Single-Shot Facial Appearance Acquisition without Statistical Appearance Priors

G. Y. Soh and A. Ghosh

Imperial College London, UK

Abstract

Single-shot in-the-wild facial reflectance acquisition has been a long-standing challenge in the field of computer graphics and computer vision. Current state-of-the-art methods are typically learning-based methods, pre-trained on a dataset of facial reflectance data. However, due to the high cost and time-consuming nature of gathering these datasets, they are usually limited in the number of subjects covered and hence are prone to biases in the dataset. To this end, we propose a novel multi-stage guided optimization with differentiable rendering to tackle this problem, without the use of statistical facial appearance priors. This makes our method immune to these biases, and we demonstrate the advantage with qualitative and quantitative evaluations against current state-of-the-art methods.

CCS Concepts

• ***Computing methodologies*** → ***Reflectance modeling;***

1. Introduction

With the introduction of facial reflectance datasets, single-view facial reflectance inference has become feasible using statistical priors and deep neural networks [LMP*21, PLMZ23]. However, a key limitation is the absence of large-scale facial appearance datasets, due to the expensive and difficult process of hiring people to a studio for capture. These methods are typically trained only a small handful of real captured subjects (less than 1000) and might be subject to bias or not represent the full human population well.

In contrast to deep-learning-based methods, optimization methods are analysis-by-synthesis approaches that aim to reconstruct the reflectance by synthesizing rendered images and comparing them with the reference image. Closest to our work is Dib et. al. [DBA*21], where differentiable rendering is used to fit a 3DMM with statistical reflectance priors onto the subject. However, we find that the reliance on low-resolution statistical priors causes their reconstruction to be blurred, and the authors also mention a bias towards Caucasian skin tones [DBA*21].

Hence, in our work, we focus on the task of obtaining the facial reflectance of a subject from a single image, without the use of any statistical facial appearance priors. We propose a new multi-staged optimization pipeline with differentiable rendering, that gradually guides the optimization towards a plausible reconstruction of the subject's facial appearance.

2. Methodology

An overview of our method can be seen in Fig. 1.

2.1. 3DMM Fitting

We first fit the face with a 3DMM model to reconstruct a mesh $\mathbf{S} \in \mathbb{R}^{3\times n}$ of the subject's face from an affine model:

$$\mathbf{S}(\mathbf{k}_{id}, \mathbf{k}_{exp}) = \bar{\mathbf{S}} + \mathbf{B}_{id}\mathbf{k}_{id} + \mathbf{B}_{exp}\mathbf{k}_{exp}, \quad (1)$$

To fit the model, we directly utilized an off-the-shelf pre-trained CNN [DYX*19] to predict these coefficients from a single image.

2.2. Reflectance

We model the BRDF with two separable components: one diffuse component representing the body reflectance of skin, and one specular lobe representing the specular reflection at the surface of the skin due to the stratum corneum. The diffuse component is modelled with a basic Lambertian model with diffuse albedo ρ_d, while the specular component is modelled with a specular albedo ρ_s and the Beckmann microfacet BRDF. To constrain our optimization, we set the index of refraction uniformly at $\eta = 1.38$, with specular roughness α defined on a per-region basis following the analysis of skin reflectance done by Weyrich et al. [WMP*06].

2.3. Illumination

We apply the distant illumination assumption and model the illumination as a distant environment map via a mixture of Spherical Gaussians (SGs). The final environment map is a sum of $N = 648$ monochrome Spherical Gaussians, which is rasterized to a 180×360 resolution bitmap.

DOI: 10.2312/egs.20251035

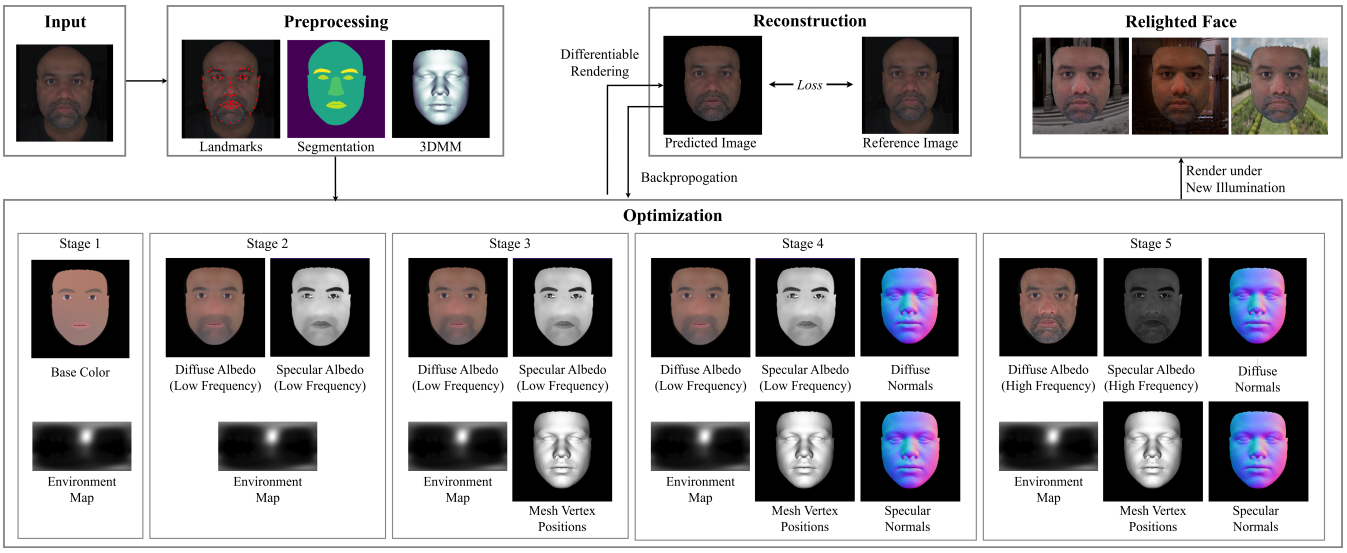

Figure 1: *An overview of our methodology*

3. Optimization Strategy

Directly recovering the shape, reflectance, and illumination of a single image is an extremely ill-posed problem due to the complex and non-convex loss landscape, and it is easy to fall into a local minima with a naive approach. Therefore, we introduce a multi-stage optimization strategy, that guides the optimization to a good solution by first heavily constraining the optimization process and iteratively relaxing the constraints until a high-fidelity reconstruction is reached.

3.1. Stage 1: Base Color & Illumination Optimization

The main goal of the first stage of optimization is to estimate the illumination of the scene by supervising the effects of the shading from the illumination on the face. We achieve this by first segmenting the face into the skin, lip, eye, and eyebrow regions using an off-the-shelf face parser [LSWP21], and optimize a uniform per region specular and diffuse albedo ρ_s^m, ρ_d^m, where $m \in \{\text{skin}, \text{lip}, \text{eye}, \text{eyebrow}\}$. The spherical Gaussian parameters ξ, λ, μ are jointly optimized at this stage as well. The purpose of the per region albedo instead of an albedo texture is to ensure that no illumination is baked into the albedo by limiting its expressivity. The optimizer is instead required to optimise the environment map to minimize $\mathcal{L}_{\text{image}}$, the L1 per-pixel image loss between the rendered and reference images.

3.2. Stage 2: Low-Frequency Albedo Optimization

The second stage of optimization focuses on the low-frequency albedo variations on the subject's face. Formally, we introduce a specular albedo map T^s_{alb} and diffuse albedo map T^d_{alb} to be optimized at 512×512 resolution. However, to prevent overfitting of the albedo maps, a smoothness regularizer is applied per texel τ in texture space.

$$\mathcal{L}_{\text{smooth}}(T) = \sum_{i \in \mathcal{H}} \sum_{j \in \mathcal{W}} |\tau_{i,j} - \tau_{i+1,j}| + |\tau_{i,j} - \tau_{i,j+1}|, \tag{2}$$

3.3. Stage 3: Shape Optimization

In the third stage, to improve shape reconstruction quality and mesh registrations for better shading cue removal, we optimize the mesh on a per-vertex basis using the gradient preconditioning technique introduced in [NJJ21]. To help guide the optimization and prevent getting stuck in local optima, a weak keypoint loss is introduced by minimizing the L2 norm between detected facial landmarks $\mathbf{p}$ with off-the-shelf landmark detectors and the 2D projection $\hat{\mathbf{p}}$ of a set of predefined 3D keypoints defined on the 3DMM. Additionally, to prevent implausible solutions for the mesh, we apply an L2 regularizer on the norm of each vertex position $\hat{\mathbf{s}}$ from their initial position on the 3DMM mesh $\mathbf{s}$:

$$\mathcal{L}_{\text{landmark}} = \sum_{i=0}^{N_p} \|\hat{\mathbf{p}}_i - \mathbf{p}_i\|_2^2, \quad \mathcal{L}_{\text{shape}} = \sum_{i=0}^{N_s} \|\hat{\mathbf{s}}_i - \mathbf{s}_i\|_2^2 \tag{3}$$

We directly optimize the texture in the image space and project it to the mesh at each iteration, instead of relying on the UV coordinates, such that both shape and the textures can be independently optimized without negatively impacting $\mathcal{L}_{\text{image}}$. Fig. 2 demonstrates the effect of the mesh optimization.

3.4. Stage 4: Normal Optimization

Stage 4 aims to recover the high-frequency specular and diffuse normals commonly used for high-fidelity photorealistic rendering. We introduce a new texture $T^s_{\mathbf{n}} \in \mathbb{R}^{\mathcal{H} \times \mathcal{W} \times 3}$ to model the specular normal perturbation vector $\epsilon^{\mathbf{s}}_{\mathbf{n}} \in \mathbb{R}^3$. The specular normal $\hat{\mathbf{n}}_{\mathbf{s}}$ is obtained by normalizing the perturbed normal:

$$\hat{\mathbf{n}}_{\mathbf{s}} = \frac{\mathbf{n} + \epsilon^{\mathbf{s}}_{\mathbf{n}}}{\|\mathbf{n} + \epsilon^{\mathbf{s}}_{\mathbf{n}}\|}, \quad \mathcal{L}_{\text{normal}} = \sum_{i \in T^s_{\mathbf{n}}} \|\epsilon^{\mathbf{s}}_{\mathbf{n}_i}\|_2^2 \tag{4}$$

where $\mathbf{n}$ is the mesh normal. The diffuse normals are obtained by blurring the perturbation with a Gaussian kernel with standard deviation σ set to 1. We further introduce an additional term to regularize the normal perturbations to prevent improbable solutions.

3.5. Stage 5: High-Frequency Albedo Optimization

Finally, we recover the high-frequency albedo variations of the subject to recover high-fidelity details of the subject. This is obtained by lowering the weight of $\mathcal{L}_{\text{smooth}}$, and refining the albedo maps.

4. Implementation Details

Our final loss function is a weighted sum of each individual loss component:

$$\begin{aligned}\mathcal{L}_{\text{total}} = {} & \lambda_{\text{image}}\mathcal{L}_{\text{image}} + \lambda^{s}_{\text{smooth}}\mathcal{L}_{\text{smooth}}(T^{s}_{\text{alb}}) \\ & + \lambda^{d}_{\text{smooth}}\mathcal{L}_{\text{smooth}}(T^{d}_{\text{alb}}) + \lambda_{\text{landmark}}\mathcal{L}_{\text{landmark}} \\ & + \lambda_{\text{shape}}\mathcal{L}_{\text{shape}} + \lambda_{\text{normal}}\mathcal{L}_{\text{normal}}\end{aligned} \tag{5}$$

The key to our method's success is by modulating the loss weights involved in each optimization stage. The details for each stage are presented in Table 1. We implemented our methodology with the physically-based differentiable render Mitsuba 3 and utilized the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For additional optimization stability, gradient clipping is used to limit the gradients in the range $[-1,1]$ on a per-element basis. The whole process takes about 20 minutes to complete on an RTX 4090 GPU.

Table 1: *Hyperparameter values per stage. Here N represents the number of iterations, λ are the loss weights, and α_i are the learning rate for each parameter i. The rest of the learning rates are set at a constant $\alpha = 0.05$ throughout the optimization process.*

	Stage 1	Stage 2	Stage 3	Stage 4	Stage 5
N	2000	2000	750	500	100
λ_{image}	1.0	1.0	1.0	1.0	1.0
$\lambda^{s}_{\text{smooth}}$	-	10.0	10.0	10.0	1.0
$\lambda^{d}_{\text{smooth}}$	-	1.0	1.0	1.0	0.1
$\lambda_{\text{landmark}}$	-	-	1.0	1.0	1.0
λ_{shape}	-	-	50.0	50.0	50.0
λ_{normal}	-	-	-	0.1	0.1
α_{ξ}	0.01	0.01	0.005	0.0	0.0
α_{λ}	0.05	0.05	0.01	0.0	0.0
α_{μ}	0.05	0.05	0.01	0.0	0.0

5. Results

Our method is able to inverse render accurate reflectance maps from in-the-wild images for use in photorealistic rendering. To demonstrate this, we compare our method's results with other facial appearance techniques in the following sections.

5.1. Comparison with State-of-the-art Models

In Fig. 3 and Fig. 4 we provide a qualitative comparison with Relightify [PLMZ23] and MoSAR [DHG*24] respectively, These methods are the state-of-the-art deep-learning methods for 3D reflectance reconstruction from a single image. Our method produces competitive results, and due to the lack of appearance priors, are able to better preserve personalized details such as makeup and skin tone. Respective results are obtained from the original papers.

Moreover, in Fig. 5 and Table 2 we provide quantitative results on Digital Emily [ARL*10]. We compare our results by relighting the appearance maps obtained under different illuminations. Note that Rainer et. al. [RBG23] relies on multiple images, whereas the rest rely on a single image. We also demonstrate the efficacy of our multi-stage pipeline by performing ablation studies in Table 3.

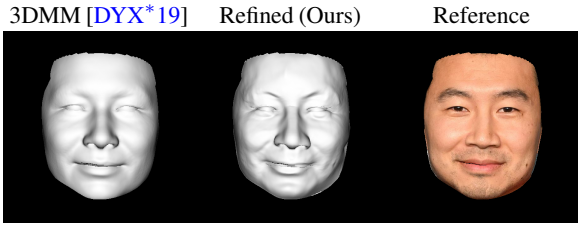

Figure 2: *Results of the face mesh recovered. Our reconstructed mesh fits our target better than the standard 3DMM prediction.*

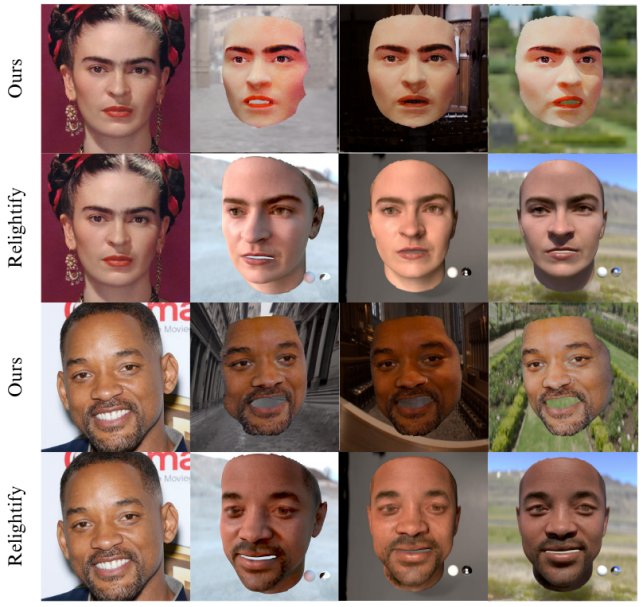

Figure 3: *Qualitative comparison with Relightify [PLMZ23]*

Figure 4: *Qualitative comparison with MoSAR [DHG*24].*

5.2. Comparison with Active Illumination Captures

In Fig. 6, we compare the reflectance maps recovered by our method with the maps recovered by a multi-view, desktop-based, active illumination setup by Lattas et. al. [LLK*22]. While their method captures higher resolution maps from multiple cameras,

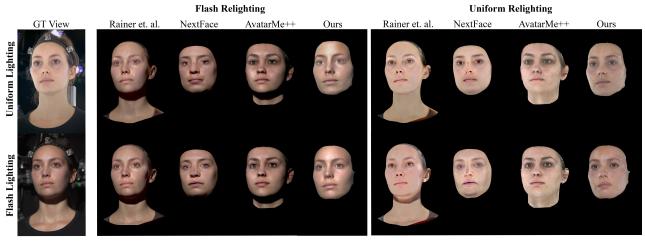

Figure 5: *Comparisons with Digital Emily [ARL*10].*

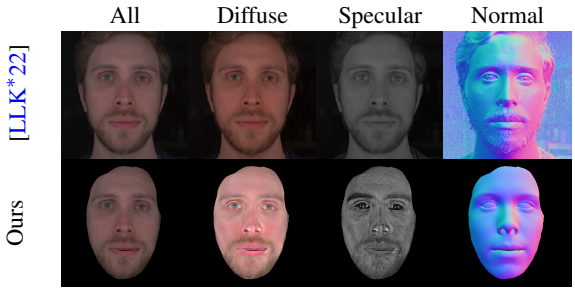

Figure 6: *Comparisons of the reflectance maps recovered by our single-view method and the multi-view, active-illumination setup by [LLK*22].*

Table 2: *PSNR and LPIPS for Fig. 5 adapted from [RBG23].*

Input	Method	Flash Relight		Uniform Relight	
		PSNR ↑	LPIPS ↓	PSNR ↑	LPIPS ↓
Uniform	[RBG23]	**25.3**	**0.11**	23.5	0.081
	[DBA*21]	19.7	0.15	15.9	0.13
	[LMP*21]	21.4	0.13	21.2	0.11
	Ours	23.4	0.12	**25.9**	**0.077**
Flash	[RBG23]	26.4	0.085	22.6	0.11
	[DBA*21]	19.6	0.16	15.6	0.13
	[LMP*21]	21.8	0.13	21.3	0.12
	Ours	**28.1**	**0.065**	**24.1**	**0.071**

Table 3: *Ablation experiment on Digital Emily [ARL*10] by disabling each step in the final pipeline. Here U → F denotes Uniform Input with Flash Relighting and vice versa, similar to Fig. 5.*

	Metric	Full	Without Stage				
			1	2	3	4	5
U→F	PSNR ↑	23.4	16.2	22.3	21.9	23.2	22.1
	LPIPS ↓	0.12	0.22	0.14	0.17	0.12	0.15
F→U	PSNR ↑	24.1	17.0	23.2	22.8	23.9	23.1
	LPIPS ↓	0.07	0.17	0.08	0.15	0.08	0.09

our method manages to recover good reflectance maps despite only supervising on a single view, under unknown lighting conditions. Additionally, we observe that our reflectance maps contain less baked-in illumination, as opposed to [LLK*22], which do not account for ambient occlusion and inter-reflections.

6. Conclusions

We have presented a novel method for single-shot facial reflectance reconstruction that, unlike previous works, do not rely on any statistical appearance priors. Our approach came about due to the high costs and inaccessibility of facial reflectance datasets, which could introduce bias in the reconstruction. To this end, our prior free method manages to preserve the identity while capturing personalized albedo such as makeup or lipstick. However due to the inherent lack of priors, our method has difficulty resolving ambiguities illumination and albedo, and is only limited to well-lit images with monochromatic lighting. Also, our model struggles with external shadows and complex illumination as well. Nevertheless, we hope that our work inspires further prior-free methods for facial appearance capture and other related applications, particularly with guided optimization processes.

References

[ARL*10] ALEXANDER O., ROGERS M., LAMBETH W., CHIANG J.-Y., MA W.-C., WANG C.-C., DEBEVEC P.: The digital emily project: Achieving a photorealistic digital actor. *IEEE Computer Graphics and Applications 30*, 4 (2010), 20–31. 3, 4

[DBA*21] DIB A., BHARAJ G., AHN J., THÉBAULT C., GOSSELIN P., ROMEO M., CHEVALLIER L.: Practical face reconstruction via differentiable ray tracing. In *Computer Graphics Forum* (2021), vol. 40, Wiley Online Library, pp. 153–164. 1, 4

[DHG*24] DIB A., HAFEMANN L. G., GOT E., ANDERSON T., FADAEINEJAD A., CRUZ R. M., CARBONNEAU M.-A.: Mosar: Monocular semi-supervised model for avatar reconstruction using differentiable shading. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2024), pp. 1770–1780. 3

[DYX*19] DENG Y., YANG J., XU S., CHEN D., JIA Y., TONG X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops* (2019), pp. 0–0. 1, 3

[LLK*22] LATTAS A., LIN Y., KANNAN J., OZTURK E., FILIPI L., GUARNERA G. C., CHAWLA G., GHOSH A.: Desktop-based high-quality facial capture for everyone. In *ACM SIGGRAPH 2022 Talks* (New York, NY, USA, July 2022), SIGGRAPH '22, Association for Computing Machinery, p. 1–2. 3, 4

[LMP*21] LATTAS A., MOSCHOGLOU S., PLOUMPIS S., GECER B., GHOSH A., ZAFEIRIOU S.: Avatarme++: Facial shape and brdf inference with photorealistic rendering-aware gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence 44*, 12 (2021), 9269–9284. 1, 4

[LSWP21] LIN Y., SHEN J., WANG Y., PANTIC M.: Roi tanh-polar transformer network for face parsing in the wild. *Image and Vision Computing 112* (2021), 104190. 2

[NJJ21] NICOLET B., JACOBSON A., JAKOB W.: Large steps in inverse rendering of geometry. *ACM Transactions on Graphics (TOG) 40*, 6 (2021), 1–13. 2

[PLMZ23] PAPANTONIOU F. P., LATTAS A., MOSCHOGLOU S., ZAFEIRIOU S.: Relightify: Relightable 3d faces from a single image via diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2023), pp. 8806–8817. 1, 3

[RBG23] RAINER G., BRIDGEMAN L., GHOSH A.: Neural shading fields for efficient facial inverse rendering. Accepted: 2023-10-09T07:34:18Z. 3, 4

[WMP*06] 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. *ACM Transactions on Graphics (ToG) 25*, 3 (2006), 1013–1024. 1