracing to recover an initial environment map, such as that exhibited in Figure 4 (a).

Due to limited image resolution, the surface normals on a face are sparsely sampled, and an environment map obtained by directly tracing the highlight



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

Fig. 4. Intermediate results of illumination estimation. (a) Traced environment map by forward warping; (b) Traced environment map by inverse warping; (c) Map after deconvolution; (d) Final environment map after illumination color rescaling.

component would be sparse as well, as shown in Figure 4 (a). To avoid this problem, we employ inverse image warping where for each pixel p in the environment map, trace back to the face to get its corresponding normal N_p and use the available face normals nearest to N_p to interpolate a highlight value of N_p. In this way, we avoid the holes and overlaps caused by directly tracing (i.e., forward warping) highlights to the environment map. The result of this inverse warping is illustrated in Figure 4 (b).

5.2 Deconvolution by the specular lobe

Next, we use the specular lobe to deconvolve the filtered environment map. This deconvolution is applied in the spherical domain, rather than in the spatial domain parameterized by latitude and longitude which would introduce geometric distortions.

Consider the deconvolution kernel K_x centered at a point $\mathbf{x} = (\theta_x, \phi_y)$ on the environment map. At a nearby point $\mathbf{y} = (\theta_y, \phi_y)$, the value of K_x is

$$K_x(\mathbf{y}) = k_s^x (L_y \cdot L_x)^{\alpha_x} \tag{4}$$

where L_x and L_y are 3D unit vectors that point from the sphere center toward $\mathbf{x}$ and $\mathbf{y}$, respectively. The terms α_x and k_s^x denote the surface roughness and specular albedo at $\mathbf{x}$.

To determine α_x and k_s^x for each pixel in the environment map, we use statistics from the MERL/ETH Skin Reflectance Database [50]. In these statistics, faces are categorized by skin type, and every face is divided into ten regions, each with its own mean specular albedo and roughness because of differences in skin properties, e.g., the forehead and nose being relatively more oily. Using the mean albedo and roughness value of each face region for the face's skin type[4], our method performs deconvolution by the Richardson-Lucy algorithm [21, 35]. Figure 4 (c) shows an environment map after deconvolution.

5.3 Rescaling illumination color

The brightness of highlight reflections often leads to saturated pixels, which have color values clipped at the maximum image intensity. As a result, the highlight

[4] Skin type is determined by the closest mean albedo to the mean value of the face's albedo layer. Extraction of the face's albedo layer is described in Sec. 5.3.

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

Fig. 5. (a) Input photo; (b) Automatically cropped face region by landmarks [53] (network input); (c) predicted highlight layer (scaled by 2); (d) highlight removal result.

intensity in these color channels may be underestimated. This problem is illustrated in Figure 5, where the predicted highlight layer appears blue because the light energy in the red and green channels is not fully recorded in the input image. To address this issue, we take advantage of diffuse shading, which is generally free of saturation and indicative of illumination color.

Diffuse reflection (i.e., the diffuse layer) is the product of albedo and diffuse shading, and the diffuse shading can be extracted from the diffuse layer through intrinsic image decomposition. To accomplish this decomposition, we finetune the intrinsic image network from [24] using synthetic face images to improve the network's effectiveness on faces. Specifically, 10,000 face images were synthesized from 50 face shapes randomly generated using the Basel Face Model [29], three different skin tones, diffuse reflectance, and environment maps randomly selected from 100 indoor and 100 outdoor real HDR environment maps. Adding this Albedo-Shading Net to our system as shown in Figure 2 (b) yields a highlight layer, albedo layer, and diffuse shading layer from an input face.

With the diffuse shading layer, we recolor the highlight layer H extracted via Highlight-Net by rescaling its channels. When the blue channel is not saturated, its value is correct and the other channels are rescaled relative to it as

$$[H'(r),\ H'(g),\ H'(b)] = [H(b) * c_d(r)/c_d(b),\ H(b) * c_d(g)/c_d(b),\ H(b)] \quad (5)$$

where c_d is the diffuse shading chromaticity. Rescaling can similarly be solved from the red or green channels if they are unsaturated. If all channels are saturated, we use the blue channel as it is likely to be the least underestimated based on common colors of illumination and skin. After recoloring the highlight layer, we compute its corresponding environment map following the procedure in Sections 5.1-5.2 to produce the final result, such as shown in Figure 4 (d).

5.4 Triangulating lights from multiple faces

In a scene where the light sources are nearby, the incoming light distribution can vary significantly at different locations. An advantage of our non-parametric illumination model is that when there are multiple faces in an image, we can

recover this spatially variant illumination by inferring the environment map at each face and using them to triangulate the 3D light source positions.

As a simple scheme to demonstrate this idea, we first use a generic 3D face model (e.g., the Basel Face Model [29]) to solve for the 3D positions of each face in the camera's coordinate system, by matching 3D landmarks on the face model to 2D landmarks in the image using the method of [53]. Highlight-Net is then utilized to acquire the environment map at each of the faces. In the environment maps, strong light sources are detected as local maxima found through non-maximum suppression. To build correspondences among the lights detected from different faces, we first match them according to their colors. When there are multiple lights of the same color, their correspondence is determined by triangulating different combinations between two faces, with verification using a third face. In this way, the 3D light source positions can be recovered.

6 Experiments

6.1 Training data

For the pretraining of Highlight-Net, we use the Basel Face Model [29] to randomly generate 50 3D faces. For each face shape, we adjust the texture map to simulate three different skin tones. These 150 faces are then rendered under 200 different HDR environment maps, including 100 from indoor scenes and 100 from outdoor scenes. The diffuse and specular components are rendered separately, where a spatially uniform specular albedo is randomly generated between $[0, 1]$. Some examples of these renderings are provided in the supplemental document. For training, we preprocessed each rendering by subtracting the mean image value and then normalizing to the range [0,1].

In finetuning Highlight-Net, the image set for each celebrity undergoes a series of commonly-used preprocessing steps so that the faces are aligned, frontal, radiometrically calibrated, and under a consistent illumination color. For face frontalization, we apply the method in [8]. We then identify facial landmarks [53] to crop and align these frontal faces. The cropped images are radiometrically calibrated by the method in [15], and their color histograms are matched by the built-in histogram transfer function in MATLAB [23] to reduce illumination color differences. We note that in each celebrity's set, images were manually removed if the face exhibits a strong expression or multiple lighting colors, since these cases often lead to inaccurate spatial alignment or poor illumination color matching. Some examples of these preprocessed images are presented in the supplementary material.

6.2 Evaluation of highlight removal

To examine highlight extraction performance, we compare our highlight removal results to those of several previous techniques [16, 40, 41, 47, 52] in Figure 6. The first two rows show results on faces with known ground truth captured by cross-polarization under an indoor directional light. In order to show fair comparisons

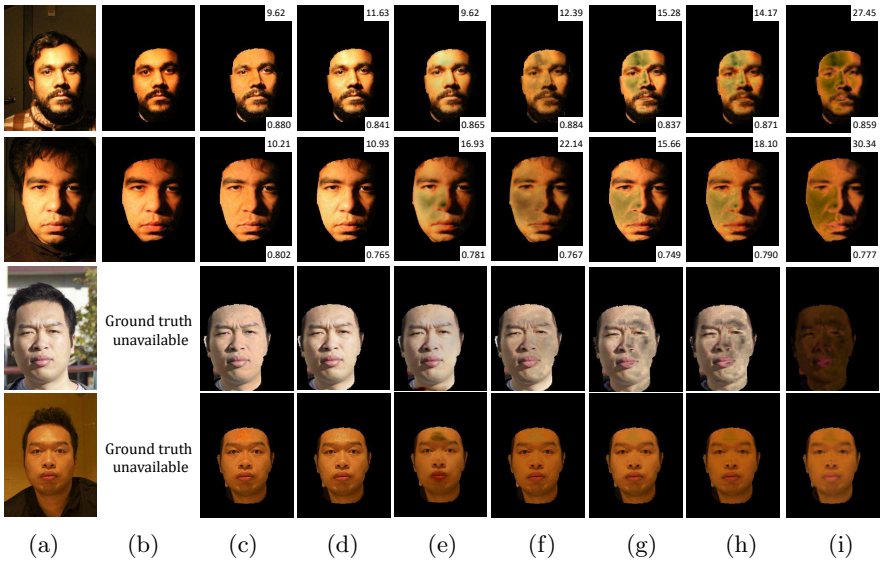


Fig. 6. Highlight removal comparisons on laboratory images with ground truth and on natural images. Face regions are cropped out automatically by landmark detection [53]. (a) Input photo. (b) Ground truth captured by cross-polarization for lab data. (c-h) Highlight removal results by (c) our finetuned Highlight-Net, (d) Highlight-Net without finetuning, (e) [41], (f) [16], (g) [40], (h) [52], and (i) [47]. For the lab images, RMSE values are given at the top-right, and SSIM [49] (larger is better) at the bottom-right.

for both absolute intensity errors and structural similarities, we use both RMSE and SSIM [49] as error/similarity metrics. The last two rows are qualitative comparisons on natural outdoor and indoor illuminations, where ground truth is unavailable due to the difficulty of cross-polarization in general settings. In all of these examples, our method outperforms the previous techniques, which generally have difficulty in dealing with the saturated pixels that commonly appear in highlight regions. We note that since most previous techniques are based on color analysis and the dichromatic reflection model [39], they cannot process grayscale images, unlike our CNN-based method. For results on grayscale images and additional color images, please refer to the supplement. The figure also illustrates the importance of training on real image data. Comparing our finetuning-based method in (c) to our method without finetuning in (d) and a CNN-based method trained on synthetic data [41] in (e) shows that training only on synthetic data is insufficient, and that our unsupervised approach for finetuning on real images substantially elevates the quality of highlight separation.

Quantitative comparisons over 100 synthetic faces and 30 real faces are presented in Table 1. Error histograms and image results are shown in the supplement.

	Synthetic data						Real data					
	Ours	[41]	[16]	[40]	[52]	[47]	Ours	[41]	[16]	[40]	[52]	[47]
Mean RMSE	**3.37**	4.15	5.35	6.75	8.08	28.00	**7.61**	8.93	10.34	10.51	11.74	19.60
Median RMSE	**3.41**	3.54	4.68	6.41	7.82	29.50	**6.75**	8.71	10.54	9.76	11.53	22.96
Mean SSIM	**0.94**	**0.94**	0.92	0.91	0.91	0.87	0.89	0.89	**0.90**	0.86	0.88	0.88
Median SSIM	**0.95**	0.94	0.92	0.91	0.91	0.87	0.90	0.90	**0.91**	0.88	0.90	0.89

Table 1. Quantitative highlight removal evaluation.

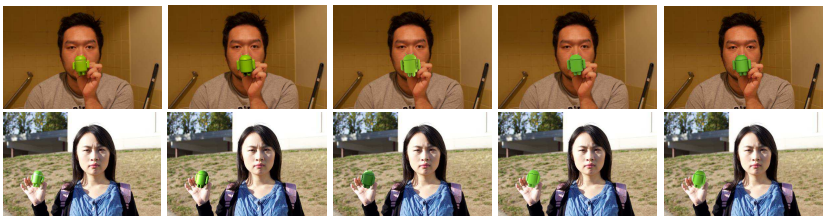
	Diffuse Bunny					Glossy Bunny				
Relighting RMSE	Ours	[9]	[5]	[19]	[12]	Ours	[9]	[5]	[19]	[12]
Mean (outdoor)	**10.78**	18.13	\	21.20	17.77	**11.02**	18.28	\	21.63	18.28
Median (outdoor)	**9.38**	17.03	\	19.95	15.91	**9.74**	17.67	\	20.49	16.30
Mean (indoor)	**13.18**	\	29.25	25.40	20.52	**13.69**	\	29.71	25.92	21.01
Median (indoor)	**11.68**	\	25.99	25.38	19.22	**11.98**	\	26.53	25.91	19.75

Table 2. Illumination estimation on synthetic data.

6.3 Evaluation of illumination estimation

Following [9], we evaluate illumination estimation by examining the relighting errors of a Stanford bunny under predicted environment maps and the ground truth. The lighting estimation is performed on synthetic faces rendered into captured outdoor and indoor scenes and their recorded HDR environment maps. Results are computed for both a diffuse and a glossy Stanford bunny (see the supplement for rendering parameters, visualization of rendered bunnies, and estimated environment maps). The comparison methods include the following: our implementation of [12] which uses a face to recover spherical harmonics (SH) lighting up to second order under the assumption that the face is diffuse; downloaded code for [19] which estimates illumination and reflectance given known surface normals that we estimate using [51]; online demo code for [9] which is designed for outdoor images; and author-provided results for [5] which is intended for indoor images.

The relighting errors are presented in Table 2. Except for [9] and [5], the errors were computed for 500 environment maps estimated from five synthetic faces under 100 real HDR environment maps (50 indoor and 50 outdoor). Since [9] and [5] are respectively for outdoor and indoor scenes and are not trained on faces, their results are each computed from LDR crops from the center of the 50 indoor/outdoor environment maps. We found [9] and [5] to be generally less precise in estimating light source directions, especially when light sources are out-of-view in the input crops, but they still provide reasonable approximations. For [5], the estimates of high frequency lighting become less precise when the indoor environment is more complicated. The experiments indicate that [19] may be relatively sensitive to surface textures and imprecise geometry in comparison to our method, which is purposely designed to deal with faces. For the Spherical Harmonics representation [12], estimates of a low-order SH model are seen to



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

Fig. 7. Virtual object insertion results for indoor (first row) and outdoor (second row) scenes. (a) Photos with real object. Object insertion by (b) our method, (c) [5] for the first row and [9] for the second row, (d) [19], (e) [12]. More results in the supplement.

Fig. 8. Object insertion results by our method.

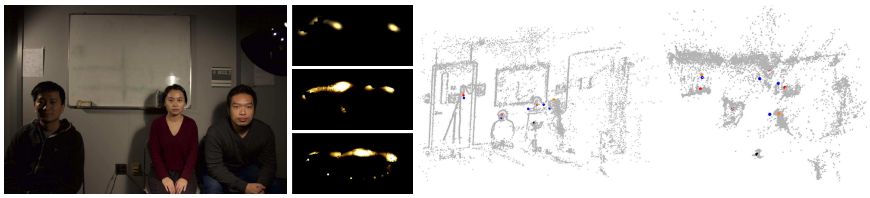
lack detail, and the estimated face albedo incorporates the illumination color, which leads to environment maps that are mostly white (see supplement for examples). Overall, the results indicate that our method provides the closest estimates to the ground truth. For a comparison of environment map estimation errors in real scenes, please refer to the supplement.

We additionally conducted comparisons on virtual object insertion using estimated illumination, as shown in Figure 7 and in the supplement. To aid in verification, we also show images that contain the actual physical object (an Android robot). In some cases such as the bottom of (c), lighting from the side is estimated as coming from farther behind, resulting in a shadowed appearance. Additional object insertion results are shown in Figure 8.

6.4 Demonstration of light source triangulation

Using the simple scheme described in Section 5.4, we demonstrate the triangulation of two local light sources from an image with three faces, shown in Figure 9 (a). The estimated environment maps from the three faces are shown in Figure 9 (b). We triangulate the point lights from two of them, while using the third for validation. In order to provide a quantitative evaluation, we use the DSO SLAM system [4] to reconstruct the scene, including the faces and light sources. We manually mark the reconstructed faces and light sources in the 3D point clouds as ground truth. As shown in Figure 9 (c-d), the results of our method are close to this ground truth. The position errors are 0.19m, 0.44m and 0.29m for the

faces from left to right, and 0.41m and 0.51m for the two lamps respectively. If the ground truth face positions are used, the position errors of the lamps are reduced to 0.20m and 0.49m, respectively.



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

Fig. 9. (a) Input image with multiple faces; (b) their estimated environment maps (top to bottom are for faces from left to right); estimated 3D positions from (c) side view and (d) top view. Black dot: camera. Red dots: ground truth of faces and lights. Blue dots: estimated faces and lights. Orange dots: estimated lights using ground truth of face positions.

7 Conclusion

We proposed a system for non-parametric illumination estimation based on an unsupervised finetuning approach for extracting highlight reflections from faces. In future work, we plan to examine more sophisticated schemes for recovering spatially variant illumination from the environment maps of multiple faces in an image. Using faces as lighting probes provides us with a better understanding of the surrounding environment not viewed by the camera, which can benefit a variety of vision applications.

Acknowledgments. This work is supported by Canada NSERC Discovery Grant 611664. Renjiao Yi is supported by scholarship from China Scholarship Council.

References

1. Barron, J.T., Malik, J.: Shape, illumination, and reflectance from shading. IEEE Trans Pattern Anal Mach Intell (PAMI) **37**(8), 1670–1687 (2015)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: ACM SIGGRAPH. pp. 187–194. ACM (1999)
3. Calian, D.A., Lalonde, J.F., Gotardo, P., Simon, T., Matthews, I., Mitchell, K.: From faces to outdoor light probes. In: Computer Graphics Forum. vol. 37, pp. 51–61. Wiley Online Library (2018)
4. Engel, J., Koltun, V., Cremers, D.: Direct sparse odometry. IEEE Transactions on Pattern Analysis and Machine Intelligence (2017)

5. Gardner, M.A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagné, C., Lalonde, J.F.: Learning to predict indoor illumination from a single image. ACM Transactions on Graphics (SIGGRAPH Asia) **9**(4) (2017)
6. Garrido, P., Valgaerts, L., Wu, C., Theobalt, C.: Reconstructing detailed dynamic face geometry from monocular video. ACM Transactions on Graphics (TOG) **32**(6), 158:1–158:10 (2013)
7. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: European Conference on Computer Vision. pp. 87–102. Springer (2016)
8. Hassner, T., Harel, S., Paz, E., Enbar, R.: Effective face frontalization in unconstrained images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4295–4304 (2015)
9. Hold-Geoffroy, Y., Sunkavalli, K., Hadap, S., Gambaretto, E., Lalonde, J.F.: Deep outdoor illumination estimation. In: IEEE International Conference on Computer Vision and Pattern Recognition (2017)
10. Kemelmacher-Shlizerman, I., Basri, R.: 3d face reconstruction from a single image using a single reference face shape. IEEE Trans Pattern Anal Mach Intell (PAMI) **33**(2), 394–405 (2011)
11. Kim, H., Jin, H., Hadap, S., Kweon, I.: Specular reflection separation using dark channel prior. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. pp. 1460–1467 (2013)
12. Knorr, S.B., Kurz, D.: Real-time illumination estimation from faces for coherent rendering. In: Mixed and Augmented Reality (ISMAR), 2014 IEEE International Symposium on. pp. 113–122. IEEE (2014)
13. Lalonde, J.F., Narasimhan, S.G., Efros, A.A.: What does the sky tell us about the camera? In: European conference on computer vision. pp. 354–367. Springer (2008)
14. Lalonde, J.F., Narasimhan, S.G., Efros, A.A.: What do the sun and the sky tell us about the camera? International Journal of Computer Vision **88**(1), 24–51 (2010)
15. Li, C., Lin, S., Zhou, K., Ikeuchi, K.: Radiometric calibration from faces in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3117–3126 (2017)
16. Li, C., Lin, S., Zhou, K., Ikeuchi, K.: Specular highlight removal in facial images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 3107–3116 (2017)
17. Li, C., Zhou, K., Lin, S.: Intrinsic face image decomposition with human face priors. In: Proceedings of European Conference on Computer Vision (2014)
18. Li, Y., Lin, S., Lu, H., Shum, H.Y.: Multiple-cue illumination estimation in textured scenes. In: Proceedings of International Conference on Computer Vision. pp. 1366–1373 (2003)
19. Lombardi, S., Nishino, K.: Reflectance and illumination recovery in the wild. IEEE Trans Pattern Anal Mach Intell (PAMI) **38**(1), 129–141 (2016)
20. Lopez-Moreno, J., Hadap, S., Reinhard, E., Gutierrez, D.: Compositing images through light source detection. Computers & Graphics **34**(6), 698–707 (2010)
21. Lucy, L.B.: An iterative technique for the rectification of observed distributions. The astronomical journal **79**, 745 (1974)
22. Mallick, S.P., Zickler, T., Belhumeur, P.N., Kriegman, D.J.: Specularity removal in images and videos: A pde approach. In: Proceedings of European Conference on Computer Vision (2006)
23. Mathworks: Matlab r2014b, https://www.mathworks.com/products/matlab.html

24. Narihira, T., Maire, M., Yu, S.X.: Direct intrinsics: Learning albedo-shading decomposition by convolutional regression. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 2992–2992 (2015)
25. Nishino, K., Nayar, S.K.: Eyes for relighting. In: ACM Transactions on Graphics (TOG). vol. 23, pp. 704–711. ACM (2004)
26. Okabe, T., Sato, I., Sato, Y.: Spherical harmonics vs. haar wavelets: Basis for recovering illumination from cast shadows. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 50–57 (2004)
27. Panagopoulos, A., Wang, C., Samaras, D., Paragios, N.: Illumination estimation and cast shadow detection through a higher-order graphical model. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2011)
28. Papadopoulo, T., Lourakis, M.I.: Estimating the jacobian of the singular value decomposition: Theory and applications. In: European Conference on Computer Vision. pp. 554–570. Springer (2000)
29. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: Advanced Video and Signal Based Surveillance, 2009. AVSS'09. Sixth IEEE International Conference on. pp. 296–301. Ieee (2009)
30. Pessoa, S., Moura, G., Lima, J., Teichrieb, V., Kelner, J.: Photorealistic rendering for augmented reality: A global illumination and brdf solution. In: Virtual Reality Conference (VR), 2010 IEEE. pp. 3–10. IEEE (2010)
31. Phong, B.T.: Illumination for computer generated pictures. Communications of the ACM **18**(6), 311–317 (1975)
32. Ramamoorthi, R., Hanrahan, P.: A signal-processing framework for inverse rendering. In: ACM SIGGRAPH. pp. 117–128. ACM (2001)
33. Rematas, K., Ritschel, T., Fritz, M., Gavves, E., Tuytelaars, T.: Deep reflectance maps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4508–4516 (2016)
34. Richardson, E., Sela, M., Or-El, R., Kimmel, R.: Learning detailed face reconstruction from a single image. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2017)
35. Richardson, W.H.: Bayesian-based iterative method of image restoration. JOSA **62**(1), 55–59 (1972)
36. Sato, I., Sato, Y., Ikeuchi, K.: Acquiring a radiance distribution to superimpose virtual objects onto a real scene. IEEE Trans Vis Comput Graph (TVCG) **5**, 1–12 (1999)
37. Sato, I., Sato, Y., Ikeuchi, K.: Illumination from shadows. IEEE Trans Pattern Anal Mach Intell (PAMI) **25**, 290–300 (2003)
38. Sela, M., Richardson, E., Kimmel, R.: Unrestricted facial geometry reconstruction using image-to-image translation. In: Proceedings of International Conference on Computer Vision (2017)
39. Shafer, S.: Using color to separate reflection components. Color Research & Application **10**(4), 210–218 (1985)
40. Shen, H.L., Zheng, Z.H.: Real-time highlight removal using intensity ratio. Applied optics **52**(19), 4483–4493 (2013)
41. Shi, J., Dong, Y., Su, H., Yu, S.X.: Learning non-lambertian object intrinsics across shapenet categories. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2017)
42. Shrivastava, A., Pfister, T., Tuzel, O., Susskind, J., Wang, W., Webb, R.: Learning from simulated and unsupervised images through adversarial training. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2017)

43. Snap. Inc.: Snapchat, https://www.snapchat.com/
44. Tan, P., Lin, S., Quan, L.: Separation of highlight reflections on textured surfaces. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. vol. 2, pp. 1855–1860 (2006)
45. Tan, P., Lin, S., Quan, L., Shum, H.Y.: Highlight removal by illumination-constrained inpainting. In: Proceedings of International Conference on Computer Vision (2003)
46. Tan, R., Ikeuchi, K.: Reflection components decomposition of textured surfaces using linear basis functions. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. vol. 1, pp. 125–131 (2005)
47. Tan, R.T., Nishino, K., Ikeuchi, K.: Separating reflection components based on chromaticity and noise analysis. IEEE transactions on pattern analysis and machine intelligence **26**(10), 1373–1379 (2004)
48. Wang, Y., Samaras, D.: Estimation of multiple illuminants from a single image of arbitrary known geometry. In: Proceedings of European Conference on Computer Vision. pp. 272–288 (2002)
49. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
50. Weyrich, T., Matusik, W., Pfister, H., Bickel, B., Donner, C., Tu, C., McAndless, J., Lee, J., Ngan, A., Jensen, H.W., et al.: Analysis of human faces using a measurement-based skin reflectance model. In: ACM Transactions on Graphics (TOG). vol. 25, pp. 1013–1024. ACM (2006)
51. Yang, F., Wang, J., Shechtman, E., Bourdev, L., Metaxas, D.: Expression flow for 3d-aware face component transfer. In: ACM Transactions on Graphics (TOG). vol. 30, p. 60. ACM (2011)
52. Yang, Q., Wang, S., Ahuja, N.: Real-time specular highlight removal using bilateral filtering. Computer Vision–ECCV 2010 pp. 87–100 (2010)
53. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on. pp. 2879–2886. IEEE (2012)