Learning Physically-based Material and Lighting Decompositions for Face Editing

Qian Zhang[1,2], Vikas Thamizharasan[1], James Tompkin
Brown University

Abstract

Lighting is crucial for portrait photography, yet the complex interactions between skin and incident light are expensive to model computationally in graphics and hard to reconstruct analytically via computer vision. Instead, to allow fast and controllable reflectance and lighting editing, we form a physically-based decomposition through deep learned priors from path-traced portrait images. Previous approaches use simplified material models or low-frequency or low-dynamic-range lighting struggle to model specular reflections, or relight directly without intermediate decomposition. Instead, we estimate surface normal, skin albedo and roughness, and high-frequency HDRI maps, and propose an architecture to estimate both diffuse and specular reflectance components. In experiments, we show that this approach can better represent the true appearance function than simpler baseline methods, leading to better generalization and higher-quality editing. *Data, code, and results:* https://github.com/brownvc/phaced.

Key Words: Intrinsic decomposition, portrait relighting, inverse rendering, deep learning.

1. Introduction

Lighting is a crucial factor in successful portrait photography. Photographers set up studio lights and reflectors to enhance the appearance of subjects, with careful consideration for the appearance of skin to avoid unwanted gloss and highlights. For casual camera users, this level of control is difficult to achieve. Editing lighting and material appearance after a photo has been taken might simplify the creation process for novices, but current tools require manual operation and skill to produce convincing effects.

Decomposing an image into useful channels could help the portrait manipulation task. Under Lambertian reflectance assumptions, intrinsic decomposition separates the material color—the *albedo*—from the received illumination at each pixel of an image. This allows edits and recombination with novel lighting or materials. However, faces are not Lambertian, and require complex lighting and material models to more accurately decompose an image into useful intermediate channels. Further, such decompositions are ill-posed, and so must consider how to incorporate assumptions or priors to produce plausible answers.

Our work focus on the problem of single image face decomposition using physically-based lighting and material models. First, we consider diffuse and specular reflectance under a Cook-Torrance SVBRDF model, consisting of separate skin albedo, and specular scaling coefficient (ρ) and roughness (m) maps. Next, we consider that high-dynamic-range lighting with high spatial frequency is critical for specular appearance. As such, we create realistic synthetic data using real-world face geometry captures, real-world reflectometer measurements of skin, and real-world HDRI illumination with self-shadowing via path tracing.

To produce plausible decompositions, we supervise training of a deep neural network to estimate from a single face image a normal map, albedo map, specular scaling and roughness maps, and an approximate HDR incoming lighting map. Then, as realistic shadowing and glossy reflection rendering is computationally expensive, we use these physically-based maps to predict diffuse shading and specular maps given the lighting as conditioning information. Finally, we reconstruct the outputs using our image formation model. Each intermediate image formation model component (and so network architecture) can be supervised explicitly for stability, with final end-to-end fine tuning.

We operate directly on linear HDR images as specular illumination components are often clipped/saturated in LDR images. This allows more accurate specular reconstruction, Linearity also makes our decompose-edit-compose pipeline possible without introducing any non-linear errors due to tone mapping, which eases later editing and compositing. For instance, such a decomposition allows relighting with plausible specular highlights, along with shading editing and gloss and sharp specular highlight editing. In comparisons to baselines with simpler lighting and material models, and to pure relighting methods that do not decompose to intermediate maps, our method is better able to reproduce specularity and shading, and so provide more control in editing and more accurate relighting.

In short, our work argues that portrait editing can benefit

[1] Equal contribution. [2] Corresponding author.

	Reflectance Model	Illumination Model	Geometry Representation	Self-Shadow	Code	Data	Input Image
Method				Yes/No?	Released?	Released?	LDR/HDR?
Yamaguchi et al. [37]	Textures; explicit	De-lighting only; implicit	$3DMM^{\triangle}$ + Displacement map	NA	No	No	LDR
Lattas et al. (AvatarMe) [19]	Textures; explicit	De-lighting only; explicit	$3DMM^{\triangle}$ + Displacement map	NA	No	Yes	LDR
Dib et al. [9]	BSDF; explicit	Area lights; explicit	3DMM	Yes	No	No	LDR
Mallikarjun et al. [22]	Reflectance field rep.	OLAT basis; explicit	3DMM	Yes	No	No	LDR
Smith et al. (AlbedoMM) [29]	BSDF; explicit	SH2; explicit	3DMM	No	Yes	No	LDR
Zhou et al. (DPR) [41]	Lambertian; implicit	SH2; explicit	Normals	No	Yes	Yes	LDR
Sun et al. [30]	None; implicit	Environment map; explicit	None	Implicit	No	No	HDR
Hou et al. [12]	Lambertian; implicit	SH2; explicit	3DMM	Yes	Yes	Yes	LDR
Sengupta et al. (SfSNet) [28]	Lambertian; explicit	SH2; explicit	Normals	No	Yes	Yes	LDR
Nestmeyer et al. [23]	Lambertian+Residual; explicit	Directional Light; implicit	Normals	Yes	No	No	HDR
Wang et al. [33]	BSDF; explicit	Environment map; explicit	Normals	Yes	No	No	LDR
Ours	BSDF; explicit	Environment map; explicit	Normals	Implicit	Yes	Yes	HDR

Table 1. Comparison of closely-related state-of-the-art face appearance modeling methods. $\triangle$: proxy for UV unwrapping. The table was created based on our best-effort understanding of published methods from their papers and presentations. The first block defines techniques that estimate explicit geometry; the second block are 'direct' relighting methods without decomposition; the third block are 2D decomposition methods and are most closely related to our work.

from learning priors for decompositions through physically-based image formation models. Our contributions are:

- A realistic synthetic face image generation pipeline using public available face assets, creating a high quality synthetic face dataset with specularity and self-occlusions under varying lighting conditions,
- A method to decompose HDR image via a physically-based image formation model, allowing editing of properties like spatially-varying specular gloss.

Our work helps to shed light on how to accomplish accurate image decomposition without access to expensive light stage captures. We will release our source code and dataset for further research in the community.

2. Related work

We discuss classic and recent methods that address closely-related problems. We also provide an additional table of closely-related work (Table 1), This relates reflectance models, illumination models, geometry models, and model features, as well as whether code and data are available for each technique.

Intrinsic decomposition These commonly assume classic monochromatic illumination (MI) [2] or Retinex constraints [18]. Li et al. [20] used statistics of skin reflectance and facial geometry as constraints in an optimization for intrinsic components. Recently, end-to-end learning approaches embed priors in neural networks via synthetic images [15, 21]. Better results can be achieved with hybrid training of synthetic and real data [28] or with high-quality real images [23, 33].

Skin reflectance models Face appearance modeling is well-studied in computer graphics. One common approach is the Torrance-Sparrow specular BRDF, as used by Weyrich et al. [35] to develop an analytic spatially-varying face rendering model with measured skin data. Recent analysis works employ it. For example, based upon 3DMM face geometry, Smith et al. [29] build a statistical model for human face appearance including both diffuse and specular albedo. Subsurface scattering is an additional skin appearance component [24] that is computationally expensive to model; similar to most concurrent works, we do not assume this appearance factor in our model.

Face decomposition With capture setups like light stages, recent approaches have trained deep neural networks to estimate physically-based reflectance from monocular images, usually in a supervised manner [37, 28, 5, 23, 19, 33]. Sengupta et al. [28] assume Lambertian reflectance, while Nestmeyer et al. [23] and Wang et al. [33] predict specularity in addition, though not with a decomposed skin model. Yamaguchi et al. [37] and Lattas et al. [19] trained deep neural networks to infer high-quality geometry, diffuse, specular albedo, and displacement map from a single image but do not explicitly model illumination. In our work, we avoid the problem of geometry estimation and work only in screen space with normal maps. Finally, differentiable ray tracing can produce accurate reconstructions [10] with more realistic self-shadows and without large databases, though it is computationally expensive.

Lighting representation Spherical harmonics (SH) capture low-frequency signals efficiently for fast rendering [26, 1], and have been used at 2nd order to cheaply model the irradiance onto the face for diffuse shading [25]. Zhou et al. [41] use SH illumination to learn to relight a single input face image. For more accurate decompositions, Kanamori et al. [16] precompute light occlusion in the SH formulation directly for human body relighting.

Environment maps store sampled light and are often in a high dynamic range (HDR) for image-based lighting [8]. Many face works choose this representation as it can sample high-frequency signals, though deep learning models often

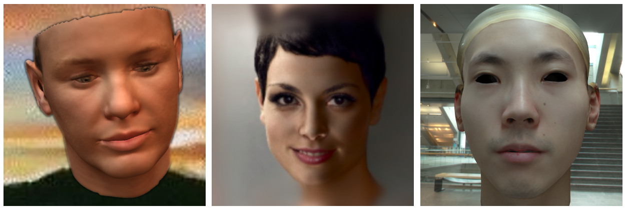

Figure 1. Training dataset comparison. From left to right: SfSNet uses normals derived from 3DMM geometry, SH2 approximated LDR lighting, and diffuse reflectance [28]. DPR uses SH2 approximated LDR lighting, Lambertian reflectance, and normals derived from 3DMM fit to CelebA. To improve upon these, our dataset uses realistic captured FaceScape geometry [38], high-frequency HDRI environment maps, and Torrance-Sparrow SVBRDF reflectance with physically-measured parameters mapped from Weyrich et al. [35], leading to increased realism.

use smaller (32 × 16) maps. Yi et al. [39] trace specular highlights into the scene and obtain an environment map through a deconvolution determined by prior knowledge of face materials. Calian et al. [3] use faces as light probes to estimate HDR lighting from a single LDR photograph, learning priors through an autoencoder. Sun et al. [30] estimate high frequency environment maps at the bottleneck for portrait image relighting. Nestmeyer et al. [23] assume directional lighting and model specularity as a non-diffuse ‘residual’ term in their image formation process.

Our work takes the decomposition approach with HDR environment maps and a physically-based model of skin. On high-quality supervised data, we show that this can improve editing quality and capability over simpler decomposition and pure relighting approaches.

3. Dataset generation

High-quality data is important for overall model quality and generalization, but is expensive to acquire via light stages and so is often proprietary. As such, the research community has created synthetic databases for face decomposition and relighting [28, 41] (Fig. 1). Our approach increases data realism; we will release scripts to generate our data for further research in face analysis and editing.

Renderer and shading model We generate our synthetic dataset in Blender [7] and use the physically-based path tracing renderer *Cycles*. Our synthetic faces are modeled with Blender’s Principled BSDF, which is based on Disney’s “PBR” shader, itself derived from the Torrance-Sparrow model [31, 32]. The rendering integral for this diffuse and specular model is:

$$
\begin{aligned}
L(x,\omega_o) &= \int_\Omega \alpha(x) L(x,\omega_i)(N\cdot\omega_i)d\omega_i \\
&+ \int_\Omega f_{sTS}(x,\omega_o,\omega_i)L(x,\omega_i)(N\cdot\omega_i)d\omega_i
\end{aligned} \tag{1}
$$

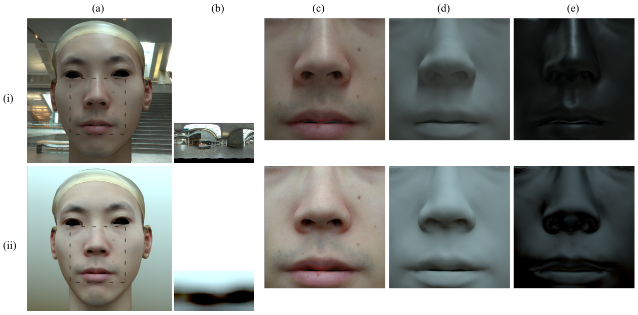

Figure 2. High frequency HDR lighting and self occlusion is required for accurate shadowing and specular reproduction. *First row:* Images rendered with HDR environment map. *Second row:* Images rendered with second order spherical harmonic approximation **(ii,b)**. Column **(a)** Rendered image, **(c)** Rendered image (zoomed in), **(d)** Diffuse shading, **(e)** Specular component.

where:

$$
f_{sTS} = \rho_s \frac{1}{4}\frac{DGF_r(\omega_o\cdot H)}{(N\cdot\omega_i)(N\cdot\omega_o)}, \tag{2}
$$

with:

$$
G = \min\{1, \frac{2(N\cdot H)(N\cdot\omega_o)}{(\omega_o,H)}, \frac{2(N\cdot H)(N\cdot\omega_i)}{(\omega_o,H)}\}. \tag{3}
$$

G is the geometry term, D is the micro-facet distribution, and F_r is the reflective Fresnel term. We have a factor of 4 in the denominator instead of π in the original Torrance-Sparrow paper [31] as we use the GGX micro-facet distribution [32]. The free variables determining specular appearance are the surface normal N, lighting L_Ω, albedo $\alpha(x)$, specular scaling coefficient ρ_s and roughness m.

Geometry and albedo Our face geometry and diffuse albedo data comes from the large-scale 3D face dataset FaceScape [38], consisting of 18,760 detailed 3D face models and high resolution albedo and displacement maps, captured from 938 subjects each with 20 expressions.

Skin reflectance We use skin reflectance statistics from the MERL/ETH Skin Reflectance Database [35], which provides per-face-region estimates. For each face in FaceScape’s captured data, we find the closest matching face regions in the MERL/ETH dataset using the per-face-region diffuse albedo, and then sample specular roughness m and scaling coefficient ρ for specular response. Rather than constant ρ and m for all face regions across all individuals [33], our approach uses spatially-varying specularity. We split the face into 10 regions and randomly sampled the Torrance-Sparrow specular reflectance parameters per face region as in Weyrich et al. [35]. The face regions and variation of parameters are shown in Fig. 15 of their paper. We manually aligned the FaceScape geometry and albedo data to have the same 10 face regions.

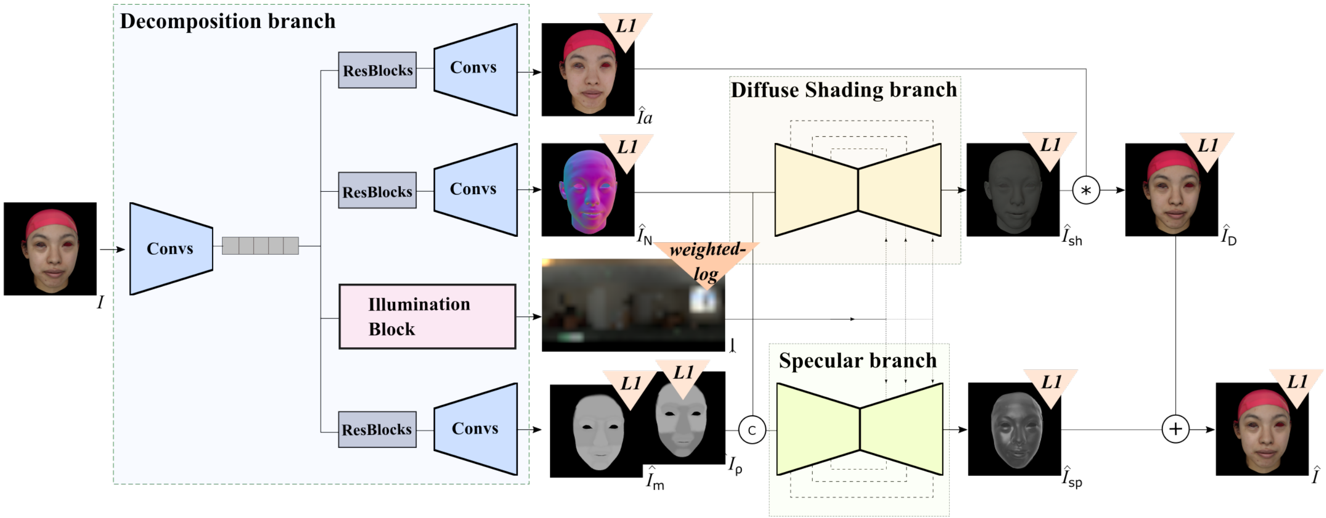

Figure 3. Our neural network architecture is composed of three blocks. *Blue:* First, a decomposition block uses a shared bottleneck to produce the constituent maps for shading. *Yellow:* Second, a diffuse shading branch uses lighting conditioning [33] in the decoder to produce a shading map (quotient image). *Green:* Third, a specular shading branch takes skin roughness and scaling maps and, again via lighting conditioning, creates the specular map. Finally, the image is linearly composited.

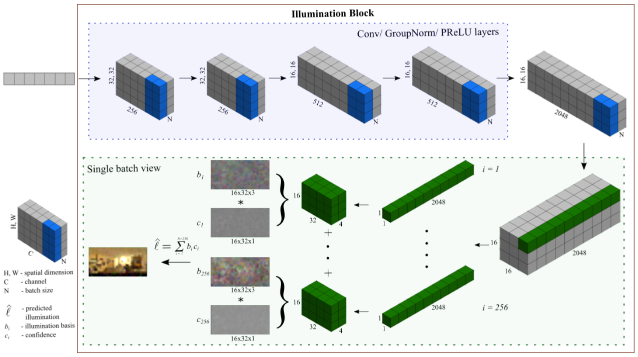

Figure 4. Illumination Block architecture. Blue highlighted cubes indicate application of group normalization [36].

Our renderings do not include subsurface scattering as these numerical parameters are not provided in the MERL/ETH dataset. Rendering low-noise subsurface scattering with a path tracer is computationally expensive, taking more than 30 seconds per image in Blender (GPU) [7] and Mitsuba1 (CPU) [14] renderers even for noisy outputs.

Lighting representation We use a $32 \times 16 \times 3$ resolution HDR image [23, 30, 33] rather than an SH2 approximation [28, 41]. SH2 approximations cannot capture illumination effects like hard shadows from self-occlusion or accurate specular reflectance (Fig. 2 compares via path tracing). However, using even low-resolution HDR maps is a trade off, as more parameters must be estimated than SH2 or spherical Gaussian models, and so a larger neural network is required. This choice of environment lighting representation was also adopted in recent works [30, 33]. Later, we will show the importance of higher-frequency illumination in the network's ability to model these complex effects (Figs. 5 and 9). Our equirectangular HDR environment maps are selected from the Laval Indoor HDR Dataset [11]. We choose the environment maps randomly, and replace only for very dark lighting conditions.

Output We render our data on a NVIDIA Quadro RTX 6000 GPU, taking $\approx$ 18 seconds per image. We export each component as a 512×512 image, in 32-bit high dynamic range where appropriate: normal I_N, albedo I_α, lighting l, scaling coefficient I_ρ, roughness I_m, as well as intermediate diffuse shading I_{sh} (sometimes called a quotient image), specular I_{sp}, albedo modulated diffuse shading I_D, and final output I. For reproduction in the paper, all HDR images are tone mapped via the Reinhard operator [27].

4. Decomposition architecture

4.1. Neural network

Given a dataset of face images with generated supervision for a physically-based deep learning approach, we take inspiration from Sengupta et al. [28] and Nestmeyer et al. [23] and design a three-stage approach with decomposition, diffuse shading, and specular branches (Fig. 3).

1. Decomposition branch This takes as input a single portrait image I and decomposes it into the diffuse albedo map ($\hat{I}_\alpha$), surface normal map ($\hat{I}_N$), specular reflectance parameter maps $\hat{I}_\rho$ and $\hat{I}_m$, and illumination ($\hat{l}$). The decomposition branch must extract all relevant information from the face, and it is important that the features embedded in

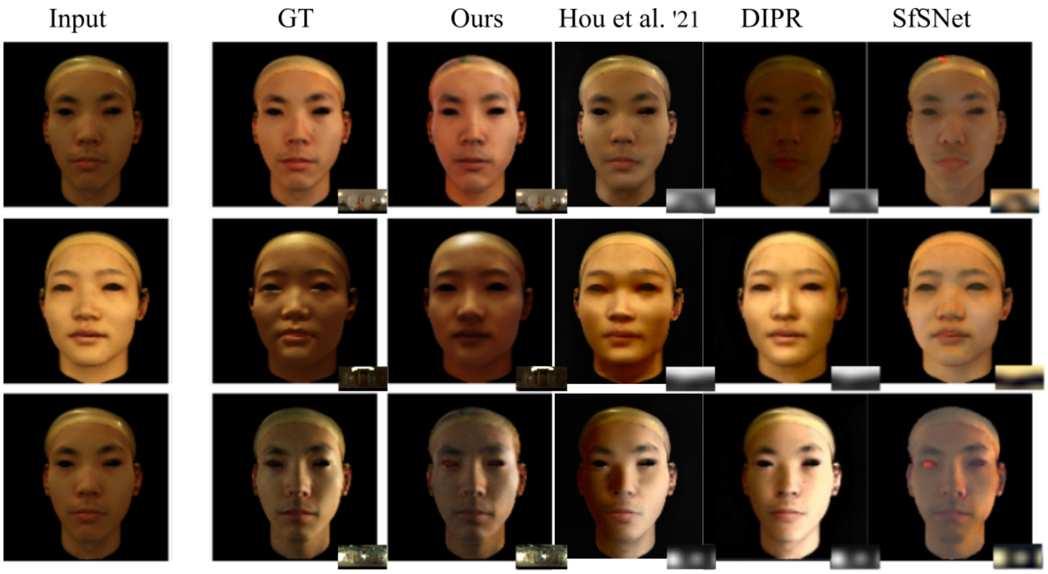

Figure 5. Comparison of our relit results with Hou et al. [12], DIPR and SfSNet

the bottleneck are not invariant to lighting as otherwise it would be impossible to predict an environment map. As such, our design encodes a large bottleneck from the input image, from which separate ResBlocks and decoders can transform and up-convolve maps back to input resolution.

1a. Illumination block Fig. 4 shows detailed architecture of the illumination block, This block estimates a high frequency $16 \times 32 \times 3$ environment map from the bottleneck encoded from the input image. Taking inspiration from Hu et al. [13] and Sun et al. [30], our method decomposes 256 localized environment maps (referred to as illumination basis in the figure) and 256 corresponding confidence maps. Then, these are combined in a weighted sum to form the estimated environment map. This network architecture was first proposed by Hu et al. [13] as a "confidence learning" approach for color constancy, and then adapted by Sun et al. [30] for low-resolution environment map prediction.

2. Diffuse shading branch This takes a normal map as input with illumination conditioning to produce the shading layer ($\hat{I}_{sh}$). It is built from a U-Net autoencoding architectures with skip connections, with additional lighting conditioning on the decoder. Inspired by Wang et al. [33], we observed that illumination is best fed as a feature-wise linear modulation at the up-convolution layers, analogous to AdaIN in StyleGAN [17], rather than concatenation at the input stage or at the bottleneck stage [30].

3. Specular branch This is also a U-Net with illumination conditioning in the decoder. It takes a normal map and specular reflectance parameter maps $\hat{I}_\rho$ and $\hat{I}_m$ as input to produce the specular layer $\hat{I}_{sp}$.

Final output We construct the final image $\hat{I}$ simply as: $\hat{I}_D = \hat{I}_{sh} * \hat{I}_\alpha$; $\hat{I} = \hat{I}_D + \hat{I}_{sp}$. Unlike Wang et al.'s lighting network [33], our final image is created from a linear combination of estimated shadings, making editing operations on the inferred maps easier and faster.

4.2. Training

Two-stage training Each branch is initially trained separately on 128×128 input images, then all three are combined and fine tuned end to end for reconstruction. The intuition behind the two-stage training process is that, for practical reasons, it is more efficient to train the hyperparameters of a large multi-decoder network, as opposed to training it end-to-end from scratch.

HDR space and data normalization Unlike previous works that operate on LDR images, we use HDR images as the linear property of pixel intensities and the environment map is critical for accurate specular reconstruction and avoiding artifacts like clipping.

Given that we are operating with HDR images, data normalization becomes critical to network training. Simple standardization will fail to reconcile the large differences in distributions between the input image, each reflectance component, and lead to unstable training. As such, following Weber et al. [34], we use normalization techniques (Alg. 1) to preserve the dynamic range prior to standardization, which can be reversed after inference.

Losses Given our rich synthetic data, we penalize L_1 supervised losses on all components. For the HDR illumination ($\hat{l}$), we penalize a weighted-log L_1 loss, weighted by

Algorithm 1 Data normalization routine. Notation: normal I_N, albedo I_α, illumination l, specular scaling coefficient map I_ρ, spcular roughness map I_m, diffuse shading (quotient image) I_{sh}, specular shading I_{sp}, diffuse lit face I_D, and input image I.

```
Input: Set $\mathcal{L}$ of HDR environment maps.
Input: Predetermined median $\hat{m} \leftarrow 0.5$.
Input: $i \leftarrow \{0, \ldots, |\mathcal{L}| - 1\}$.
Compute median $m_i \forall l_i \in \mathcal{L}$.
Compute exposure correction $e_i \leftarrow m_i / \hat{m}$.
$l^i \leftarrow l^i * e_i$.
Compute mean exposure $\bar{m} \forall e_i$.
$l^i \leftarrow \log(l^i)$.
Apply $\bar{m}$ to $I_N, I_\alpha, I_\rho, I_m, I_{sh}, I_{sp}, I_D, I$.
Standardize.
```

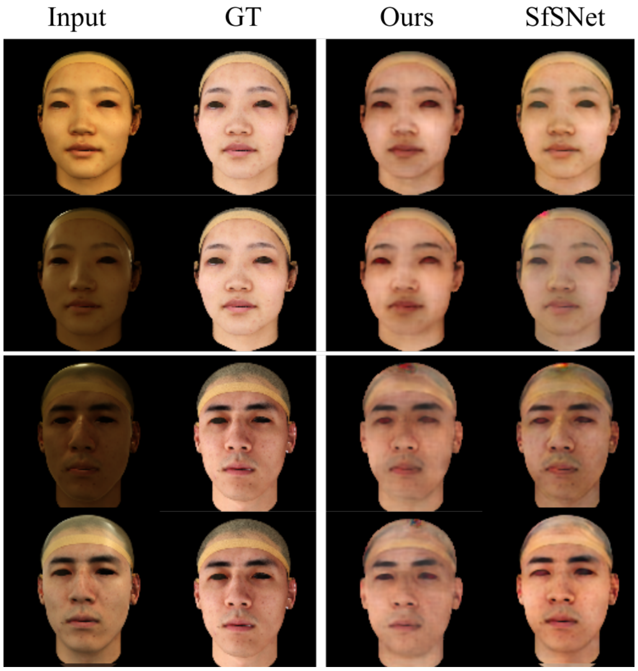

Figure 6. **Diffuse albedo estimation consistency**: Our network predicts more consistent diffuse albedo across different illuminations for the same individual (top pair, bottom pair), while SfSNet is less consistent.

the solid angle of each pixel over the sphere [34]. We set λ_ρ, λ_m = 1.0, λ_N, λ_α, λ_{sh}, λ_D, λ_{sp}, λ_I = 0.8 and λ_l = 0.1, *where λ_i is the weight on the respective loss for component i.*

5. Experiments

Dataset We render 100 face identities each under 25 random (of 100) different illumination environments, producing 2,500 training samples. Then, we render a test set of 100 other face identities each under 10 random (of a set of 20) test illuminations, producing 1,000 test samples. For our qualitative results, we show examples from only ten authorized identities [38], none of which are in the training set or the test set used for numeric comparisons.

Baselines For decompositions, we compare to public baselines. We consider SfSNet from Sengupta et al. [28], which is a 2D single-image decomposition approach that assumes SH2 lighting and diffuse reflectance only. We retrain SfSNet on our more realistic image data. We also compare to AlbedoMM from Smith et al [29], which is a geometric fitting approach based on 3DMM with diffuse and specular statistical model components. This cannot be retrained on our data. Finally, for relighting, we compare to DIPR from Zhou et al. [41], which uses a SH2 bottleneck to relight without decomposition.

Ablation—Neither ρ nor m Fig. 10 demonstrate that estimating a specular map directly from normal and bypassing separate ρ and m maps produces substantially less accurate specular results for our architecture.

Results—Quantitative evaluation In Table 3, we quantitatively compare our decomposition and reconstruction results with SfSNet using $L1$ (the loss that both methods penalize), mean-squared error, and perceptual LPIPS [40] metrics. Our method produces more accurate reconstructions overall, with equivalent albedo estimates and better shading estimates. Given SfSNet's assumptions, we also show results when only diffuse effects are in the input (column 'Diffuse') for reconstruction without specular effects. Here, our method shows smaller gains over SfSNet. For specularity, we compare to AlbedoMM [29] in Table 2. We use their probabilistic fitting pipeline [1] to estimate 3DMM parameters. The estimated specular maps are of lower quality, partly because of geometric fitting inaccuracies.

Results—Qualitative evaluation
Albedo estimation: Ours vs. SfSNet In Figs. 6 and 8, we qualitatively compare our diffuse albedo estimates from our decomposition branch with SfSNet's. In Fig. 6, we show that our network can predict consistent albedo for the same individual across different illumination conditions, while SfSNet is less able to do so despite having been trained on the same data [2]. While not yet equal to the ground truth, our results are closer than baseline approaches. We reason that this is due to our network's ability to model the distribution of higher-frequency HDR illumination. In Fig. 8, we show the importance of modeling specularity. Without explicit specular handling, SfSNet tends to bake specular

[1] Pipeline published here. https://github.com/waps101/AlbedoMM/

[2] *Error in photogrammetry*: Some of FaceScape's GT diffuse albedo have illumination baked in (Fig. 6)

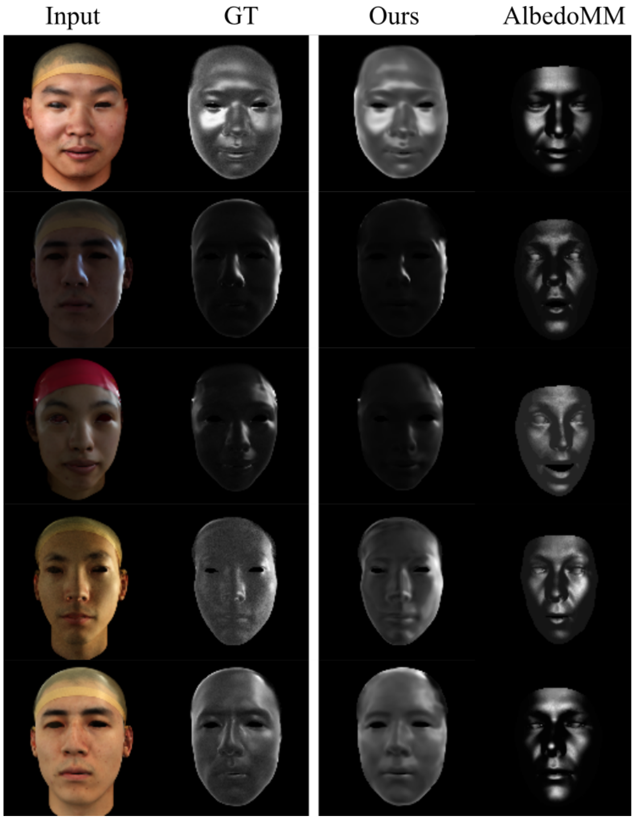

Figure 7. Our specular estimation compared to AlbedoMM (a 3D morphable model fitting method) shows more accurate reproduction of both sharp and broad highlights.

effects into the albedo layer, making them look more like the input images. Our approach does not do this as it explicitly reconstructs specular, which we will later show is important for editing and relighting tasks.

Diffuse shading estimation: Ours vs. SfSNet In Fig. 9, we compare our shading layer estimation to SfSNet on hand-picked illumination conditions where the light is causing self-occlusions. Our network's approach of using high frequency illumination is better able to construct more complex diffuse shading with self-occlusion effects, which is important for capturing realistic illumination. SfSNet's SH2 lighting assumption prevents this model from being able to capture these effects.

Specular estimation: Ours vs. AlbedoMM In Fig. 7, we compare our specular layer estimation to AlbedoMM, which is one of the only published specular models and is attempting to solve a 3D fitting problem. We show our model's ability to capture specular effects under varying illumination from a 2D image.

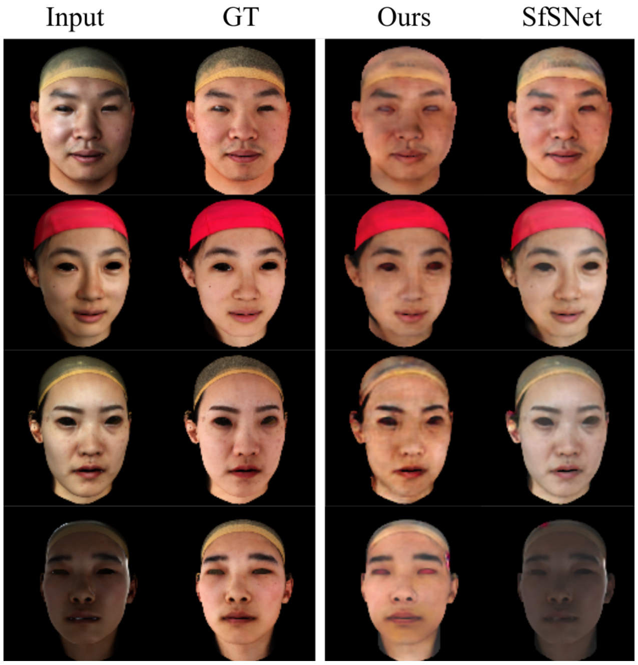

Figure 8. **Specular separation**: With its explicit specular handling, our model does not bake specular effects into the albedo. Without explicitly modeling specular, SfSNet tends to bake this appearance into the albedo image, causing it to look closer to the reconstruction.

Table 2. Our specular map estimates vs. AlbedoMM [29].

Methods	Specular		
	L1 ↓	*MSE* ↓	*LPIPS* ↓
AlbedoMM	0.2537	0.0866	1.1404
Ours	**0.0596**	**0.0071**	**0.2366**

5.1. Applications

Relighting: Ours vs. DIPR vs. SfSNet Relighting takes as input a portrait image and a target illumination; some approaches tackle this through decomposition and others attempt to more directly learn a relighting function [41]. We compare our results with DIPR [41] and SfSNet. Besides the limitation that both approaches use 2nd-order SH lighting, DIPR also assumes monochromatic lighting. Fig. 5 shows relighting under various target illuminations. Both DIPR and SfSNet fail to successfully unbake illumination effects: DIPR does not explicitly model the reflectance components and SfSNet has a Lambertian assumption. Our approach fares better.

Specular reflectance editing Finally, we show editing of the specular maps as another application of our approach in Fig. 11. Notice because of our image formation model, we are able to selectively edit the desired component and preserve all components we chose not alter.

Table 3. Decomposition quantitative evaluation performance by L1, MSE, and perceptual LPIPS [40] metrics. † SfSNet was re-trained on our data. With specular effects in the input images ('Reconstruction'), our approach is better than SfSNet. Without specular in the input ('Diffuse'), we see that our reconstruction quality is slightly improved thanks to more accurate shading estimation, though albedo estimates are slightly better for SfSNet. *Note:* We show LPIPS on Albedo and Shading for completeness, though this perceptual metric may be less meaningful on these less natural images.

Method	Reconstruction			Albedo			Shading			Diffuse		
	L1 ↓	*MSE* ↓	*LPIPS* ↓	*L1*	*MSE*	*LPIPS*	*L1*	*MSE*	*LPIPS*	*L1*	*MSE*	*LPIPS*
SFSNet†	0.0699	0.0178	0.1287	**0.0470**	**0.0080**	**0.0612**	0.099	0.0433	0.2819	0.0480	0.0089	0.0860
Ours	**0.0496**	**0.0099**	**0.0869**	0.0480	0.0085	0.0851	**0.0483**	**0.0193**	**0.2731**	**0.0417**	**0.0076**	**0.0705**

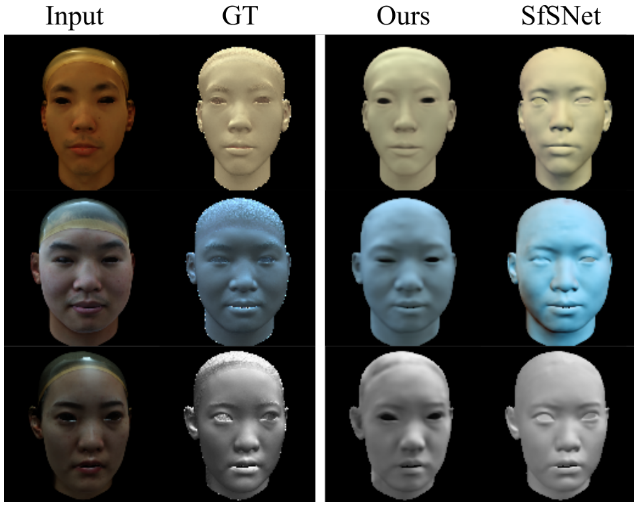

Figure 9. Modeling higher-frequency illumination than SH2 produces shading that is closer to the path-traced ground truth.

5.2. Additional results

We show additional results of our decomposition approach and compare it with SfSNet [28] and ground truth in Figs. 12 and 13 at the end of the paper. We also show an example on a real-world image of an unseen individual under unknown illumination taken from the FaceScape dataset (Fig. 15). To mitigate the synthetic-to-real domain gap, we choose a frontal view and scale the photo to approximately match our synthetic HDR image range.

6. Limitations

The FaceScape [38] dataset contains mostly Asian faces of limited skin tone variation, which restricts the ability of the priors learned from the data to be useful for other skin tones. Rather than build a practical system that one might deploy in the real world, our work only attempts to show academically that better synthetic data and image formation can improve face decomposition. Further, capturing albedo maps from human subjects is complex, and some of the data we use still has baked illumination components from shadowing in fine geometric detail.

Overcoming domain gaps is still a challenge. Chandran and Winberg et al. [4] propose a technique to project ray-traced faces to StyleGAN's latent space. This lets synthetic renderings retrain generated skin surface details, and let StyleGAN fill in missing details like eyes, inner mouth, hair, and background, all while respecting global scene illumination and camera pose. Such an approach allows supervised synthetic training to generalize to real world data.

Another limitation is missing subsurface scattering effects, which depend on the incoming illumination and the variable diffusivity of skin itself. While Blender supports subsurface effects, and while it is possible to map Christensen-Burley's d parameter (approximate scattering distance) to the scattering coefficient σ_s' and absorption coefficient σ_a captured by Weyrich et al. [35], we do not use Christensen-Burley's [6] approximation of subsurface scattering because a) there is a lack of captured priors from ETH/MERL, which makes realistic parameter selection difficult, and b) it is expensive to path trace computationally with low noise. We show an example of the difference between images with and without subsurface scattering (Fig. 14). Effects are typically most visible in thin regions of skin, such as the ears or nostrils.

Due to the path tracing, slight render noise is present in the training data. One side effect of using convolutional neural networks of limited capacity is that they learn this random high frequency patterning very late (if at all), and as such our output maps are not noisy.

Finally, our work estimates face decompositions. Moving to the more general case of portraits requires modeling geometric occlusion from external objects (world, hair, trees, etc.) which we cannot deal with, and other complex optical effects from face accessories like glasses. Future work should investigate how to combine differentiable path tracing with face modeling to capture refraction effects.

7. Conclusion

We present a method to decompose a single face image into physically-based channels that are useful for editing ap-

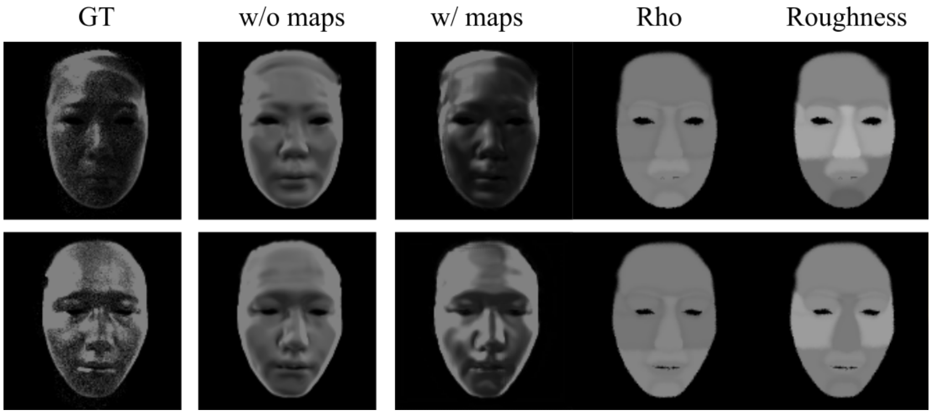

Figure 10. Predicting a specular layer with rho and roughness maps as input is beneficial. *w/o maps* indicates the specular branch was trained with only normal as input, while *w/ maps* indicates training with both rho and roughness map provided as input along with normal.

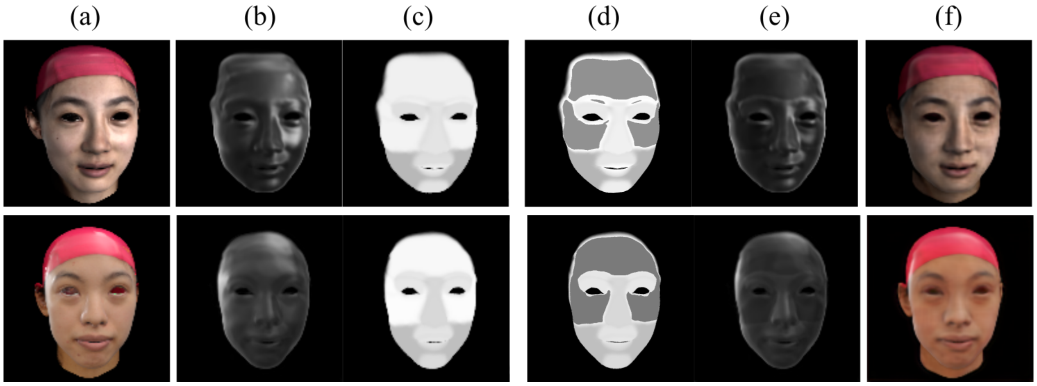

Figure 11. Left to right; **(a)** Input, **(b)** estimated specular layer and **(c)** its ρ map. **(d)** edited ρ map and **(e)** the estimated specular layer with **(d)** as input to specular branch. **(f)** Reconstruction using decomposition from **(a)** except replacing the specular layer with **(e)**.

plications. Our approach renders a more realistic dataset than previously available, and then uses supervised deep learning to encode priors that predict individual image formation components. We demonstrate that this approach is more successful than three recent methods with public codebases, particularly for specular reflections. Going forward, our work demonstrates the value of structuring deep learning frameworks around physically-based image formation models for more accurate reconstruction and editing.

8. Declarations

Consent for publication: We commit to the License Agreement of the FaceScape Dataset; all visages shown in our paper are from the stated publishable list.

Availability of data and materials: Image samples, code, and results released on https://github.com/brownvc/phaced.

Funding: We thank Brown University, Qian Zhang thanks an Andy van Dam PhD Fellowship, and we declare no competing interests.

Acknowledgments: We thank the reviewers for their detailed feedback. Qian Zhang thanks Kai Wang, Aaron Gokaslan, and Xianghao Xu for initial explorations. Finally, we all thank Chloe LeGendre, Henrique Weber, and Kalyan Sunkavalli for fruitful discussions.

References

[1] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 2003. 2

[2] A. Bousseau, S. Paris, and F. Durand. User-assisted intrinsic images. *ACM Transactions on Graphics (TOG)*, 28, 2009. 2

[3] D. A. Calian, J.-F. Lalonde, P. Gotardo, T. Simon, I. Matthews, and K. Mitchell. From faces to outdoor light probes. In *Computer Graphics Forum*, volume 37. Wiley Online Library, 2018. 3

[4] P. Chandran, S. Winberg, G. Zoss, J. Riviere, M. Gross, P. Gotardo, and D. Bradley. Rendering with style: combining traditional and neural approaches for high-quality face rendering. *ACM Transactions on Graphics (TOG)*, 40, 2021. 8

[5] A. Chen, Z. Chen, G. Zhang, K. Mitchell, and J. Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2

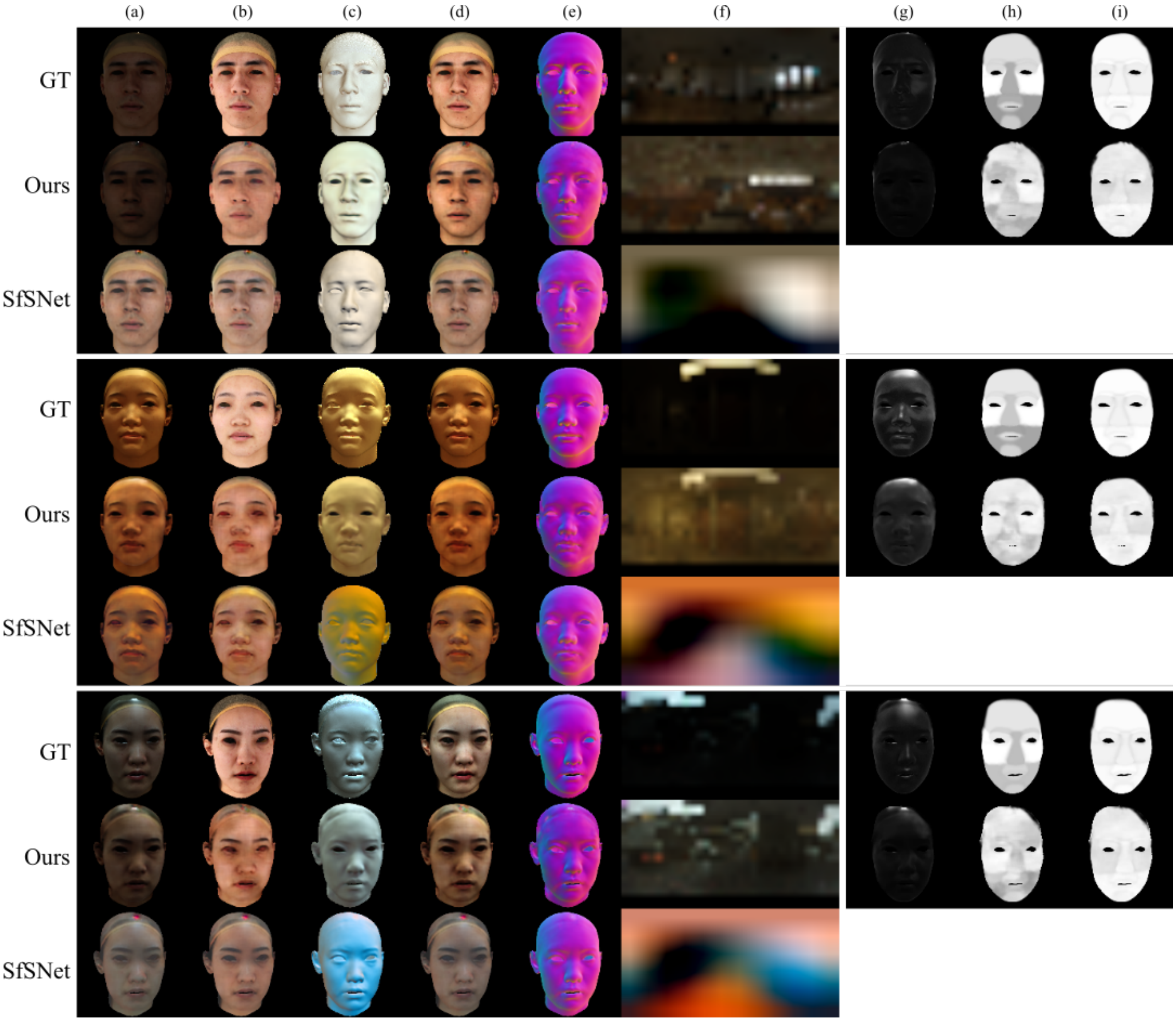

Figure 12. Comparisons of Ours and SfSNet (re-trained on our data) with Ground Truth (GT). Row **GT**: Column **(a)** is the Input image. Rows **Ours** and **SfSNet** left to right: **(a)** Reconstruction, **(b)** Albedo, **(c)** Shading, **(d)** Diffuse, **(e)** Normal, **(f)** Illumination, **(g)** Specular, **(h)** Rho Map, **(i)** Roughness Map. *Note:* Reconstructing the full range of radiances in the ground truth images is difficult, causing the ground truth images to have a larger HDR range. Due to subsequent tone mapping, the ground truth HDR environment maps look darker overall. The estimated environment maps have lower intensity range, and so appear more evenly exposed after tone mapping.

[6] P. H. Christensen. An approximate reflectance profile for efficient subsurface scattering. In *ACM SIGGRAPH 2015 Talks*. 2015. 8

[7] B. O. Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 3, 4

[8] P. Debevec. Image-based lighting. In *ACM SIGGRAPH 2006 Courses*. 2006. 2

[9] A. Dib, G. Bharaj, J. Ahn, C. Thebault, P. Gosselin, and L. Chevallier. Face reflectance and geometry modeling via differentiable ray tracing. *ArXiv*, abs/1910.05200, 2019. 2

[10] A. Dib, G. Bharaj, J. Ahn, C. Thébault, P. Gosselin, M. Romeo, and L. Chevallier. Practical face reconstruction via differentiable ray tracing. In *Computer Graphics Forum*, volume 40. Wiley Online Library, 2021. 2

[11] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J.-F. Lalonde. Learning to predict indoor illumination from a single image. *ACM Transactions on Graphics (TOG)*, 36, 2017. 4

[12] A. Hou, Z. Zhang, M. Sarkis, N. Bi, Y. Tong, and X. Liu. Towards high fidelity face relighting with realistic shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. 2, 5

[13] Y. Hu, B. Wang, and S. Lin. Fc 4: Fully convolutional color

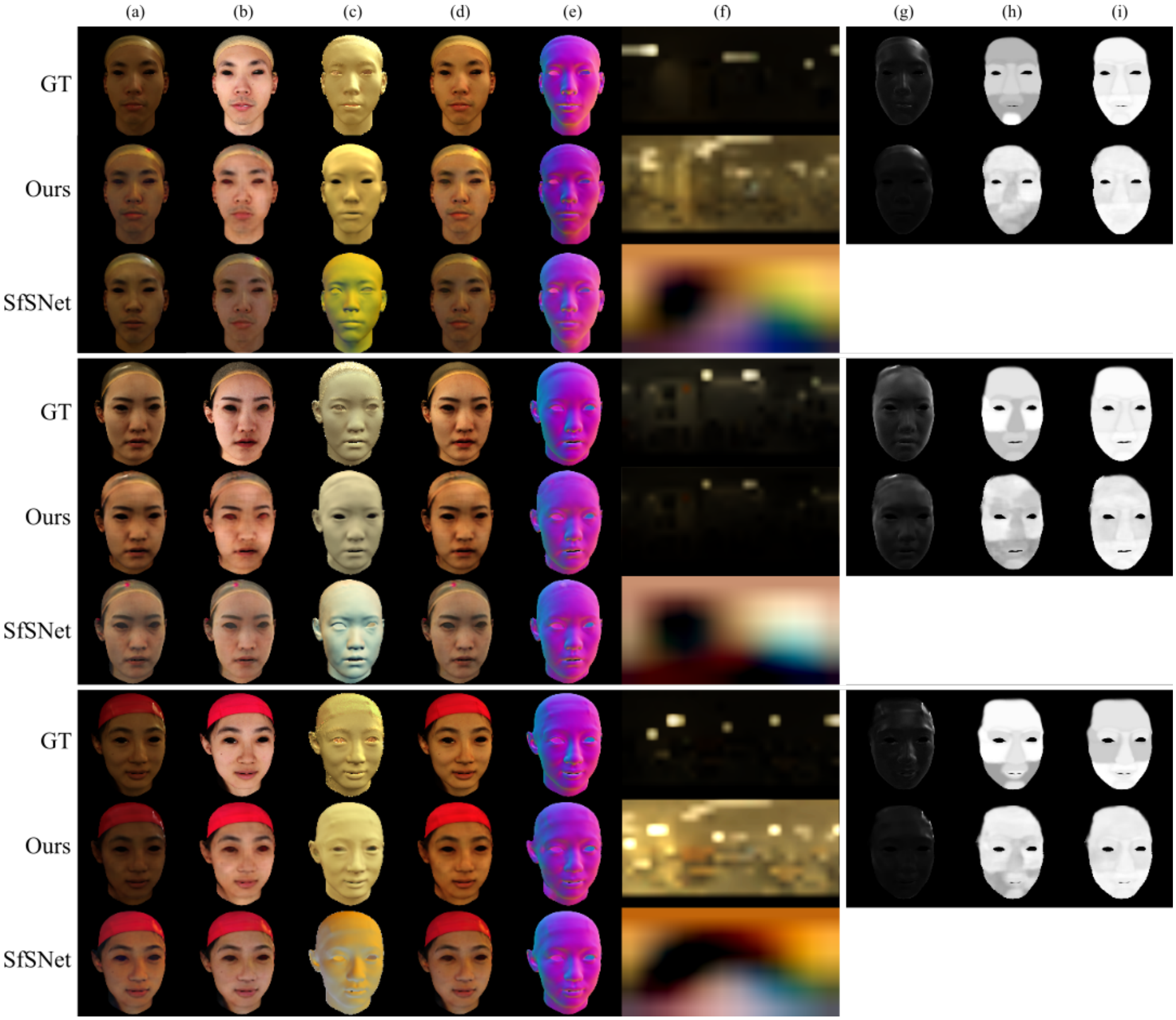

Figure 13. Comparisons of Ours and SfSNet (re-trained on our data) with Ground Truth (GT). Row **GT**: Column **(a)** is the Input image. Rows **Ours** and **SfSNet** left to right: **(a)** Reconstruction, **(b)** Albedo, **(c)** Shading, **(d)** Diffuse, **(e)** Normal, **(f)** Illumination, **(g)** Specular, **(h)** Rho Map, **(i)** Roughness Map. *Note:* Reconstructing the full range of radiances in the ground truth images is difficult, causing the ground truth images to have a larger HDR range. Due to subsequent tone mapping, the ground truth HDR environment maps look darker overall. The estimated environment maps have lower intensity range, and so appear more evenly exposed after tone mapping.

constancy with confidence-weighted pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 5

[14] W. Jakob. Mitsuba renderer, 2010. http://www.mitsuba-renderer.org. 4

[15] M. Janner, J. Wu, T. D. Kulkarni, I. Yildirim, and J. Tenenbaum. Self-supervised intrinsic image decomposition. In *Advances in Neural Information Processing Systems*, 2017. 2

[16] Y. Kanamori and Y. Endo. Relighting humans: occlusion-aware inverse rendering for full-body human images. *ACM Transactions on Graphics (TOG)*, 37, 2018. 2

[17] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019. 5

[18] E. H. Land and J. J. McCann. Lightness and retinex theory. *Josa*, 61, 1971. 2

[19] A. Lattas, S. Moschoglou, B. Gecer, S. Ploumpis, V. Triantafyllou, A. Ghosh, and S. Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction" in-the-wild". In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020. 2

[20] C. Li, K. Zhou, and S. Lin. Intrinsic face image decompo-

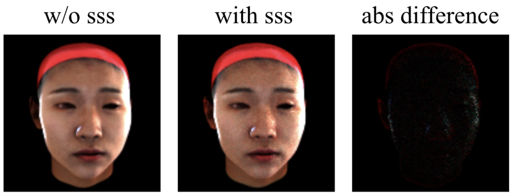

Figure 14. Subsurface scattering effect comparison. From left to right: rendered image with diffuse and specular effects; rendered image with image with added subsurface scattering effects; absolute difference image. In this environment with strong directional lighting, the difference is most significant on the left ear.

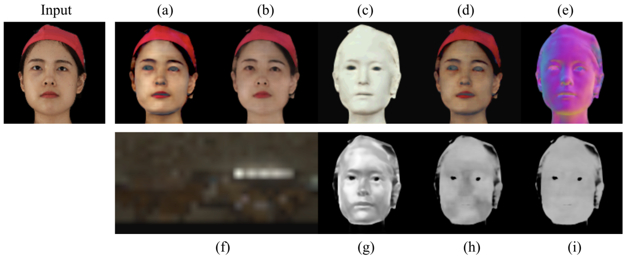

Figure 15. Results of real photograph from FaceScape [38] multiview dataset. We choose the frontal view and scale the photo to approximately match our synthetic HDR image range. Top left is the input photo (no ground truth). On the right side: **(a)** Reconstruction, **(b)** Albedo, **(c)** Shading, **(d)** Diffuse, **(e)** Normal, **(f)** Illumination, **(g)** Specular, **(h)** Rho Map, **(i)** Roughness Map.

sition with human face priors. In *European conference on computer vision*. Springer, 2014. 2

[21] Z. Li and N. Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 2

[22] B. Mallikarjun, A. Tewari, T.-H. Oh, T. Weyrich, B. Bickel, H.-P. Seidel, H. Pfister, W. Matusik, M. Elgharib, and C. Theobalt. Monocular reconstruction of neural face reflectance fields. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2021. 2

[23] T. Nestmeyer, J.-F. Lalonde, I. Matthews, and A. Lehrmann. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 2, 3, 4

[24] F. E. Nicodemus, J. C. Richmond, J. J. Hsia, I. W. Ginsberg, and T. Limperis. Geometrical considerations and nomenclature for reflectance. *Final Report National Bureau of Standards*, 1977. 2

[25] R. Ramamoorthi and P. Hanrahan. On the relationship between radiance and irradiance: determining the illumination from images of a convex lambertian object. *JOSA A*, 18, 2001. 2

[26] R. Ramamoorthi and P. Hanrahan. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*. ACM, 2001. 2

[27] E. Reinhard, M. Stark, P. Shirley, and J. Ferwerda. Photographic tone reproduction for digital images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, 2002. 4

[28] S. Sengupta, A. Kanazawa, C. D. Castillo, and D. W. Jacobs. Sfsnet: Learning shape, reflectance and illuminance of facesin the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 3, 4, 6, 8

[29] W. A. P. Smith, A. Seck, H. Dee, B. Tiddeman, J. Tenenbaum, and B. Egger. A morphable face albedo model. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. Fitting pipeline published at https://github.com/waps101/AlbedoMM/tree/master/scala. 2, 6, 7

[30] T. Sun, J. T. Barron, Y.-T. Tsai, Z. Xu, X. Yu, G. Fyffe, C. Rhemann, J. Busch, P. E. Debevec, and R. Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38, 2019. 2, 3, 4, 5

[31] K. E. Torrance and E. M. Sparrow. Theory for off-specular reflection from roughened surfaces. *Josa*, 57, 1967. 3

[32] B. Walter, S. R. Marschner, H. Li, and K. E. Torrance. Microfacet models for refraction through rough surfaces. *Rendering techniques*, 2007. 3

[33] Z. Wang, X. Yu, M. Lu, Q. Wang, C. Qian, and F. Xu. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics (TOG)*, 39, 2020. 2, 3, 4, 5

[34] H. Weber, D. Prévost, and J.-F. Lalonde. Learning to estimate indoor lighting from 3d objects. In *2018 International Conference on 3D Vision (3DV)*. IEEE, 2018. 5, 6

[35] T. Weyrich, W. Matusik, H. Pfister, B. Bickel, C. Donner, C. Tu, J. McAndless, J. Lee, A. Ngan, H. W. Jensen, et al. Analysis of human faces using a measurement-based skin reflectance model. In *ACM Transactions on Graphics (TOG)*, volume 25. ACM, 2006. 2, 3, 8

[36] Y. Wu and K. He. Group normalization. In *Proceedings of the European conference on computer vision (ECCV)*, 2018. 4

[37] S. Yamaguchi, S. Saito, K. Nagano, Y. Zhao, W. Chen, K. Olszewski, S. Morishima, and H. Li. High-fidelity facial reflectance and geometry inference from an unconstrained image. *ACM Transactions on Graphics (TOG)*, 37, 2018. 2

[38] H. Yang, H. Zhu, Y. Wang, M. Huang, Q. Shen, R. Yang, and X. Cao. Facescape: a large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 3, 6, 8, 12

[39] R. Yi, C. Zhu, P. Tan, and S. Lin. Faces as lighting probes via unsupervised deep highlight extraction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 3

[40] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018. 6, 8

[41] H. Zhou, S. Hadap, K. Sunkavalli, and D. W. Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019. 2, 3, 4, 6, 7