

S2F2: Self-Supervised High Fidelity Face Reconstruction from Monocular Image

Abdallah Dib* Junghyun Ahn* Cédric Thébault Philippe-Henri Gosselin Louis Chevallier

InterDigital R&I

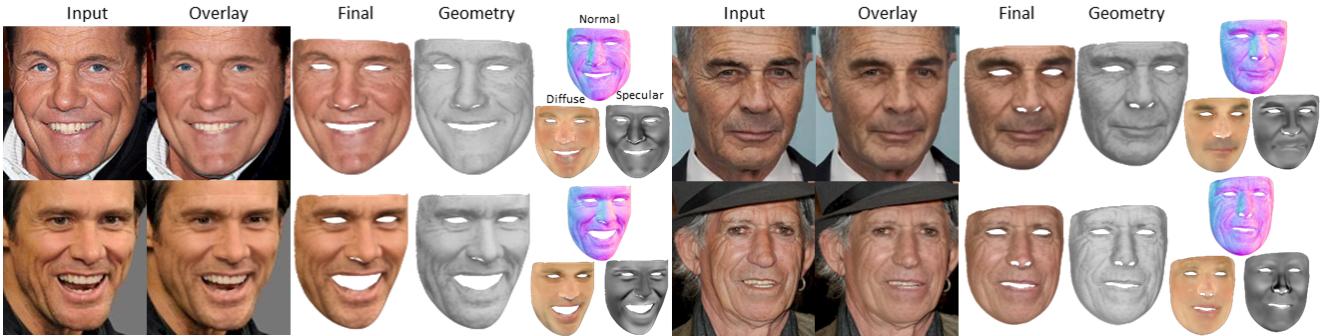


Figure 1. Given a single image, our method achieves appealing 3D face reconstruction and estimates a dense detailed face geometry, spatially varying face reflectance (diffuse and specular albedos) and high frequency scene illumination.

Abstract

We present a novel face reconstruction method capable of reconstructing detailed face geometry, spatially varying face reflectance from a single monocular image. We build our work upon the recent advances of DNN-based auto-encoders with differentiable ray tracing image formation, trained in self-supervised manner. While providing the advantage of learning-based approaches and real-time reconstruction, the latter methods lacked fidelity. In this work, we achieve, for the first time, high fidelity face reconstruction using self-supervised learning only. Our novel coarse-to-fine deep architecture allows us to solve the challenging problem of decoupling face reflectance from geometry using a single image, at high computational speed. Compared to state-of-the-art methods, our method achieves more visually appealing reconstruction.

1. Introduction

Fast, robust and high fidelity 3D face reconstruction has a wide range of applications in many domains such as interactive face editing, video-conferencing, XR, Metaverse applications, and visual effects for movies post-production. Several approaches such as [1, 2, 3, 4, 5, 6] achieve high fi-

dentity reconstruction, but require complex hardware setups (multi-view, lightstage). They are therefore not easily usable in most of the aforementioned applications. Significant progress was made to achieve high quality reconstruction from monocular image/video using optimization-based frameworks. Such methods [7, 8, 9, 10] are generally slow, of limited robustness and not suitable for interactive scenarios. Also, their performances in challenging conditions (non-uniform lighting, extreme poses) are limited.

Deep-based analysis-by-synthesis approaches have been investigated to leverage the generalization capabilities of machine learning. However, these methods [11, 12, 13, 14] generally sacrifice reconstruction quality. Methods combining CNNs with differentiable rendering trained in a self-supervised manner have been introduced by Tewari *et al.* [11, 12, 15]. These methods directly regress the parameters of a statistical morphable model and achieve real-time performance but fall short of the quality and fidelity because their estimated geometry and reflectance are bound by the statistical prior space which limits its generalization regarding the real diversity of face geometry and reflectance.

More recently, many works aim to improve the realism and fidelity of deep-based methods by capturing either detailed geometry or reflectance but not both, which we discuss next. First, to capture detailed geometry, and because of the complexity of the problem, several methods rely on ground truth dataset obtained either from multi-view recon-

*Equal contribution

struction setup (and/or lightstage) [16, 17, 18, 19], from synthetic data [20, 21] or from a mixture of both [22, 23]. Feng *et al.* [24] is the only self-supervised method that captures detailed geometry. However, this method only captures medium-scale geometry details and misses high-frequency geometry variations. Additionally, its estimated reflectance is restricted to the statistical prior space which limits its generalization regarding the real diversity in face geometry and reflectance. Second, and to improve the reflectance, Dib *et al.* [25] combined ray tracing and self-supervised learning to capture medium-scale reflectance details. However, this method restricts the estimated geometry to a parametric face model preventing high-frequency facial details (such as wrinkles, folds...) to be captured. To our knowledge, there is no existing self-supervised methods that can jointly estimate detailed geometry and reflectance.

The first contribution of this work, is the introduction of the first self-supervised method that jointly estimate detailed geometry and reflectance. This is accomplished via our novel coarse-to-fine architecture, with an adapted training strategy which allows our method to efficiently solve the ambiguous and complex problem of separating detailed geometry from reflectance from a single image taken under uncontrolled lighting conditions.

The second contribution, is the combination, for the first time, of differentiable ray tracing with vertex-based renderer at training time to overcome the problem of edge discontinuities of the ray tracing. This allow our method to benefit from both renderers. On one hand, ray tracing accurately models self-shadows and on the second hand, the vertex-based renderer evaluates correctly the whole geometry including boundaries. This leads to a significant improvement in the estimated geometry compared to Dib *et al.* [25] that uses only ray tracing.

Finally, the aforementioned contributions enable to take a big leap forward in reconstruction quality for self-supervised methods and lead to superior face reconstruction when compared to recent state-of-the art methods. To our knowledge, this is the first time a self-supervised method reaches this level of fidelity and realism. Our robust face attributes estimation (diffuse, specular and normal) leads to practical applications such as face attribute editing and relighting.

2. Related works

Methods such as [1, 3, 4, 5] deliver high-fidelity face reconstruction from multi-view or light-stage setup, but they are generally expensive and not applicable for in-the-wild conditions (many cameras, specific lighting). In this work, we are interested in face reconstruction/tracking methods that only use image or video as input and do not require any external hardware setup beyond the camera. These methods can be split into two categories: optimization-based and

learning-based approaches.

Geometry and reflectance modeling Statistical 3D Morphable Models - 3DMM - is the main building block for a wide range of optimization-based and learning-based methods [26, 27, 28, 29]. This statistical model adds a lot of structure and priors to face reconstruction problem from monocular image or video and makes it tractable. However, due to the low-dimensional space of 3DMM, subject specific medium and high frequency geometry and albedo details cannot be modeled. Additionally, the skin reflectance model of 3DMM can only model the diffuse albedo and may bake shadows/specularity in the albedo. [30] proposes a drop-in replacement for the basic lambertian reflectance model of 3DMM incorporating a diffuse and specular priors. In this work, we base our reconstruction on the 3DMM geometry with the statistical diffuse and specular prior of [30] and we train a novel multi-stage deep network to capture fine diffuse and geometry details.

Most of optimization-based methods like [9, 7, 8, 31, 32, 10, 33, 34] rely on the same 3DMM parametrization, they provide generally precise reconstruction at the expense of a high computation cost and are sensitive to difficult lighting conditions.

Among the learning-based methods, deep convolution neural networks (CNN) are effective at direct face reconstruction [13, 14, 11, 12, 15, 35, 21, 36, 37, 38, 39, 40, 41]. Tewari *et al.* [11] proposed the first self-supervised autoencoder-like method to estimate face attributes based on 3DMM. The advantage of these self-supervised methods is that they can be trained on large corpus of unlabeled images. However they generally fall short of reconstruction precision because of their simplified underlying scene parameterization (pure-Lambertian BRDF to model skin reflectance and low-order spherical harmonics to model light). Their inability to model self-shadows is also a possible reason for their instability under challenging lighting conditions. Dib *et al.* [25] proposes a self-supervised method that significantly improves over these methods and solves many of these limitations. For instance, it uses a cook-torrance BRDF to model skin reflectance, and captures personalized albedos outside the statistical prior space. It also uses a differentiable ray tracing to model self-shadows. However the reconstructed geometry of their method is still limited by 3DMM space.

Detailed geometry reconstruction Capturing fine detailed geometry on top of global face structure is a pre-condition to achieve high fidelity face reconstruction. However, because of the complexity of the problem, methods such as [6] uses specific hardware setup which is not applicable in-the-wild. Others, such Cao *et al.* [42] relies on ground truth dense data [1], or data captured using a lightstage (or multi-view) such as [19, 43, 44, 16, 17, 18]. However, acquiring

such ground truth data is not always possible.

Optimization based methods like [45, 7, 9] use shape-from-shading [46] to capture fine geometry details. However these methods may not generalize well and are computationally expensive.

Some deep-based methods rely partially on synthetic data ([20, 47, 22]) or a mix of labeled and unlabeled data [23] to capture fine detailed geometry. The bias introduced by these methods may impede the resulting precision due to the mismatch with real data distribution. They also do not estimate face reflectance. Sengupta *et al.* [21] uses synthetic data to train their network which estimates a normal map and skin reflectance (limited to pure-lambertian BRDF) but does not capture high frequency geometry details. More recently, Feng *et al.* [24] learns an ‘expression-dependent’ displacement model in-the-wild and is the only method that relies only on unlabeled images for end-to-end training. However this method only captures medium-frequency displacement map and their estimated reflectance is restricted to the statistical albedo prior space which limits their reconstruction quality. To our knowledge, our method is the first self-supervised method that jointly estimates: geometry at high frequency, spatially varying personalized skin reflectance with diffuse and specular albedos and high frequency illumination from a single monocular image.

Differentiable rendering Differentiable rendering is a key block in the context of analysis-by-rendering and several implementations exist. Tewari *et al.* [11] proposed an efficient vertex-based differentiable rendering that can only handle pure Lambertian BRDF and cannot capture self-shadows. Works such as [48] propose a differentiable shadow computation method for this type of renderer.

Dib *et al.* [49, 10] introduced a method which uses differentiable ray tracing for face reconstruction within a classic optimization framework. The key advantage of ray tracing over vertex-based renderer is the capacity of ray tracing to handle self-shadows where a visibility mask is calculated for each surface point with respect to each light during direct illumination pass. However, differentiable ray tracing is computationally expensive and memory consuming. Recently, Dib *et al.* [25] uses differentiable ray tracing in conjunction with a deep neural architecture for direct face reconstruction. In this scheme, inference does not incur the speed penalty of ray tracing and delivers near real-time performance with robust reconstruction in challenging lighting conditions. Regarding the optimization process, a limitation of ray tracing is the noise on gradients originating at the objects boundaries as they are sampled by very few points. Solutions exist but remain expensive ([50, 51]). In this work, we combine a vertex-based renderer with a ray tracing renderer together with a deep neural architecture that takes advantage of both. On one hand, ray tracing accurately models self-shadows, and on the second hand, the vertex-based

renderer evaluates the whole geometry including the edges.

3. Method

Our goal is to obtain a high fidelity face reconstruction with faithful separation between reflectance and geometry attributes. To solve this ill-posed problem, we propose a novel multi-stage deep architecture, wherein different stages progressively refine the reconstruction.

Our network architecture, depicted in Figure 2, is composed of 3 stages denoted: ‘Coarse’, ‘Medium’ and ‘Fine’. The ‘Coarse’ reconstruction relies on the statistical geometry and albedo priors space. This base reconstruction lacks important geometry and albedo (diffuse and specular) details because it is restricted by the low dimensional space of the underlying model. The ‘Medium’ stage improves the previously estimated albedos without the limitations of the statistical prior. Finally, the ‘Fine’ stage adds diffuse albedo and fine geometry details.

In the next, we introduce the scene attributes used by our formulation and we describe the network architecture.

3.1. Scene attributes

Face geometry Shape identity is modeled using [26, 34]’s statistical face model, and is given by $e = a_s + \Sigma_s \alpha$. e is a vector of face geometry vertices with N vertices. The identity-shape space is spanned by $\Sigma_s \in \mathbb{R}^{3N \times K_s}$ composed of $K_s = 80$ principal components of this space. $\alpha \in \mathbb{R}^{K_s}$ weights each coefficient of the 3DMM and $a_s \in \mathbb{R}^{3N}$ is the mean face mesh vertices. We use linear blendshapes to represent face expressions over the neutral identity $e: v = e + \Sigma_e \delta$, where v is the final vertex position displaced from e by blendshape weights vector $\delta \in \mathbb{R}^{K_e}$ and $\Sigma_e \in \mathbb{R}^{3N \times K_e}$ composed of $K_e = 75$ components of the expression space.

Face reflectance Similar to [25], we use a simplified Cook-Torrance BRDF [52, 53] with a constant roughness term. This BRDF model has the advantage of modelling specular reflections compared to the pure Lambertian BRDF. For each vertex, we define a diffuse $c_i \in \mathbb{R}^3$ and a specular $s_i \in \mathbb{R}^3$ albedos. The statistical diffuse $c \in \mathbb{R}^{3N}$ and specular $s \in \mathbb{R}^{3N}$ albedos are obtained from statistical prior of [30], where $c = a_r + \Sigma_r \beta$ and $s = a_b + \Sigma_b \beta$ with $\Sigma_r, \Sigma_b \in \mathbb{R}^{3N \times K_r}$ are the PCA for diffuse and specular reflectance with $K_r = 80$. a_r and a_b are the average skin diffuse and specular reflectance. We use the same coefficient β to sample the diffuse and specular albedos as in [30].

Illumination Similar to [25], we use nine spherical harmonics (SH) bands to model light. Dib *et al.* [10] showed that this high order SH parameterization provides a better light and shadows estimation when used with ray tracing compared to the widely used low-order 3 SH bands. An environment map of 64×64 is derived from SH to use with ray tracing. We define $\gamma \in \mathbb{R}^{9 \times 9 \times 3}$ the light coefficients.

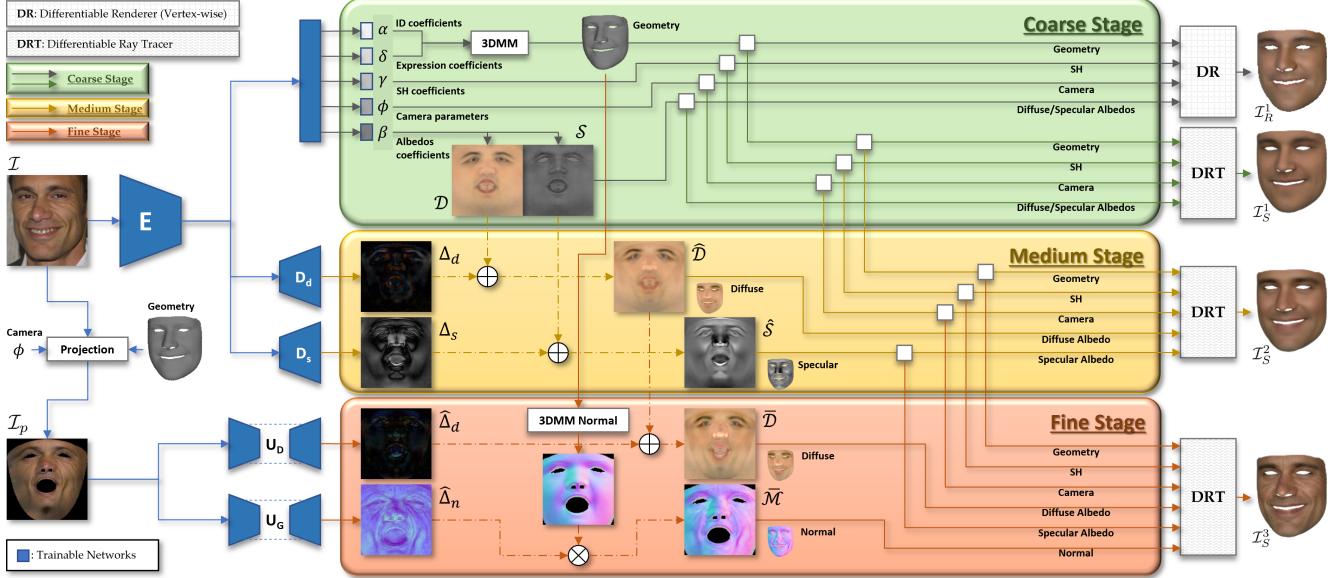


Figure 2. Our network architecture, trained end-to-end in a self-supervised manner, estimates face attributes (reflectance and detailed geometry) in a coarse-to-fine fashion (refer to section 3).

Camera We use the pinhole camera model and define $\phi = \{T, R\}$ as the camera parameters with rotation $R \in SO(3)$ and translation $T \in \mathbb{R}^3$.

3.2. Coarse stage

On Figure 2, the network **E** projects the input image I into the latent scene representation followed by a fully connected layer that predicts the semantic attributes vector $\chi = \{\alpha, \delta, \phi, \gamma, \beta\}$. Diffuse D and specular S texture representations are derived from β . These parameters are fed to a differentiable ray tracer (DRT) and to a vertex-wise differentiable renderer (DR) to generate two images I_S^1 and I_R^1 respectively. The following loss function is minimized during training:

$$E_d(\chi) + E_p(\alpha, \beta) + E_b(\delta), \quad (1)$$

where E_d is the data term equal to:

$$E_d(\chi) = E_{ph}^S(\chi) + w_{dr} E_{ph}^R(\chi) + w_{lm} E_{land}(\chi), \quad (2)$$

with E_{ph}^S is the pixel-wise photo-consistency loss between input and ray traced pixels $p_i, p_i^S \in \mathbb{R}^3$:

$$E_{ph}^S(\chi) = \sum_i |p_i^S(\chi) - p_i|, \quad (3)$$

where $p_i^S = \mathcal{F}(\chi)$, with \mathcal{F} , the Monte Carlo estimator of the rendering equation [54]. E_{ph}^R is the vertex-based photo-consistency loss between the projected mesh and the corresponding pixels in the input image, defined as follows:

$$E_{ph}^R(\chi) = \sum_{i=1}^N |\mathcal{B}(n_i, c_i, R_i) - I(\Pi \circ C(v_i))|, \quad (4)$$

where N is the number of vertices, $C(v_i)$ is the projection of vertex v_i in the real image, equal to: $R^{-1}(v_i - T)$. Π is the perspective camera matrix that projects a 3D vertex to a 2D pixel. \mathcal{B} is the final irradiance equal to the sum of the diffuse and specular terms weighted by the specular intensity s_i (details on \mathcal{B} in supplementary material section A). E_{land} is the landmark loss, which measures the distance between the $L = 68$ predicted facial landmarks and the projection of their corresponding vertex on the input image. These landmarks are obtained using the landmarks detector of [55]. E_p is the statistical prior that regularizes against implausible face geometry and reflectance [10]. $E_b(\delta)$ is a soft-box constraint that maintains δ in the range $[0, 1]$.

Edge discontinuities Ray tracing can naturally models self-shadows by building a visibility mask for each surface point with respect to each light. However, the major drawback of differentiable ray tracing is the discontinuities along geometry edges ([50, 51]). In fact, when solving for the rendering equation via Monte Carlo ray tracing [54], very few points are sampled on these areas. As a result, back-propagation fails to handle geometry edges accurately during the optimization. Several solutions have been proposed to overcome this limitation but they are generally very computationally expensive ([48, 51]). For instance, [51] explicitly samples the geometry edges, which extremely penalizes the training time as it needs to calculate the silhouette edges of the geometry. While landmarks are mainly used to guide the training, in the particular case of ray tracing they can help mitigating the aforementioned limitation. However the geometry is not as precise as it could be. As an efficient solution, we introduce in this work a new loss term E_{ph}^R (eq. 4)

which relies on a vertex-based differentiable renderer. This leads to a more accurate reconstruction, by taking advantage of ray tracing (which can model self-shadows) and vertex-based rendering (for better gradients on geometry edges) without significant cost. For instance, it only takes 370 ms to process (forward-backward) an image using our method compared to 42 seconds for the method of [51].

3.3. Medium stage

The albedos (diffuse and specular) and geometry estimated by the previous stage are bound by the statistical prior space and can only capture low frequency components of the skin reflectance and geometry. Our goal is to obtain personalized albedos (outside this space) with detailed geometry. Estimating these parameters jointly in a self-supervised manner is challenging. For this, we proceed with a coarse-to-fine strategy and we start by capturing personalized medium diffuse and specular albedos outside the statistical prior space. The challenge here is to avoid mixing diffuse and specular albedos and also to avoid baking unexplained shadows in these albedos. For this, we use the same technique of Dib *et al.* [25] which estimates personalized shadow-free albedos. For this, we train two additional decoders, \mathbf{D}_d and \mathbf{D}_s , in a self-supervised way, that estimate diffuse Δ_d and specular Δ_s increments to be added on top of the previously estimated textures, \mathcal{D} and \mathcal{S} , respectively. The resultant textures, $\hat{\mathcal{D}}$ and $\hat{\mathcal{S}}$, are used to generate a new image \mathcal{I}_s^2 . We note that the second stage has only access to latent space of \mathbf{E} allows this stage to focus on separating medium diffuse from specular albedo without worrying about high frequency geometry details that are discarded naturally by design. We define $\hat{\chi} = \{\alpha, \delta, \phi, \gamma, \hat{\mathcal{D}}, \hat{\mathcal{S}}\}$ and we minimize the following loss function:

$$\mathsf{E}_d(\hat{\chi}) + \mathsf{E}_{sc}(\hat{\mathcal{A}}, \mathcal{A}) + w_m \mathsf{E}_m(\hat{\mathcal{A}}) + w_b \mathsf{E}_b(\hat{\mathcal{A}}), \quad (5)$$

where $\mathsf{E}_{sc}(\hat{\mathcal{A}}, \mathcal{A}) = w_s \mathsf{E}_s(\hat{\mathcal{A}}) + w_c \mathsf{E}_c(\hat{\mathcal{A}}, \mathcal{A})$, and \mathcal{A} is either \mathcal{D} or \mathcal{S} . E_s and E_c are the symmetry and consistency regularizers (similar to [25]) used to avoid baking residual shadows in the personalized albedos. E_m is a constraint term that ensures local smoothness at each vertex, with respect to its first ring neighbors.

3.4. Fine stage

While the previous stage allows to obtain more personalized diffuse and specular albedos, these albedos remain generally blurry and still miss details. Also the geometry is restricted to the low-dimensional space of 3DMM. For this, we leverage the U-net based architecture (with skip connections) which are very efficient at capturing these fine details. First, using the pose and the geometry produced by the first stage, we project the input image \mathcal{I} in the uv-space. This projection, denoted as \mathcal{I}_p , is passed to two U-net networks,

\mathbf{U}_G and \mathbf{U}_D . \mathbf{U}_G predicts a normal map $\hat{\Delta}_n$ used to displace the original normal vectors of the coarse mesh. We denote $\bar{\mathcal{M}}$ the final normal map used for shading, where each vector $\bar{\mathbf{n}}_i$ in $\bar{\mathcal{M}}$ equal to: $\bar{\mathbf{n}}_i = \mathbf{T}_i \otimes \tilde{\mathbf{n}}_i$, with $\tilde{\mathbf{n}}_i$ sampled from $\hat{\Delta}_n$ and \mathbf{T}_i is a column-wise matrix that stack the original normal \mathbf{n}_i , tangent \mathbf{t}_i and bi-tangent \mathbf{b}_i vectors in the camera coordinate system (more on normal mapping with ray tracing in [56]). \mathbf{U}_D predicts an increment $\hat{\Delta}_d$ that is added to the estimated diffuse albedo $\bar{\mathcal{D}}$ of the previous stage. We denote $\bar{\mathcal{D}}$ as the resulting texture.

We define $\bar{\chi} = \{\alpha, \delta, \phi, \gamma, \bar{\mathcal{M}}, \bar{\mathcal{D}}\}$ and we train \mathbf{U}_G and \mathbf{U}_D in a self-supervised manner by minimizing the following loss function:

$$\mathsf{E}_d(\bar{\chi}) + \mathsf{E}_{sc}(\bar{\mathcal{D}}, \hat{\mathcal{D}}) + w_m^f \mathsf{E}_m(\bar{\mathcal{A}}) + w_b^f \mathsf{E}_b(\bar{\mathcal{A}}), \quad (6)$$

where $\mathsf{E}_{sc}(\bar{\mathcal{D}}, \hat{\mathcal{D}}) = w_s^f \mathsf{E}_s(\bar{\mathcal{D}}) + w_c^f \mathsf{E}_c(\bar{\mathcal{D}}, \hat{\mathcal{D}})$, and $\bar{\mathcal{A}}$ is either $\bar{\mathcal{D}}$ or $\bar{\mathcal{M}}$. The regularization terms E_s and E_c play an important role in avoiding baking unexplained shadows in diffuse texture $\bar{\mathcal{D}}$ in case our lighting model did not recover the light correctly. Also these regularizers contribute in producing a good separation between diffuse and geometry details. However, they sacrifice some albedo details (please refer to the ablation section 5). Finally, we note that we experimented using an additional U-net to capture a specular increment $\hat{\Delta}_s$ (similar to the diffuse) but we did not obtain substantial improvements in the reconstruction quality.

3.5. Training strategy

We proceed with the following training strategy. We first train \mathbf{E} (with the fully-connected layer) for 30 epochs, in a non-supervised manner, to directly regress χ by minimizing eq. 1. Next, we train \mathbf{D}_d , \mathbf{D}_s , and \mathbf{E} for 10 epochs while minimizing the loss in eq. 5. We follow the same training strategy proposed by [25] to separate diffuse and specular albedos, which consists in starting with a high regularization value of w_c for diffuse texture (in eq. 5), and then in progressively relaxing its value during training to allow for the diffuse albedo to capture more details. Next, we fix \mathbf{D}_d and \mathbf{D}_s , then we train \mathbf{U}_G , \mathbf{U}_D and \mathbf{E} for 5 epochs, to estimate a normal map and an enhanced diffuse map respectively, by minimizing the photo-consistency loss in eq. 6. To avoid over-fitting one component, and to obtain a plausible separation between these attributes, we start with a high weight for w_c^f and we progressively relax this constraint to allow $\bar{\mathcal{D}}$ to capture more albedo details.

Finally, we note that the vertex-based loss in eq. 4 is used in the whole training process with the goal to assist and guide the pixel-wise photo-consistency loss of ray tracing (eq. 3) at different stages, so to alleviate the problem of noisy edge gradients that ray tracing exhibits.



Figure 3. Results with challenging face details. From left to right: Input image, overlay of our final reconstruction on the input image, final reconstruction, shading, normal, diffuse and specular.

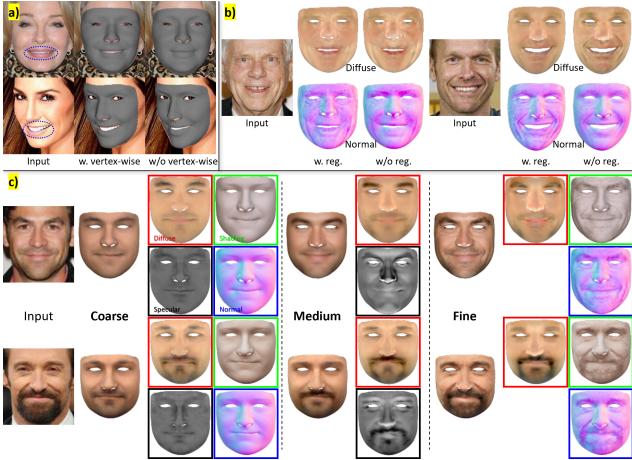


Figure 4. ablation studies (section 5)

4. Results

For training, we use CelebA dataset [57] where images were cropped to 256×256 . Output textures of different networks have the same resolution as the input image. We implemented our network using PyTorch [58]. Ray tracing is based on the method of [51]. During training, we use 8 samples per pixels for ray tracing. More implementation details are in supp. material, section B.

Figure 1, Figure 3 and supp. material (section E) show successful face reconstruction of more than 100 subjects with challenging face details, extreme lighting conditions, challenging head pose/expression and different skin type (makeup, beards)... For all these subjects, our method successfully estimates personalized albedos and captures fine detailed geometry, which leads to appealing reconstruction at high fidelity. Even in challenging lighting conditions, our method successfully estimates shadow-free maps (diffuse, specular and normal). All this, at very high computational speed, where at inference, our method takes 131 ms to process an image on a Nvidia RTX 2080 Ti. We note that while ray tracing penalizes our training time, it is not needed at inference time and our estimated attributes are compatible with existing rendering engines. Finally, our robust estimation of scene light, face reflectance and geometry provides explicit controls over the face attributes which leads to practical applications such as face attribute editing (aging) and relighting as shown in supp. material (section E). Finally, in supp. video, we also show reconstruction on video sequence.

5. Ablation

Importance of vertex-based renderer In this experiment, we study the importance of the vertex-based renderer to overcome the problem of noisy edge gradients of the ray tracer. For this, we trained \mathbf{E} by dropping the vertex-based

renderer based loss term (eq 4) from equation 2. We compare the estimated mesh to the one that use the full energy term (ray tracing + vertex-wise). The results in Figure 4 a) show that the estimated meshes using both the vertex-based renderer and ray tracer are more accurate than the ones obtained using ray tracing only (especially around the mouth edges). Quantitatively, we evaluate both methods on 100 subjects from Facescape dataset [19] with various type of facial expressions. We compute the vertex position error with respect to ground truth mesh, and we obtain 2.288/1.671 mm (mean error/std deviation) for the 'vertex-based + ray tracing' compared to 2.831/1.782 mm for 'ray tracing only' which show that combining ray tracing with the vertex-based renderer improves the reconstructed geometry. We note that reason why the improvement may look small is that 'vertex-based + ray tracing' aims to improve the geometry on very small area around the edges.

Regularization In this experiment, we study the importance of the symmetry and consistency regularizers (E_{sc}) used in equation 6 to separate the diffuse and geometric details faithfully. For this, we trained \mathbf{U}_G and \mathbf{U}_D by dropping these two regularizers. For both subjects in Figure 4 b), without these regularizers, some geometric details get baked in the albedo and leads to sub-optimal separation. Adding these regularizers produce more convincing separation. While these regularizers play an important role in obtaining a correct separation between diffuse and geometry details, they sacrifice some albedo details.

Multi-Stage reconstruction In this experiment we show the importance of the 'Medium' and 'Fine' stages to improve the realism of the 'Coarse' reconstruction. Figure 4 c) shows the reconstruction obtained from the 'Coarse' stage with the estimated statistical albedo priors. For the 'Medium' stage, we show the corresponding reconstruction and the enhanced diffuse and specular albedos. For the 'Fine' stage, we show the final reconstruction with the final diffuse and normal maps. We note that the diffuse albedo in 'Medium' stage is generally blurry and lacks some details and is significantly enhanced in the 'Fine' stage. Also, the detailed geometry captured in the 'Fine' stage significantly improves the realism of the final reconstruction. Quantitatively, we calculate the SSIM between the reconstruction of each stage and the original input image. On 1000 images, we obtain an average of 0.89/0.91/**0.97** for the coarse/medium/fine stages respectively (higher is better). This shows that the 'Fine' stage significantly improves the fidelity of the reconstruction.

6. Comparison

In this section, we compare, qualitatively and quantitatively, our method to the state-of-the-art methods.

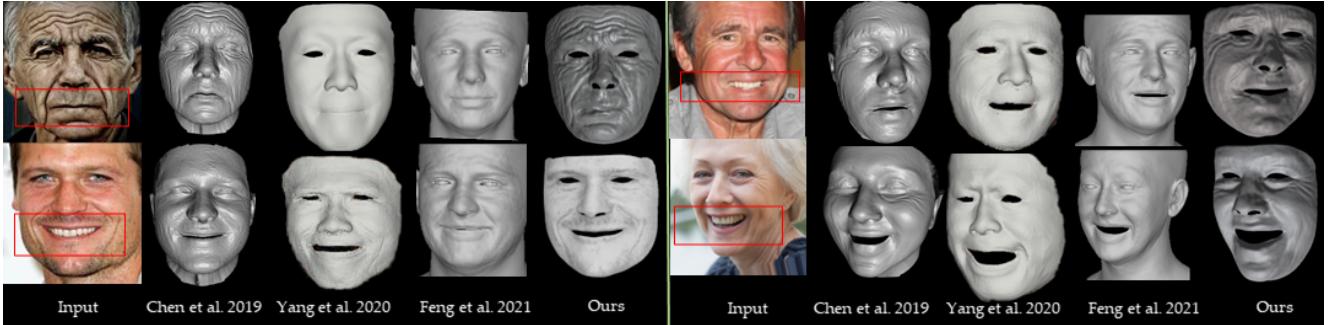


Figure 5. Comparison against Chen *et al.* [17], Yang *et al.* [19] and Feng *et al.* [24]

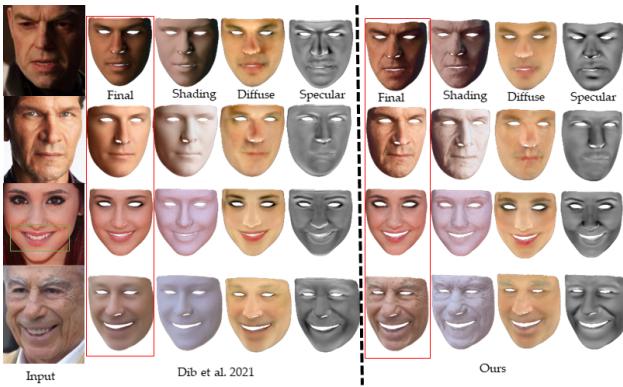


Figure 6. Comparisons against Dib *et al.* [25] (top subject from [25])

6.1. Qualitative comparison

Figure 5 shows comparison against Chen *et al.* [17], Yang *et al.* [19] and Feng *et al.* [24]. For [17] and [24], results are from authors’ open implementation. For [19] results are generated by the authors. The methods of [17] and [19] use ground truth (GT) data to train their generative network while our method and [24] are self-supervised methods and do not require any GT data. For all subjects, our method successfully capture most of geometry details especially for the top subjects that present challenging details. These details are barely captured or missed by other methods. Also our method shows better results on the wrinkles formed by the zygomaticus muscles (smile wrinkles around nose and mouth). Our method has a significantly better shape and expression recovery than all other methods, especially around the mouth (as highlighted in red rectangles), where the expression is incorrectly captured by other methods.

Compared to Dib *et al.* [25] (Figure 6), our method achieves robustness against challenging lighting conditions similar to [25] and produces shadow free albedos (first two subjects). In addition, our method estimates better diffuse map and captures detailed geometry, while [25] restricts the geometry reconstruction to the low-dimensional space of 3DMM.

This leads to a superior and high fidelity reconstruction of our method compared to [25] (highlighted in red rectangle). Finally, for third subject, our method that combine ray tracing and vertex-based rendering has a better expression recovery around the mouth than [25] that uses only ray tracing, which also confirms our earlier ablation study (highlighted in green rectangle).

Figure 7 show comparaison against the method of Abrevaya *et al.* [23] on subjects with challenging facial details. The method of Abrevaya *et al.* [23] uses a combination of labeled and unlabeled data to train the network that predicts normal map of the face. It also produces a complete normal map for the entire head (including eyes and moth interior) while our method restricts the reconstruction to the frontal region of the face (without eyes and mouth interior). However, our method captures more facial details (especially arround the eyes) than [23]. Finally we note that [23] only predicts a normal map (in camera space and not in uv space), and other face attributes are not estimated. Our method, estimates rich face attributes maps in uv space (normal, diffuse and specular), face geometry and scene light. More comparisons against other recent methods can be found in supp. material, section C.

6.2. Quantitative comparison

Geometric evaluation We first compare our estimated geometry to the state-of-the-art methods of Chen *et al.* [17], Feng *et al.* [24], and Dib *et al.* [25] on the Superfaces dataset [59] composed of 20 high resolution 3D ground truth (GT) face meshes (Table 1 and Figure 8). Table 1 reports, for each method, on all subjects, the average error μ and standard deviation σ for vertex position error. As shown in Figure 8, the same mask is used for all methods. Feng *et al.* [24] achieves slightly better results than our method on average error while ours has a smaller standard deviation. Our method achieves better score than Chen *et al.* [17] and Dib *et al.* [25]. As depicted in Figure 8, our approach measures lowest error around the mouth. Finally, we note that our method, which combines ray tracing with vertex-based renderer, has lower error than Dib *et al.* [25] that only uses

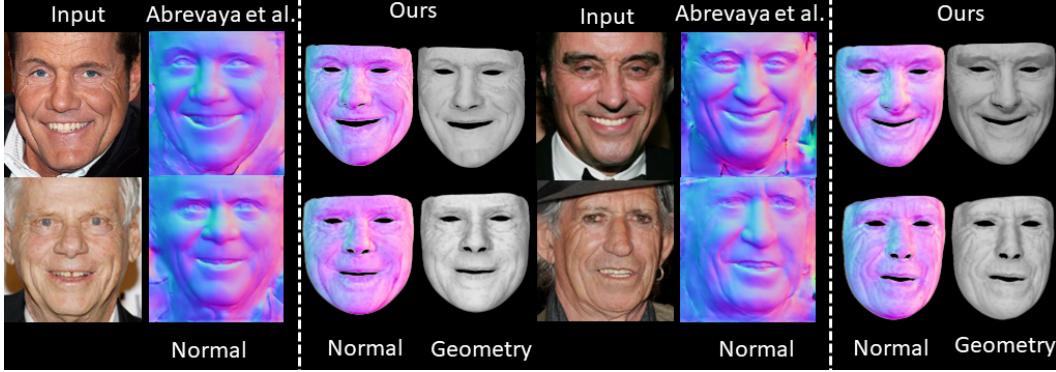


Figure 7. Comparison against Abrevaya *et al.* [23]

Table 1. Position error (mean/stdev) on Superfaces dataset [59]

Name	Chen <i>et al.</i> [17]	Feng <i>et al.</i> [24]	Dib <i>et al.</i> [25]	Ours
Mean err μ (mm)	1.847	1.234	1.379	1.287
Std dev σ (mm)	1.512	1.206	1.159	1.025

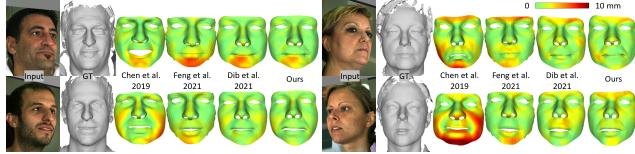


Figure 8. Vertex position error compared to GT mesh of Superfaces dataset [59].

Table 2. Position error (median/mean/stdev) on Now dataset [60]

Name	Feng <i>et al.</i> [24]	Dib <i>et al.</i> [25]	Ours
Median err (mm)	1.09	1.26	1.24
Mean err (mm)	1.38	1.57	1.54
Std dev (mm)	1.18	1.31	1.29

ray tracing, which again confirms our earlier ablation study (Section 5).

We also evaluate our method on the NoW dataset [60], which is composed of 80 subjects with a total of 1702 images. Results are reported in Table 2. We note that this dataset only evaluates mesh in neutral pose, so expression accuracy is not evaluated. Nevertheless, our method achieves competitive results and a better score than Dib *et al.* [25]. Feng *et al.* [24] achieves the top score (more scores against other methods are on NoW website). We note that for both datasets, for our method, the mesh used for comparison is the base 3DMM geometry estimated by the “coarse” stage as our method only estimates a normal map to model the finer details.

Normal evaluation We also compare our estimated normal map to Chen *et al.* [17] and Feng *et al.* [24] on Emily [61] and on realistic 3D head model [62] which we call ‘Male014’ in the next. Results are reported in Figure 9 and Table 3. Since each method uses a different UV map parametrization, the comparison was done on the rendered image and not on the unwrapped texture using the same mask (as shown in Figure 9). For ‘Male014’, our method

Table 3. Mean angular error (degrees) and percentage of errors below 20°, 25° and 30°. (Top-row: **Male014**; Bottom-row: **Emily**)

Name	Mean±Std	< 20°	< 25°	< 30°
Ours	22.9±15.3	51.6%	67.0%	77.1%
Feng et al. [24]	24.2±14.5	43.8%	59.7%	72.4%
Chen et al. [17]	26.2±15.6	39.4%	55.8%	68.5%
Ours	16.8±9.9	71.2%	84.7%	92.8%
Feng et al. [24]	15.2±12.2	77.2%	84.5%	89.4%
Chen et al. [17]	17.0±13.9	72.4%	82.7%	88.2%

has the lowest mean error. For Emily, [24] has a better average error but our method has a lower standard deviation. Our method scores the best in error percentage under 20°, 25° and 30° for both subjects except for Emily 20°. We note that the method of Chen *et al.* [17] does not correctly recover the mouth expression of both subjects (Figure 9). Finally, the last column in Figure 9 shows the final reconstruction of our method overlaid on the input image.

7. Limitations, Future works and Conclusion

Limitations and Future works First, separating light color from skin color using a single image is not solved in this work and remains an open challenge, therefore, our method may sometimes bake some albedo color in the estimated light, that is why our shading may look reddish sometimes. Second, since we do not use symmetry and consistency regularizers for the normal map, some shadows may get baked in the normal map when the method fails to recover to correct light of the scene. Our experiments show that adding these regularizers can mitigate this but sacrifices important geometry details. Third, regularizers used in eq. 6 allow our method to obtain a correct separation between diffuse albedo and geometry details but this is at the expense of some albedo details (cf. section 5). Finally, our method cannot handle occlusions, so face props (such as glasses) get baked in the estimated maps. Work such [63] is important to tackle this. As future work, our method does not connect expression to the geometry details as in [24] which is important to obtain realistic animated rigs. Such a function could

leverage the rich representation produced by our pipeline. Finally, our method estimates view and light dependent face attributes and can be extended to video/multi-view based reconstruction which can significantly improve the estimated facial attributes and alleviate a lot of ambiguity introduced by using a single image only. Some limitations are shown in supp. material, section D.

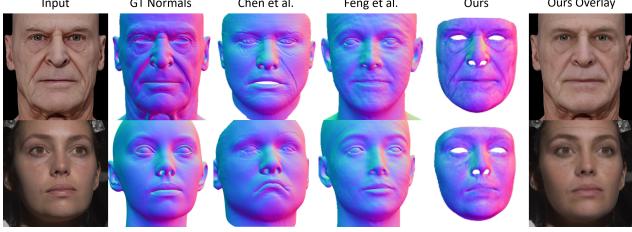


Figure 9. Estimated normal map compared to GT normal.

Conclusion In this work, we push to a new level the principle of analysis-by-rendering within self-supervised learning framework by refining the modelling, the training as well as the rendering stages. First, by combining, for the first time, ray tracing with the vertex-based renderer at training time, to solve the problem of edge-discontinuities of ray tracing which significantly improves the overall geometry and suggests an improvement on the original ray tracing algorithm. Second, by introducing a coarse-to-fine deep architecture, with adapted training strategy, that solve, for the first time, the highly challenging problem of separating detailed face attributes (reflectance and geometry) from a single image, within a self-supervised setup, and under uncontrolled lighting conditions. Compared to recent state-of-the art, our method achieves superior reconstruction quality and produces more visually appealing results and define a new baseline for self-supervised monocular face reconstruction methods '*in-the-wild*'.

8. Acknowledgements

We would like to thank authors in [24], [17] and [23] for sharing their code source publicly and authors of [19] for processing our images with their method.

References

- [1] Thabo Beeler, Fabian Hahn, Derek Bradley, Bernd Bickel, Paul Beardsley, Craig Gotsman, Robert W Sumner, and Markus Gross. High-quality passive facial performance capture using anchor frames. In *ACM Transactions on Graphics (TOG)*, volume 30, page 75. ACM, 2011. [1](#), [2](#)
- [2] Chenglei Wu, Kiran Varanasi, Yebin Liu, Hans-Peter Seidel, and Christian Theobalt. Shading-based dynamic shape refinement from multi-view video under general illumination. In *2011 International Conference on Computer Vision*, pages 1108–1115. IEEE, 2011. [1](#)
- [3] Levi Valgaerts, Chenglei Wu, Andrés Bruhn, Hans-Peter Seidel, and Christian Theobalt. Lightweight binocular facial performance capture under uncontrolled lighting. *ACM Trans. Graph.*, 31(6):187–1, 2012. [1](#), [2](#)
- [4] Abhijeet Ghosh, Graham Fyffe, Borom Tunwattanapong, Jay Busch, Xueming Yu, and Paul Debevec. Multiview face capture using polarized spherical gradient illumination. In *ACM Transactions on Graphics (TOG)*, volume 30, page 129. ACM, 2011. [1](#), [2](#)
- [5] P Gotardo, J Riviere, D Bradley, et al. *Practical Dynamic Facial Appearance Modeling and Acquisition*. ACM SIGGRAPH Asia, 2018. [1](#), [2](#)
- [6] Jérémie Rivière, Paulo Gotardo, Derek Bradley, Abhijeet Ghosh, and Thabo Beeler. Single-shot high-quality facial geometry and skin appearance capture. 2020. [1](#), [2](#)
- [7] Pablo Garrido, Levi Valgaert, Chenglei Wu, and Christian Theobalt. Reconstructing detailed dynamic face geometry from monocular video. *ACM Transactions on Graphics*, 2013. [1](#), [2](#), [3](#)
- [8] Supasorn Suwajanakorn, Ira Kemelmacher-Shlizerman, and Steven M Seitz. Total moving face reconstruction. In *European Conference on Computer Vision*, pages 796–812. Springer, 2014. [1](#), [2](#)
- [9] Pablo Garrido, Michael Zollhöfer, Dan Casas, Levi Valgaerts, Kiran Varanasi, Patrick Perez, and Christian Theobalt. Reconstruction of Personalized 3D Face Rigs from Monocular Video. *{ACM} Trans. Graph. (Presented at SIGGRAPH 2016)*, 35(3):28:1–28:15, 2016. [1](#), [2](#), [3](#)
- [10] Abdallah Dib, Gaurav Bharaj, Junghyun Ahn, Cédric Thébault, Philippe-Henri Gosselin, Marco Romeo, and Louis Chevallier. Practical face reconstruction via differentiable ray tracing. *Computer Graphics Forum*, 2021. [1](#), [2](#), [3](#), [4](#)
- [11] Ayush Tewari, Michael Zollöfer, Hyeongwoo Kim, Pablo Garrido, Florian Bernard, Patrick Perez, and Theobalt Christian. MoFA: Model-based Deep Convolutional Face Autoencoder for Unsupervised Monocular Reconstruction. In *The IEEE International Conference on Computer Vision (ICCV)*, 2017. [1](#), [2](#), [3](#)
- [12] Ayush Tewari, Michael Zollhöfer, Pablo Garrido, Florian Bernard, Hyeongwoo Kim, Patrick Pérez, and Christian Theobalt. Self-supervised multi-level face model learning for monocular reconstruction at over 250 hz. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 2549–2559. IEEE Computer Society, 2018. [1](#), [2](#)
- [13] Luan Tran and Xiaoming Liu. Nonlinear 3d face morphable model. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7346–7355, 2018. [1](#), [2](#)
- [14] Luan Tran, Feng Liu, and Xiaoming Liu. Towards high-fidelity nonlinear 3d face morphable model. In *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Long Beach, CA, June 2019. [1](#), [2](#)
- [15] A Tewari, F Bernard, P Garrido, G Bharaj, et al. *FML: Face Model Learning from Videos*. CVPR, 2019. [1](#), [2](#)

- [16] S Yamaguchi, S Saito, et al. *High-fidelity Facial Reflectance and Geometry Inference from an Unconstrained Image*. ACM TOG, 2018. 2
- [17] Anpei Chen, Zhang Chen, Guli Zhang, Kenny Mitchell, and Jingyi Yu. Photo-realistic facial details synthesis from single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9429–9439, 2019. 2, 8, 9, 10
- [18] Alexandros Lattas, Stylianos Moschoglou, Baris Gecer, Stylianos Ploumpis, Vasileios Triantafyllou, Abhijeet Ghosh, and Stefanos Zafeiriou. Avatarme: Realistically renderable 3d facial reconstruction” in-the-wild”. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 760–769, 2020. 2
- [19] Haotian Yang, Hao Zhu, Yanru Wang, Mingkai Huang, Qiu Shen, Ruigang Yang, and Xun Cao. Facescape: A large-scale high quality 3d face dataset and detailed riggable 3d face prediction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 2, 7, 8, 10
- [20] Matan Sela, Elad Richardson, and Ron Kimmel. Unrestricted facial geometry reconstruction using image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1576–1585, 2017. 2, 3, 13
- [21] Soumyadip Sengupta, Angjoo Kanazawa, Carlos D Castillo, and David W Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces in the wild’. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6296–6305, 2018. 2, 3
- [22] Xiaoxing Zeng, Xiaojiang Peng, and Yu Qiao. Df2net: A dense-fine-finer network for detailed 3d face reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2315–2324, 2019. 2, 3
- [23] Victoria Fernández Abrevaya, Adnane Boukhayma, Philip HS Torr, and Edmond Boyer. Cross-modal deep face normals with deactivable skip connections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4979–4989, 2020. 2, 3, 8, 9, 10
- [24] Yao Feng, Haiwen Feng, Michael J Black, and Timo Bolkart. Learning an animatable detailed 3d face model from in-the-wild images. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 2, 3, 8, 9, 10
- [25] Abdallah Dib, Cedric Thebault, Junghyun Ahn, Philippe-Henri Gosselin, Christian Theobalt, and Louis Chevallier. Towards high fidelity monocular face reconstruction with rich reflectance using self-supervised learning and ray tracing. In *The IEEE International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 5, 8, 9
- [26] Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. 2, 3
- [27] Tianye Li, Timo Bolkart, Michael J Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4d scans. *ACM Transactions on Graphics (ToG)*, 36(6):194, 2017. 2
- [28] Bernhard Egger, William AP Smith, Ayush Tewari, Stefanie Wuhrer, Michael Zollhoefer, Thabo Beeler, Florian Bernard, Timo Bolkart, Adam Kortylewski, Sami Romdhani, et al. 3d morphable face models-past, present, and future. *ACM Transactions on Graphics (TOG)*, 39(5):1–38, 2020. 2
- [29] Michael Zollhöfer, Justus Thies, Derek Bradley, Pablo Garrido, Thabo Beeler, Patrick Pérez, Marc Stamminger, Matthias Nießner, and Christian Theobalt. State of the Art on Monocular 3D Face Reconstruction, Tracking, and Applications. 2018. 2
- [30] William AP Smith, Alassane Seck, Hannah Dee, Bernard Tiddeman, Joshua B Tenenbaum, and Bernhard Egger. A morphable face albedo model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5011–5020, 2020. 2, 3
- [31] Chenglei Wu, Derek Bradley, Markus Gross, and Thabo Beeler. An anatomically-constrained local deformation model for monocular face capture. *ACM transactions on graphics (TOG)*, 35(4):1–12, 2016. 2
- [32] Curtis Andrus, Junghyun Ahn, Michele Alessi, Abdallah Dib, Philippe Gosselin, Cédric Thébault, Louis Chevallier, and Marco Romeo. Facelab: Scalable facial performance capture for visual effects. In *The Digital Production Symposium, DigiPro ’20*, New York, NY, USA, 2020. Association for Computing Machinery. 2
- [33] Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2387–2395, 2016. 2
- [34] Thomas Gerig, Andreas Morel-Forster, Clemens Blumer, Bernhard Egger, Marcel Luthi, Sandro Schönborn, and Thomas Vetter. Morphable face models—an open framework. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 75–82. IEEE, 2018. 2, 3
- [35] Tatsuro Koizumi and William A. P. Smith. Look ma, no landmarks! unsupervised, model-based dense face alignment. In *IEEE European Conference on Computer Vision (ECCV)*, 2020. 2
- [36] Yu Deng, Jiaolong Yang, Sicheng Xu, Dong Chen, Yunde Jia, and Xin Tong. 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*, pages 0–0, 2019. 2
- [37] Stylianos Ploumpis, Evangelos Ververas, Eimear O’Sullivan, Stylianos Moschoglou, Haoyang Wang, Nick Pears, William Smith, Baris Gecer, and Stefanos P Zafeiriou. Towards a complete 3d morphable model of the human head. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 2
- [38] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael J Black. Learning to regress 3d face shape and expression from an image without 3d supervision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7763–7772, 2019. 2
- [39] Xiaoguang Tu, Jian Zhao, Zihang Jiang, Yao Luo, Mei Xie, Yang Zhao, Linxiao He, Zheng Ma, and Jiashi Feng. Joint 3d

- face reconstruction and dense face alignment from a single image with 2d-assisted self-supervised learning. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019. 2
- [40] Jianzhu Guo, Xiangyu Zhu, and Zhen Lei. 3ddfa. <https://github.com/cleardusk/3DDFA>, 2018. 2
- [41] Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. Towards fast, accurate and stable 3d dense face alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 2
- [42] Chen Cao, Derek Bradley, Kun Zhou, and Thabo Beeler. Real-time high-fidelity facial performance capture. *ACM Transactions on Graphics (ToG)*, 34(4):1–9, 2015. 2
- [43] Timur Bagautdinov, Chenglei Wu, Jason Saragih, Pascal Fua, and Yaser Sheikh. Modeling facial geometry using compositional vaes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3877–3886, 2018. 2
- [44] Loc Huynh, Weikai Chen, Shunsuke Saito, Jun Xing, Koki Nagano, Andrew Jones, Paul Debevec, and Hao Li. Mesoscopic Facial Geometry Inference Using Deep Neural Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8407–8416, 2018. 2
- [45] Luo Jiang, Juyong Zhang, Bailin Deng, Hao Li, and Ligang Liu. 3d face reconstruction with geometry details from a single image. *IEEE Transactions on Image Processing*, 27(10):4756–4770, 2018. 3
- [46] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 21(8):690–706, 1999. 3
- [47] Elad Richardson, Matan Sela, Roy Or-El, and Ron Kimmel. Learning detailed face reconstruction from a single image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1259–1268, 2017. 3, 13
- [48] Linjie Lyu, Marc Habermann, Lingjie Liu, Ayush Tewari, Christian Theobalt, et al. Efficient and differentiable shadow computation for inverse problems. *arXiv preprint arXiv:2104.00359*, 2021. 3, 4
- [49] Abdallah Dib, Gaurav Bharaj, Junghyun Ahn, Cedric Thebault, Philippe-Henri Gosselin, and Louis Chevallier. Face reflectance and geometry modeling via differentiable ray tracing. *ACM SIGGRAPH European Conference on Visual Media Production (CVMP)*, 2019. 3
- [50] Guillaume Loubet, Nicolas Holzschuch, and Wenzel Jakob. Reparameterizing discontinuous integrands for differentiable rendering. *Transactions on Graphics (Proceedings of SIGGRAPH Asia)*, 38(6), Dec. 2019. 3, 4
- [51] Tzu-Mao Li, Miika Aittala, Frédéric Durand, and Jaakko Lehtinen. Differentiable monte carlo ray tracing through edge sampling. *ACM Trans. Graph.*, 37(6):222:1–222:11, Dec. 2018. 3, 4, 5, 7
- [52] Robert L Cook and Kenneth E. Torrance. A reflectance model for computer graphics. *ACM Transactions on Graphics (TOG)*, 1(1):7–24, 1982. 3
- [53] Bruce Walter, Stephen Marschner, Hongsong Li, and Kenneth Torrance. Microfacet models for refraction through rough surfaces. pages 195–206, 01 2007. 3
- [54] James T. Kajiya. The rendering equation. In *Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH ’86*, pages 143–150, New York, NY, USA, 1986. ACM. 4
- [55] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *International Conference on Computer Vision*, 2017. 4
- [56] Eric Veach. *Robust Monte Carlo methods for light transport simulation*. Stanford University, 1998. 5
- [57] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 7, 13
- [58] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. 2017. 7, 13
- [59] Stefano Berretti, Alberto Del Bimbo, and Pietro Pala. Surfaces: A super-resolution model for 3d faces. In Andrea Fusello, Vittorio Murino, and Rita Cucchiara, editors, *Computer Vision – ECCV 2012. Workshops and Demonstrations*, pages 73–82, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. 8, 9
- [60] Soubhik Sanyal, Timo Bolkart, Haiwen Feng, and Michael Black. Learning to regress 3D face shape and expression from an image without 3D supervision. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, pages 7763–7772, June 2019. 9
- [61] Emily. The wikihuman project. <https://vgl.ict.usc.edu/Data/DigitalEmily2/>, 2017. Accessed: 2020-05-21. 9
- [62] 3D Scan Store. Captured assets for digital artists. www.3dscanstore.com, 2021. [Online; accessed 14-November-2021]. 9
- [63] Anh Tuan Tran, Tal Hassner, Iacopo Masi, Eran Paz, Yuval Nirkin, and Gérard Medioni. Extreme 3d face reconstruction: Seeing through occlusions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3935–3944, 2018. 9
- [64] Dhruv Mahajan, Ravi Ramamoorthi, and Brian Curless. A theory of frequency domain invariants: Spherical harmonic identities for brdf/lighting transfer and image consistency. *IEEE transactions on pattern analysis and machine intelligence*, 30(2):197–213, 2007. 13
- [65] Xuaner Zhang, Jonathan T. Barron, Yun-Ta Tsai, Xiuming Zhang, Ren Ng, and David E. Jacobs. Portrait shadow manipulation. volume 39, 2020. 14

A. Vertex-based implementation

Here we provide the implementation details for the final irradiance \mathcal{B} used by the vertex-based renderer (please refer to eq. 4 in main document).

The final irradiance \mathcal{B} is equal to the sum of the diffuse and specular terms weighted by the specular intensity $s_i \in \mathbb{R}$:

$$\mathcal{B}(\mathbf{n}_i, \mathbf{c}_i, \mathbf{R}_i) = (1 - s_i) \cdot \mathcal{B}_d(\mathbf{n}_i, \mathbf{c}_i) + s_i \cdot \mathcal{B}_s(\mathbf{R}_i) \quad (7)$$

The diffuse irradiance \mathcal{B}_d , is obtained by multiplying the SH coefficients, B_{lm} of the light with SH coefficients (A_l) of the *half-cosine* function ([64]):

$$\mathcal{B}_d(\mathbf{n}_i, \mathbf{c}_i) = \mathbf{c}_i \cdot \sum_{l=0}^8 \sum_{m=-l}^l A_l \cdot B_{lm} \cdot Y_{lm}(\mathbf{n}_i) \quad (8)$$

with $\mathbf{c}_i \in \mathbb{R}^3$ and $\mathbf{n}_i \in \mathbb{R}^3$ are the diffuse albedo and normal vector for each vertex respectively.

Similarly, The specular irradiance is obtained using a spatial convolution of the SH light coefficients with the BRDF kernel of the roughness, which is constant in our simplified Cook-Torrance BRDF model:

$$\mathcal{B}_s(\mathbf{R}_i) = \sum_{l=0}^8 \sum_{m=-l}^l S_l \cdot B_{lm} \cdot Y_{lm}(\mathbf{R}_i), \quad (9)$$

with \mathbf{R}_i is the reflection direction of the viewing vector \mathbf{W}_i with respect to the surface normal, and S_l are the SH coefficients of the BRDF function [64].

B. Implementation details

We use PyTorch [58] with Cuda-enabled backend gpu. Two GPUs were used for training with a total of 24GB of memory. We used Adam optimizer with default parameters. Celeba dataset [57] is used for training with 3K images kept for validation. Images are aligned and cropped to a 256x256 resolution. Because ray tracing is generally slow, it takes 10 hours to do a single epoch. However, our method does not need ray tracing at test time and achieves fast inference (131 ms to process an image). We use 8 samples per pixel for ray tracing and a batch size of 8. For \mathbf{E} , we use a pre-trained *ResNet-152*. The architecture for \mathbf{D}_d and \mathbf{D}_s is given in Table 4. For \mathbf{U}_G and \mathbf{U}_D , we use a U-net with skip connections with a pre-trained *vgg16* backbone from here¹. The last layer of \mathbf{U}_D , \mathbf{U}_G , \mathbf{D}_d and \mathbf{D}_s is initialized to output zero increment at the beginning of the training. Landmarks weight $w_{lm} = 0.1$, vertex-based renderer weight $w_{dr} = 0.5$. For the 'Medium' stage, smoothness regularizer $w_m = 0.0001$ for diffuse and specular albedos. The symmetry regularizer $w_s = 20$ for medium diffuse

¹<https://github.com/mkisantal/backboned-unet>

Layer	Architecture
1	ct(256, 160, 3),bn=ELU,c2d(160, 256, 1),bn,ELU
2	ct(256, 256, 3),bn,ELU=c2d(256, 128, 3),bn,ELU,c2d(128, 192, 3),bn,ELU
3	ct(192, 192, 3),bn,ELU,c2d(192, 96, 3),bn,ELU,c2d(96, 128, 3),bn,ELU
4	ct(128, 128, 3),bn,ELU,c2d(128, 64, 3),bn,ELU,c2d(64, 64, 3),bn,ELU
5	ct(64, 64, 3),bn,ELU,c2d(64, 32, 3),bn,ELU,c2d(32, 42, 3),bn,ELU
6	ct(42, 42, 3),bn,ELU,c2d(42, 21, 3),bn,ELU,c2d(21, 3, 3),Tanh

Table 4. Architecture for \mathbf{D}_d and \mathbf{D}_s (ct: ConvTranspose2d, c2d: Conv2D, bn:BatchNorm2d (pytorch)).

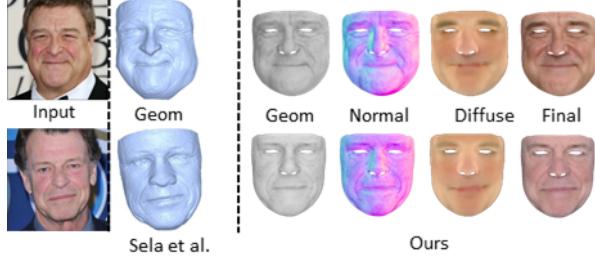


Figure 10. Comparison against Sela et al. [20]. Subjects from authors' manuscript.

and specular albedos. Consistency regularizer for specular albedo $w_c = 0.01$. For diffuse albedo, we start with a value of 0.2 and we relax it by a factor of 2 at each epoch. For the fine layer, the smoothness regularizer $w_m^f = 0.0001$ for diffuse and normal maps. The symmetry regularizer $w_s^f = 10$ for the final diffuse. The consistency regularizer, we start with $w_c^f = 1.0$ and we relax it by a factor of two at each epoch.

C. More qualitative comparison

In this section, we show more qualitative comparison against state-of-the art methods.

Figure 10 and Figure 11 show comparison results against the method of Sela et al. [20] and Richardson et al. [47]. Subjects are taken from authors' manuscript. We note that these two methods only estimate detailed geometry while our method estimates geometry, reflectance and scene light.

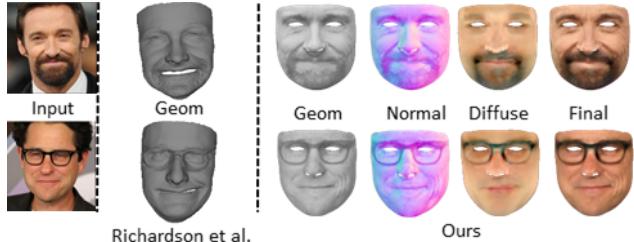


Figure 11. Comparison against Richardson et al. [47]. Subjects from authors' manuscript.

D. Limitations example

Figure 12 shows some limitations of our method. The description of the examples are as follows:

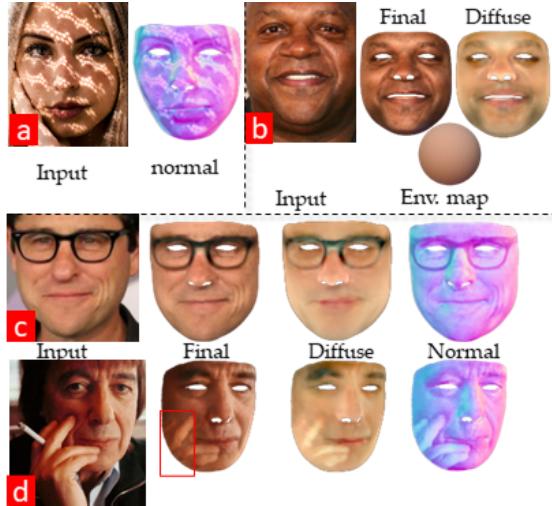


Figure 12. Limitations of our method.

- Subject a): Under external (foreign) shadows, our method fails to recover the light and leads to shadows baking in the estimated normal map. Tackling foreign shadows is a challenging problem, and the method such as Zhang *et al.* [65] described about the difficulties of handling this issue. We also note that our spherical harmonics based lighting model has its own limitations because it can only model infinite lights.
- Subject b): Separating light color from skin color using a single image remains challenging and our method does not solve for this. As a result, some skin color can get baked in the estimated light (represented here as an environment map).
- Subject c) and d): Occlusions and face props can get baked in the estimated maps. In case of severe occlusions (subject d), our method may fail to correctly estimate the geometry (highlighted in red box).

E. Results and relighting/aging applications

- Figure 3 and Figure 13 show reconstruction for subjects with challenging details on the face.
- Figure 14 show reconstruction for subjects under challenging lighting conditions.
- Figure 15 and Figure 16 show reconstruction for subjects with make-up, beards and face props.
- Figures 17, 18, 19 and 20 show reconstruction for various subjects with different ethnicity, skin color, difficult expression, challenging head pose.

For all these subjects, our method generalizes very well and achieves appealing 3D reconstruction with high fidelity

compared to the input image (first column vs second column). It also successfully estimates meaningful face attributes with faithful separation between reflectance and detailed geometry. Even under challenging lighting conditions, our method estimates plausible face reconstruction and produces shadow-free maps. In the supp. video, we show animated reconstruction with relighting and also reconstruction from video sequence.

Figure 21 shows relighting of different subjects under novel lighting conditions (second and third column). Because our method can successfully estimate shadow-free face attributes, and faithfully separates reflectance from detailed geometry, all this allow our method to perform relighting even for subjects under challenging light (last two subjects in Figure 21). The last three columns in Figure 21 show the estimated detailed geometry by our method rendered with OpenGL under different viewing angles.

Another practical applications for our method are face attribute(s) editing, attribute(s) transfer, aging and de-aging... We show in Figure 22 an 'Aging' application that consists simply on transferring the estimated normal map from source (A) to target (B).

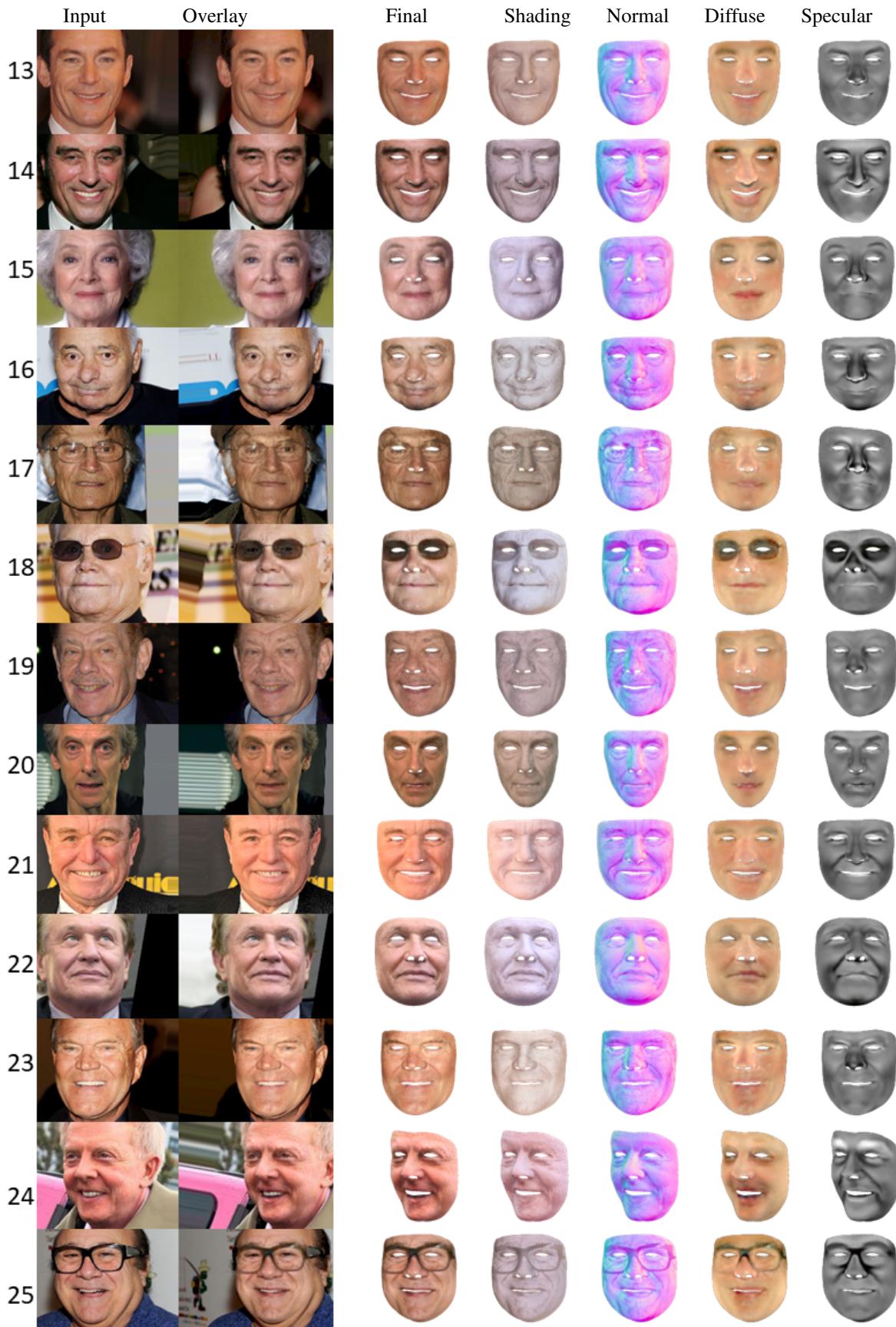


Figure 13. Results with challenging face details. From left to right: Input image, overlay of our final reconstruction on the input image, final reconstruction, shading, normal, diffuse and specular.

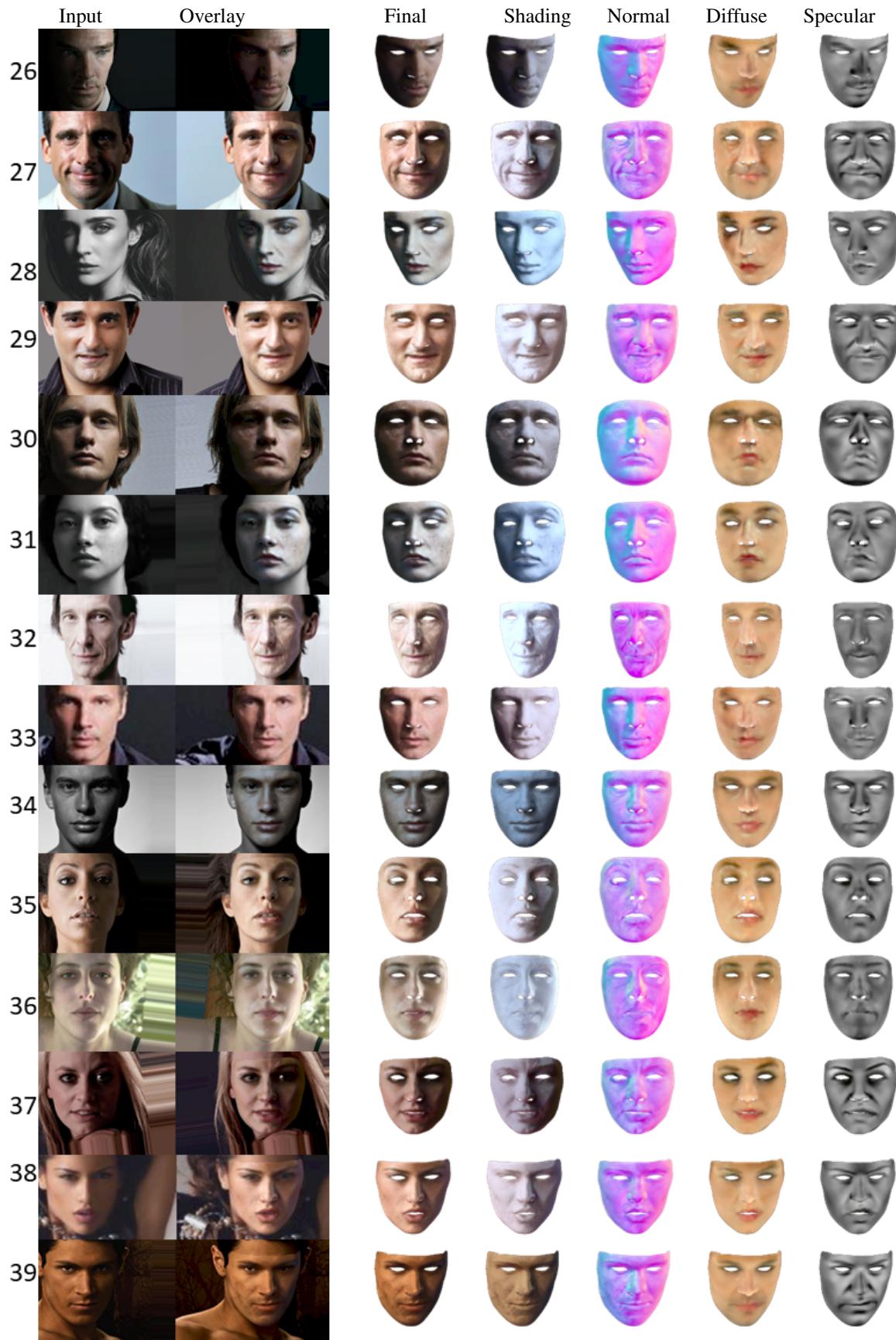


Figure 14. Results with challenging lighting conditions. From left to right: Input image, overlay of our final reconstruction on the input image, final reconstruction, shading, normal, diffuse and specular.



Figure 15. Results for subjects with make-up, beards and face props. From left to right: Input image, overlay of our final reconstruction on the input image, final reconstruction, shading, normal, diffuse and specular.



Figure 16. Results for subjects with make-up, beards and face props. From left to right: Input image, overlay of our final reconstruction on the input image, final reconstruction, shading, normal, diffuse and specular.

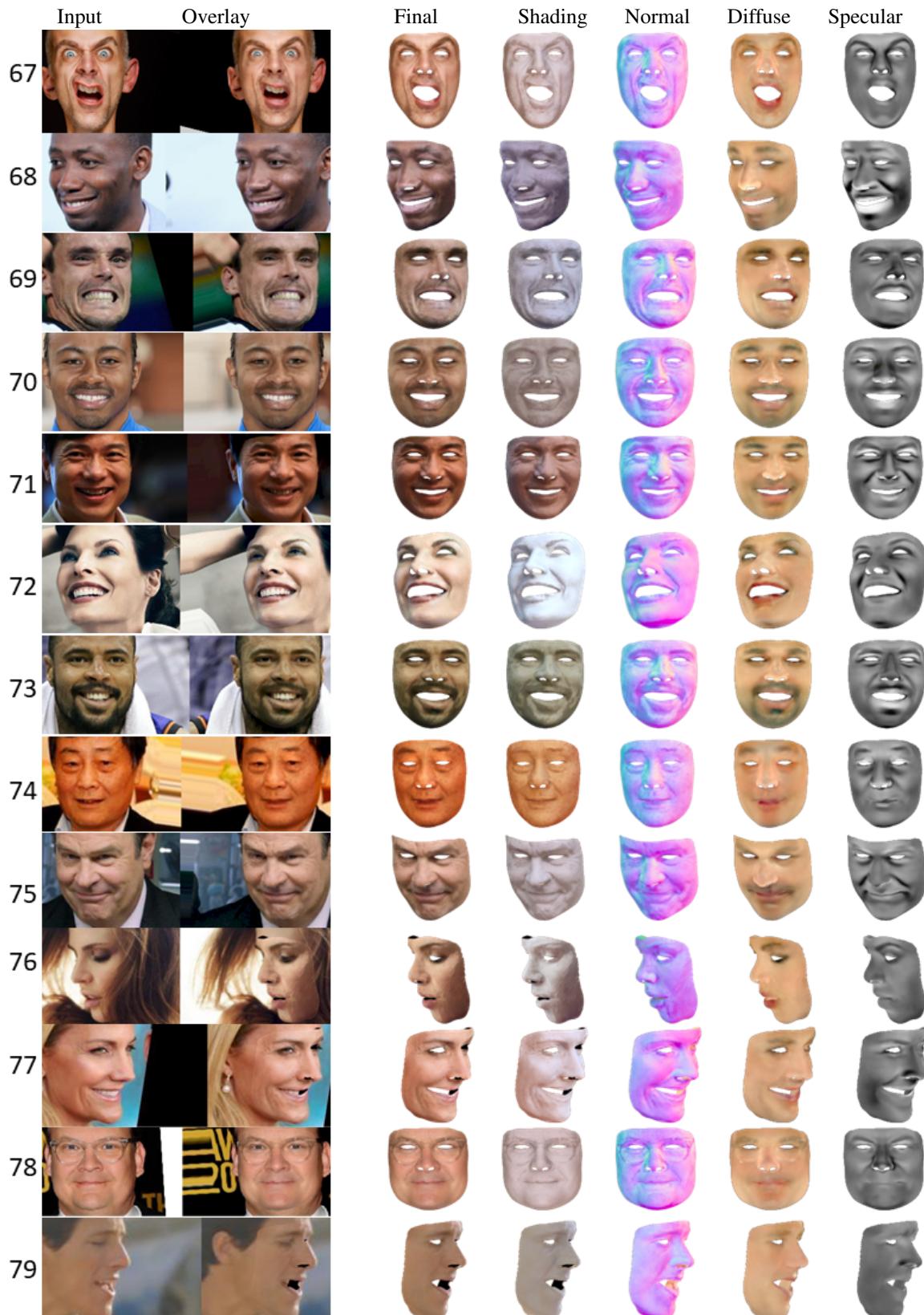


Figure 17. Results for subjects with different ethnicity, skin color, difficult expression and challenging head pose. From left to right: Input image, overlay of our final reconstruction on the input image, final reconstruction, shading, normal, diffuse and specular.

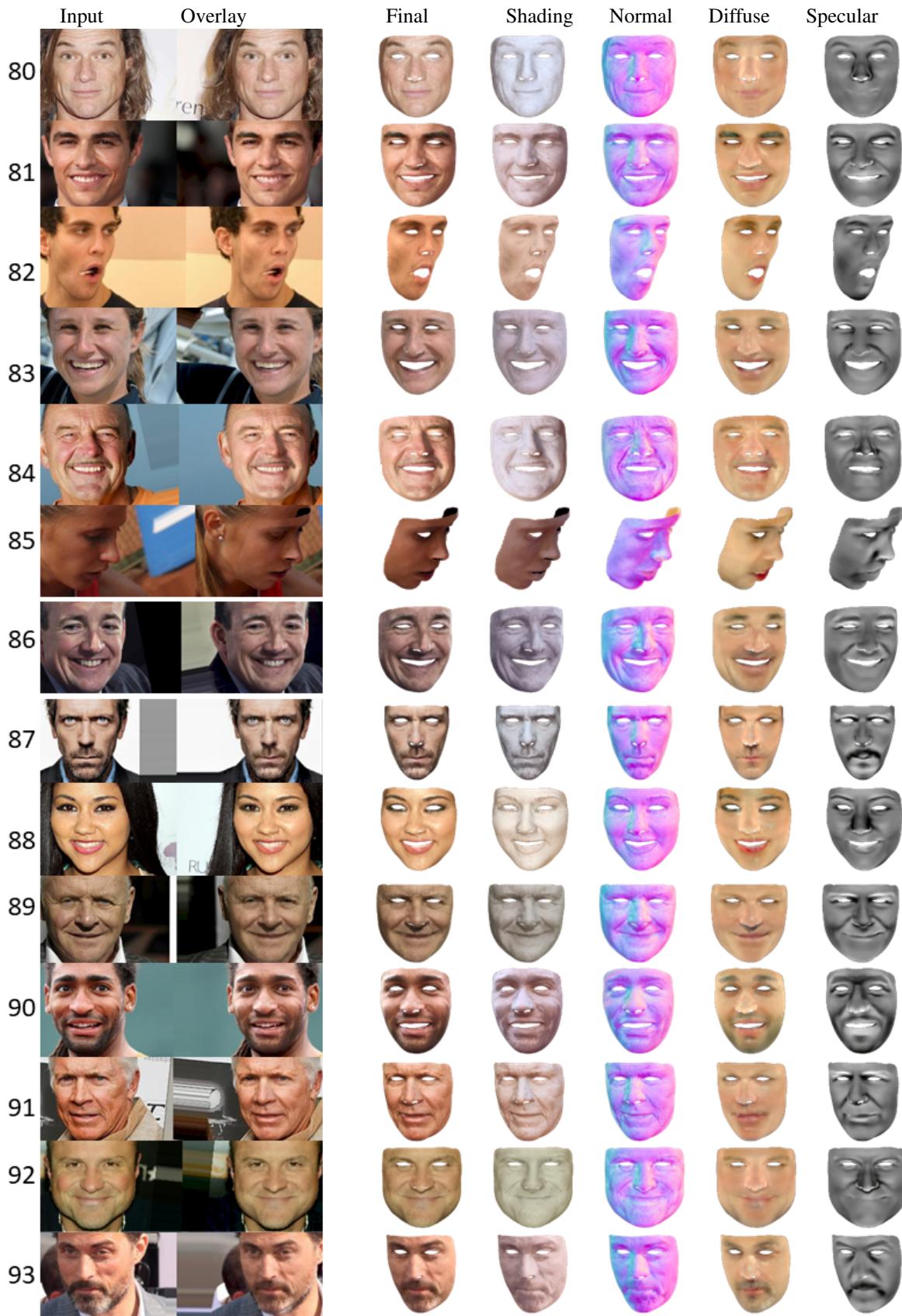


Figure 18. Results for subjects with different ethnicity, skin color, difficult expression and challenging head pose. From left to right: Input image, overlay of our final reconstruction on the input image, final reconstruction, shading, normal, diffuse and specular.

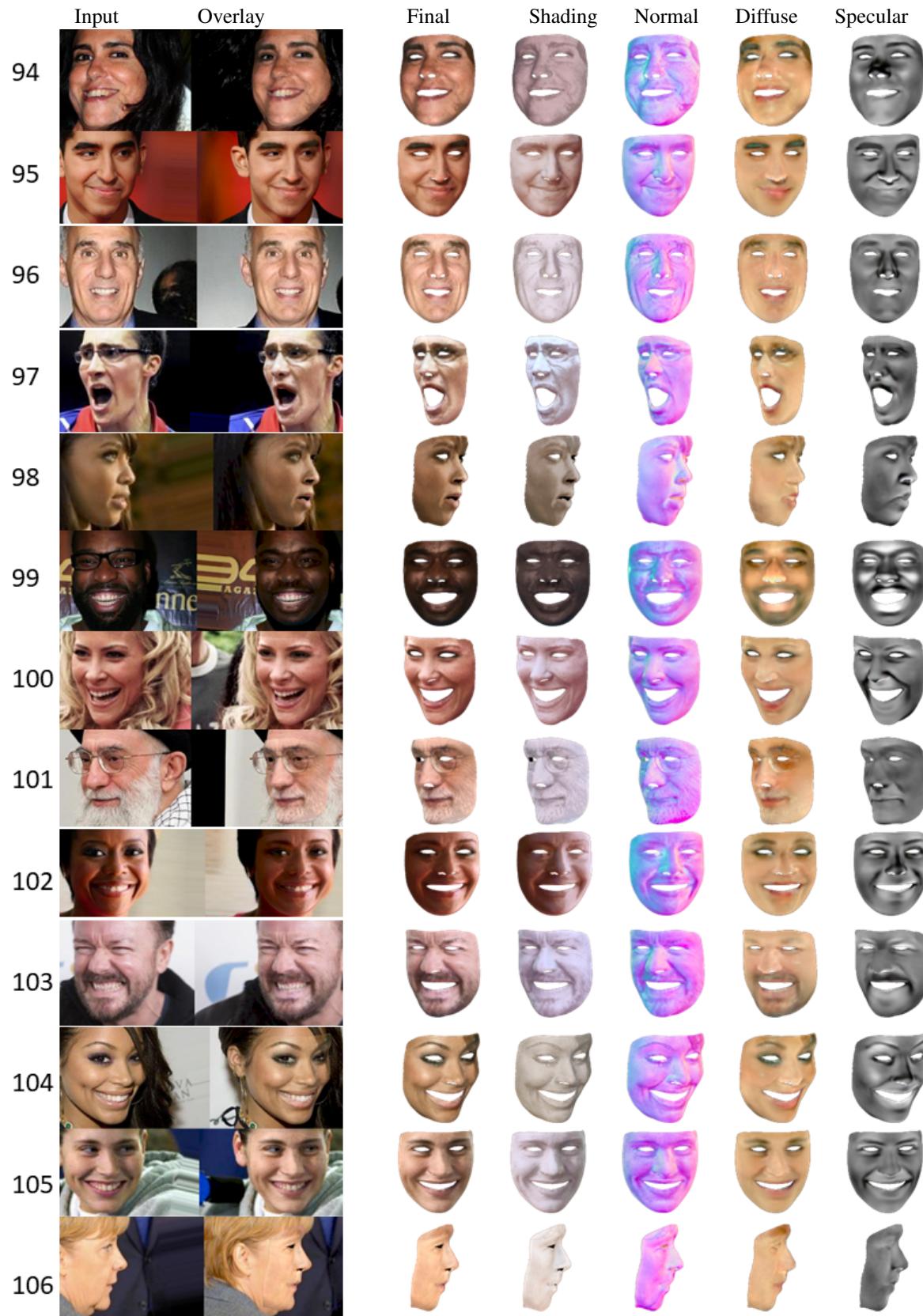


Figure 19. Results for subjects with different ethnicity, skin color, difficult expression and challenging head pose. From left to right: Input image, overlay of our final reconstruction on the input image, final reconstruction, shading, normal, diffuse and specular.

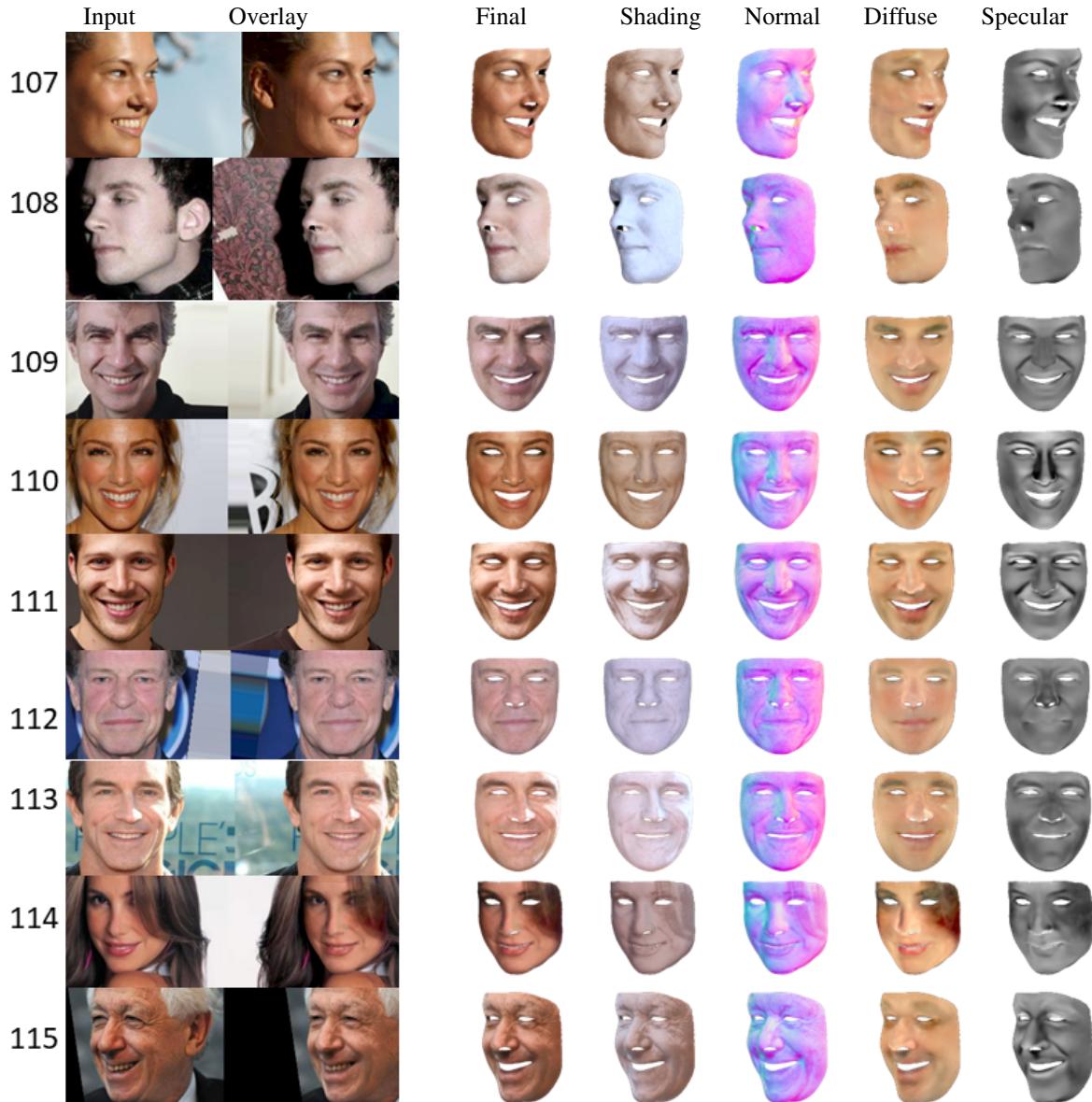


Figure 20. Results for subjects with different ethnicity, skin color, difficult expression and challenging head pose. From left to right: Input image, overlay of our final reconstruction on the input image, final reconstruction, shading, normal, diffuse and specular.

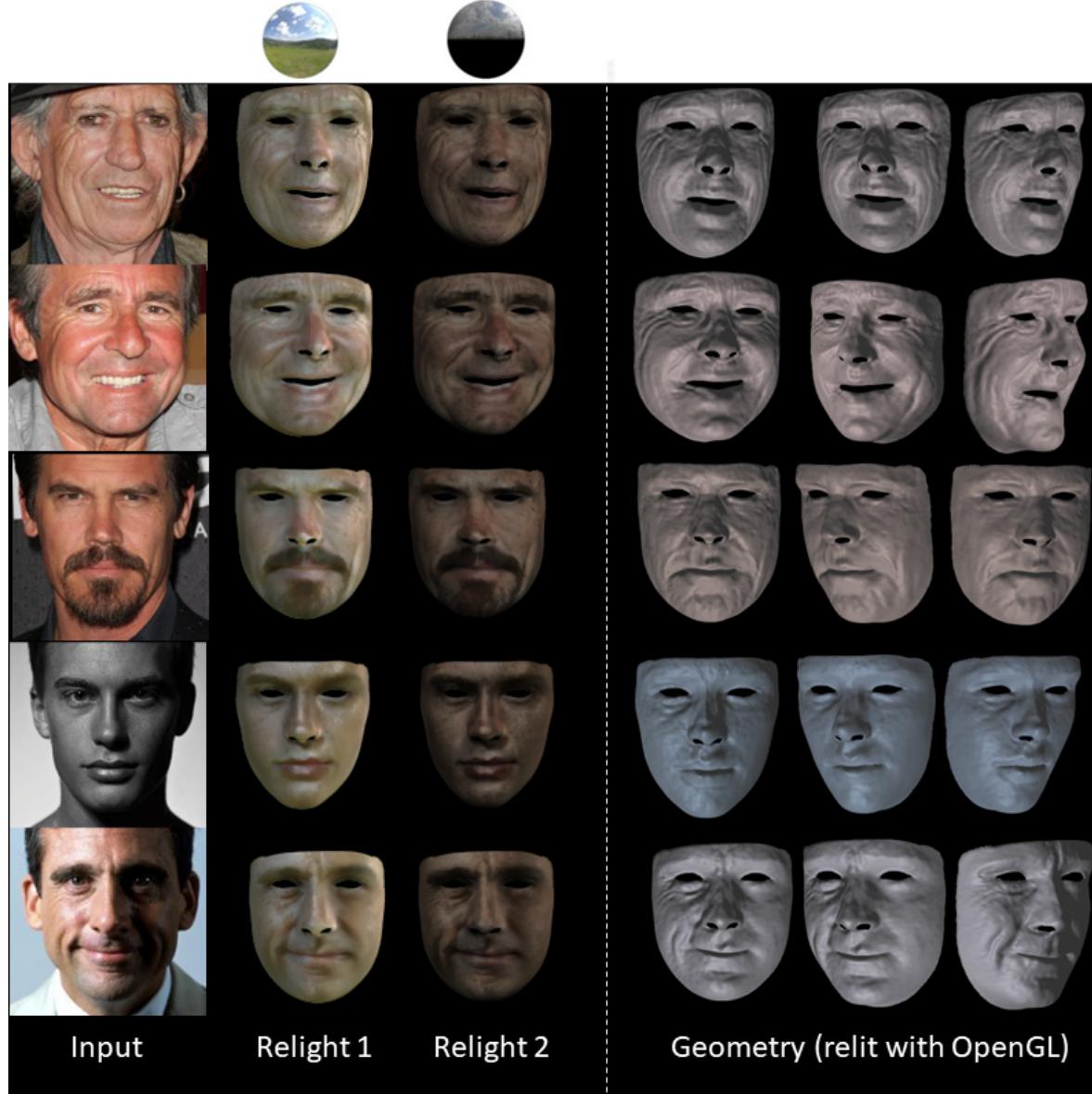


Figure 21. Our robust face attributes estimation gives explicit control over these attributes and allow for relighting even for subjects under challenging lighting conditions (last two subjects). The last three columns show the estimated geometry by our method for each subject rendered with OpenGL under different viewing angles.

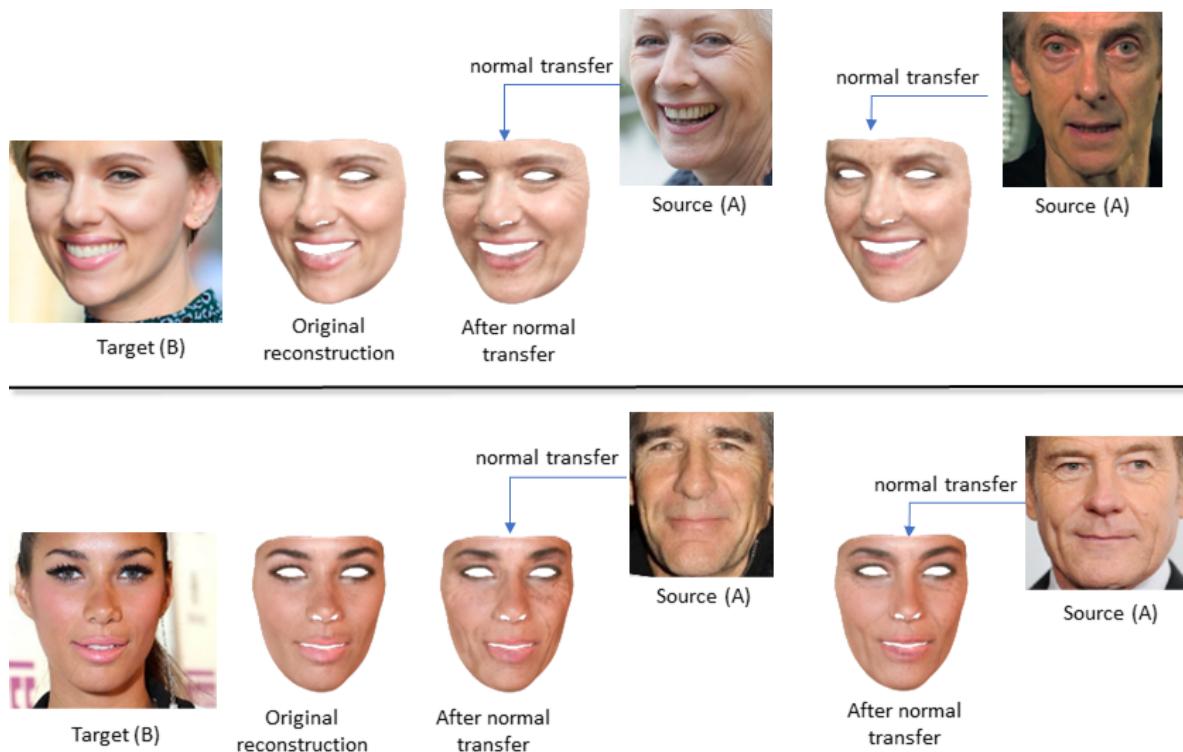


Figure 22. Face attributes edition: Our robust face attributes estimation allow for practical applications such as normal transfer from source (A) to target (B) which leads to aging/de-aging (here we show aging). Please note the wrinkles/folds that appears on the aged face.